UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

CENTRO DE CIENCIAS BÁSICAS

TESINA

Generación de un prototipo de sistema que utilice la teoría Record Linkage (Empate de Registros) en la conformación de Empresas a partir de un Directorio de Unidades Económicas.

PRESENTA

L.I. Sara Josefina Palacio Gámez

para obtener el grado de

Maestría en Informática y Tecnologías Computacionales

DIRECTORA DE TESIS

Dra. Laura A. Garza González

SINODALES

M.C. Jorge Eduardo Macías Luévano

M.C. Cesar E. Velázquez Amador

Cd. Universitaria

Aguascalientes, Ags. Mayo 2010

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES
Conmemoración del Bicentenario del inicio de la Independencia de México y del Centenario de la Revolución Mexicana

Centro de Ciencias Básicas

L.I. SARA JOSEFINA PALACIO GÁMEZ
PASANTE DE LA MAESTRÍA EN INFORMÁTICA
Y TECNOLOGÍAS COMPUTACIONALES
P R E S E N T E .

Estimado (a) Alumno (a) Palacio:

Por medio de este conducto me permito comunicar a Usted que habiendo recibido los votos aprobatorios de los revisores de su trabajo de tesis y/o trabajo práctico titulado: **"Generación de un prototipo de sistema que utilice la teoría Record Linkage (Empate de Registros) en la conformación de Empresas a partir de un Directorio de Unidades Económicas",** hago de su conocimiento que puede imprimir dicho documento y continuar con los trámites para la presentación de su examen de grado.

Sin otro particular me permito saludarle muy afectuosamente.

A T E N T A M E N T E
Aguascalientes, Ags., 31 mayo de 2010
"LUMEN PROFERRE"
EL DECANO

DR. FRANCISCO JAVIER ÁLVAREZ RODRIGUEZ

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES
CENTRO DE CIENCIAS BÁSICAS

c.c.p.- Archivo

Aguascalientes, Ags. Mayo de 2010

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

Dr. Francisco Javier Álvarez Rodríguez
Decano del Centro de Ciencias Básicas
Universidad Autónoma de Aguascalientes

P R E S E N T E.

Por este conducto, hago de su conocimiento que el tesista Sara Josefina Palacio Gámez, egresado de la Maestría en Informática y Tecnologías Computacionales del Centro de Ciencias Básicas, ha completado satisfactoriamente el documento de tesis titulado: "Generación de un prototipo de sistema que utilice la teoría Record Linkage (Empate de Registros) en la conformación de Empresas a partir de un Directorio de Unidades Económicas", por lo que se extiende el voto aprobatorio para que se haga la impresión del mismo, en cumplimiento a los requisitos de contenido y forma exigidos por la Universidad Autónoma de Aguascalientes .

A T E N T A M E N T E

Dra. Laura Garza González
Director de Tesis

M.C. Jorge E. Macías Luévanos
Revisor de Tesis

M.C. Cesar E. Velázquez Amador
Revisor de Tesis

Dedicatoria

Dedico este documento de investigación con mucho amor a mis hijos y a mi esposo (mis tres amores), que no solo me han apoyado e impulsado en mi desarrollo como mujer, madre, compañera y esposa, sino también como profesionista. Comprendiendo y tolerando los tiempos compartidos con un sinfín de actividades en estos últimos 3 años.

A mis pequeños Iván e Ian espero que comprendan el significado tan grande de su amor y apoyo, estando presentes en mi vida y en mis logros.

A mis padres quienes han sido los pilares y guías en mi formación y mi vida. A mis hermanos por sus palabras de aliento y el caminar siempre a mi lado.

A mis mejores amigas y fortalecidas mujeres quienes siempre me han enseñado el ver hacia adelante, las que en cada caída y tropezón me sostuvieron y con las que he contado en todo momento.

Por siempre agradecida.

Agradecimientos Especiales

A la MC Mónica Gladis Pérez Miranda, una gran mujer, amiga y compañera por su indiscutible apoyo y orientación en esta investigación.

A mi directora de tesis Dra. Laura A. Garza González, que con su gran experiencia y sabiduría me permitió culminar exitosamente este documento de investigación. Agradezco infinitamente el compartir conmigo sus conocimientos atesorados.

A mis compañeros de trabajo quienes dieron un toque especial en esta travesía por su apoyo y experiencia compartida, este documento no solo refleja un trabajo de investigación sino también la inquietud de más de alguno de ustedes por mejorar los proceso de trabajo.

Mil gracias a todos ustedes con mucho cariño.

Sara

Índice de Contenido

Índice de Anexos

Índice de Figuras

Índice de Tablas

Índice de Cuadros

Título.

Generación de un prototipo de sistema que utilice la teoría Record Linkage (Empate de Registros) en la conformación de un directorio de empresas a partir de un directorio de unidades económicas.

Tema.

Generación de un prototipo de un sistema que basado en la teoría Record Linkage conforme un directorio de empresas, a partir de un directorio de unidades económicas, en donde el uso del método estadístico permita identificar unidades económicas bajo la misma denominación de Razón Social y/o Nombre de Establecimiento.

Resumen.

A lo largo de los años algunas Instituciones u Organismos se han enfrentado con la problemática de EMPATAR[1] dos o más ficheros de longitud *n*, donde *n* puede o no ser igual, con información similar en algunos campos o variables, por ejemplo si se tuvieran dos ficheros con contenido de información de personas, en uno de ellos las variables o campos que lo forman sean: nombre, apellido paterno, apellido materno, calle, número y colonia y en otro solo nombre y domicilio, donde nombre este formado por apellido paterno, apellido materno y nombre(s) y domicilio de calle, número exterior, número interior, colonia, código postal entidad y municipio, aunque en ambos el contenido de sus campos es similar su distribución y orden no es el mismo, si no se cuenta con un identificador único que los relacione uno a uno el proceso de identificar cuales de ellos se encuentran en ambos ficheros se dificulta y debe realizarse la comparación por medio del contenido de sus variables.

[1] EMPATAR.- Proceso de relacionar dos ficheros mediante variables similares en contenido.

La necesidad de relacionar los ficheros uno a uno como una función inyectiva, o bien, el de agrupar toda la información en un solo fichero evitando que la información se duplique, representa un alto costo si los ficheros son de extensas longitudes y el procesamiento es manual o semiautomático.

El Instituto Nacional de Estadística y Geografía (INEGI)[2], quien es un organismo responsable de normar y coordinar el Sistema Nacional de Información Estadística y Geográfica, así como de realizar los censos nacionales, integrar el sistema de cuentas nacionales, y elaborar los índices nacionales de precios.

No se exenta de este tipo de problemáticas. La falta de un identificador único entre el Directorio Nacional de Unidades Económicas (DNUE) y los directorios externos de fuentes oficiales implica un costo y dificultad al momento de localizar registros de un fichero a otro, enfrentando la misma necesidad de realizar EMPATES por medio de cadenas de caracteres de las cuales no se tiene la certeza de que exista algún error tipográfico o bien, errores de omisión, traslape o adición de uno o más caracteres dentro de ellas.

Es por esto que se ha considerado y se propone la adopción de una de las metodologías derivadas del concepto RECORD LINKAGE ("Empate de Registros"), la cual básicamente consiste en relacionar dos o más ficheros por medio de variables de contenido similares, cuando no existe un identificador único que los relacione.

Con el propósito de facilitar el empate del DNUE con directorios externos, se propone la adopción de la teoría y conceptos relacionados.

[2] **LEY DEL SISTEMA NACIONAL DE INFORMACIÓN ESTADÍSTICA Y GEOGRÁFICA,** TÍTULO TERCERO, DIARIO OFICIAL DE LA FEDERACIÓN, SEGUNDA SECCIÓN, SECRETARÍA DE HACIENDA Y CRÉDITO PÚBLICO, DECRETO por el que se expide la Ley del Sistema Nacional de Información Estadística y Geográfica. TITULO TERCERO DE LA ORGANIZACIÓN Y FUNCIONAMIENTO DEL INSTITUTO, CAPÍTULO I, Del Instituto Nacional de Estadística y Geografía, Artículos 53 y 59.

Asimismo se pretende, que la metodología apoye la conformación de empresas con la identificación de registros que se encuentren bajo la misma denominación de Razón Social y/o Nombre del Establecimiento que se almacenan dentro de un directorio de unidades económicas.

Está tesina propone la utilización de esta teoría mediante el prototipo de un sistema que permita identificar las bondades de su adopción para la formación del Directorio de Empresas.

Tomando como base teórica para el inicio de esta investigación el artículo de Fellegi y Sunter en 1969. “A Theory for Record Linkage.”

1.- Introducción.

La información de los establecimientos y empresas captados en los Censos Económicos, provee el insumo fundamental para la creación de marcos y muestras de unidades económicas, las cuales son utilizadas como información base para encuestas e investigaciones ya sea a nivel empresa o establecimiento realizadas dentro del INEGI, contribuyendo así la parte imprescindible para llevar a cabo la misión del Instituto, la cual es "*Generar, integrar y proporcionar información estadística y geográfica de interés nacional, así como normar, coordinar y promover el desarrollo de los Sistemas Nacionales Estadístico y de Información Geográfica, con objeto de satisfacer las necesidades de información de los diversos sectores de la sociedad".*

Es por ello que se da la necesidad de mantener esta información actualizada y sin duplicidades ya que como se menciona anteriormente, es la información fuente requerida para el diseño de encuestas en donde la unidad de observación sean empresas o establecimientos que recaban, almacenan y procesan los datos necesarios para la publicación de índices y parámetros nacionales; información estratégica para el país.

De aquí la importancia de dar mantenimiento y actualización al Directorio Nacional de Unidades Económicas (DNUE)[3], paralelamente al directorio de empresas.

El segundo es integrado por la información de unidades económicas llamadas *únicos* y las empresas conformadas por más de una unidad económica (matriz y sucursales).

El DNUE para esta investigación lo definiremos como un sistema de información integrado en una base de datos que contiene la información de las unidades

[3] El DNUE es un directorio integrado con la información de unidades económicas a nivel de establecimientos y empresas para los sectores de transportes y construcción proveniente de información de los Censos Económicos.

económicas del país. La fuente principal de este directorio proviene de la captada en los Censos Económicos, y puede ser actualizada con información de otras fuentes oficiales, o bien del resultado de encuestas especiales o tradicionales y problemática o movimientos detectados en operativos de campo en períodos inter-censales.

El INEGI a través de la Dirección General de Estadística se encarga de mantener los Directorios y Marcos actualizados, tanto de población y vivienda como de unidades económicas que sirven como base para la generación de información estadística nacional.

Los Censos Económicos Nacionales[4] como el mismo Instituto lo define es *"la fuente más completa de información económica sobre el estado que guarda la economía mexicana en un momento determinado. Y esa valiosa información que se obtiene a través de los censos, sobre la totalidad de las unidades económicas que llevan a cabo sus actividades en el país sirve como punto de partida para que, tanto en el Sector Público como en el Sector Privado, realicen otras mediciones: encuestas e investigaciones económicas, estudios de planeación y de mercado y las Cuentas Económicas Nacionales, entre muchas otras".* Siendo este uno de los principales compromisos de los Censos Económicos con la sociedad mexicana.

Dentro de la Dirección General de Estadística se encuentra la Dirección Adjunta de Investigación y Normatividad, de ella depende la Dirección de Área de Diseño y Marcos Estadísticos y es a esta dirección de área a la que compete mantener actualizados los marcos de población y vivienda y el marco de unidades económicas, específicamente esta tarea la desarrolla la Subdirección de Diseño Muestral de Unidades Económicas, tal y como lo muestra la figura 1.

[4] Metodología de los Censos Económicos 2004

Organigrama Institucional

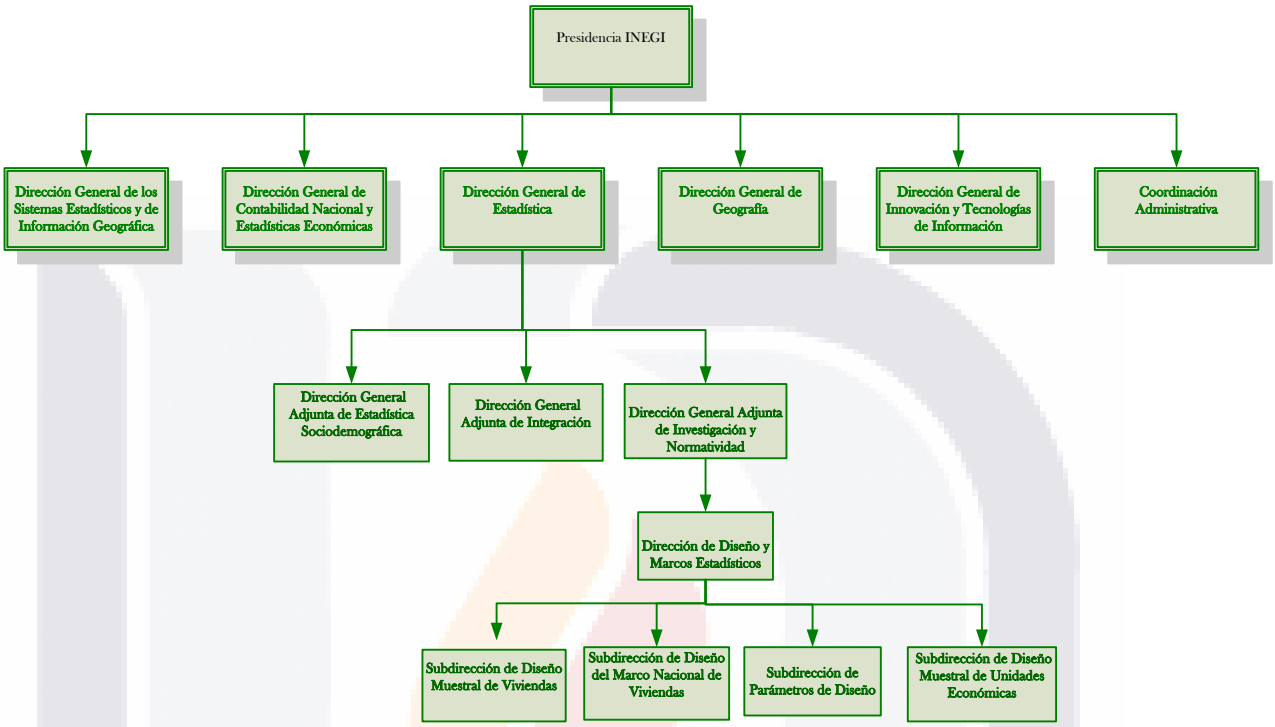


Figura 1. Instituto Nacional de Estadística y Geografía. Organigrama institucional.

2.- Problema.

En algunos organismos nacionales e internacionales de estadística que guardan y almacenan información, pueden enfrentar la necesidad de relacionar los registros de dos o más ficheros que carecen de un identificador único común que los relacione, pero que cuentan con información similar en el contenido de alguna(s) de su(s) variable(s) o campo(s).

Las necesidades de relacionarlos pueden ser variadas, desde identificar si tiene registros comunes, detectar registros duplicados dentro de cada fichero o bien al utilizar fuentes de actualización para un fichero maestro.

A este tipo de operaciones entre ficheros por medio de campos o variables similares se le denomina empate de ficheros o empate de registros.

Ésta actividad puede ser tediosa con cierto porcentaje de error, ya que depende de la perspectiva y el conocimiento de la persona que la está realizando o cuando los procesos son manuales o semiautomáticos, aun cuando existan definidos procedimientos previos. Al no contar con un identificador único, se tiene la necesidad de localizar la información de un fichero en otro por medio del contenido de las variables que componen el registro, que en la mayoría de los casos es información de tipo carácter. Adicionalmente, se tiene la posibilidad de que haya problemas en el contenido de los campos: errores de omisión, traslape o agregación de algún carácter a la cadena real derivado del proceso de captura de la información.

Un problema más posible de enfrentar, es que los ficheros o bases de datos tengan una longitud muy extensa, y haya diferencias entre un fichero y otro, en el contenido de la información como en la composición de las variables

Debido a lo anterior, el tiempo y personal involucrados en la realización de un empate entre ficheros de forma manual o semiautomática, podría variar en función de la longitud de los ficheros, sin tomar en cuenta la falta de precisión que se obtiene de esta manera; ya que implica que al comparar ficheros el número de pares por revisar sería de magnitud ($fichero1_{long\_n1} * fichero2_{long\_n2}$).

La Teoría Record Linkage gira alrededor de las investigaciones de empates o match's de registros basados en métodos estadísticos automatizados entre ficheros de longitud n1 y n2, respectivamente, cuando no existe un identificador único que los relacione uno a uno, los ficheros deben compartir información similar o común en el contenido de alguno(s) de sus campos.

Las metodologías RECORD LINKAGE se dividen en determinísticas y probabilísticas, dependiendo de los procedimientos empleados para la localización o empates de registros comunes entre ficheros.

El INEGI no es la excepción, no se exenta de la problemática de empatar ficheros, al no contar con un identificador único entre la información que se recaba de las unidades económicas que se encuentran en el DNUE y los directorios externos oficiales o de registros administrativos que pudieran actualizarlo.

El INEGI obtiene cada cinco años un Directorio Nacional de Unidades Económicas, como resultado del levantamiento de los Censos Económicos, que en el periodo inter-censal debe actualizarse, utilizando directorios externos derivados de registros administrativos y mediante la información procedente de los establecimientos visitados durante los operativos de encuestas o investigaciones.

Esta situación nos llevó a la propuesta de un prototipo de un sistema que genere un proceso automático que permita empatar el DNUE con los directorios externos,

identificar duplicados, o permitir ingresar actualizaciones de registros administrativos, y de la misma manera paralelamente mantener actualizado un directorio de empresas con la misma dinámica con la que se actualiza el DNUE.

Dentro del DNUE las empresas de origen censal se encuentran identificadas por una variable denominada *Folio*, esta variable identifica en el levantamiento del Censo Económico a todos las unidades económicas que forman una empresa integrada por más de un establecimiento. Asimismo se identifica a la matriz y a las sucursales de estas empresas.

La variable Folio solo proviene del levantamiento censal y ningún otro directorio lo tiene, es por ello que al actualizar con información proveniente de algún directorio externo o de información de registros administrativos, no se cuenta con un identificador único que relacione la información original con los directorios externos, surgiendo la necesidad de realizar empates que permitan identificar las nuevas unidades económicas pertenecientes a las empresas ya conformadas para agregarlas, el surgimiento de nuevas empresa integradas por más de un establecimiento, la identificación de unidades económicas que mueren y es necesario identificarlas para darlas de baja de las empresas ya conformadas y actualizar la información en el directorio de empresas.

En el caso de las empresas conformadas por más de una unidad económica la problemática de la actualización y mantenimiento del directorio de empresa se agrava por las siguientes situaciones que surgen a lo largo del tiempo, dada la dinámica de vida de las unidades económicas:

- Empresas que por la actualización de los establecimientos que la conforman, el establecimiento matriz es dado de baja y no conocemos la matriz actual o hubo cambio de establecimiento matriz. Esta problemática se presenta cuando se da de baja el establecimiento matriz y en algún

directorio externo con posibles altas se identifica un establecimiento que por sus características pudiera ser el establecimiento matriz de esa empresa y no cuenta con un identificador que lo relacione sino que se identifica por encontrarse bajo la denominación de Razón Social y/o Nombre del Establecimiento.

- Unidades Económicas que son posibles altas y deba verificarse que efectivamente no se encuentren ya dentro del directorio de empresas como parte de una empresa formada por más de una unidad económica, pues duplicaría la información.

- Unidades Económicas que son bajas definitivas y ya no deban ser parte del directorio de empresas ya sea únicos o como parte de una empresa formada por más de una unidad económica.

- Unidades Económicas con cambios, movimientos o problemáticas que afecten al directorio de empresas, o a las empresas formadas por más de una unidad económica.

3.- Justificación.

Aunque los organismos internacionales recomiendan que los directorios de empresas se evalúen y revisen cada 6 meses o cada año, la necesidad de contar con un directorio de empresas actualizado y depurado en el momento en que se solicite un marco, para alguna encuesta o investigación con cobertura nacional o de algún o algunos sectores económicos en específico, nos lleva a la difícil tarea de realizar actualizaciones en el momento o bien a la utilización de un marco no actual para algún requerimiento.

Este procedimiento para generar un marco de empresas actualizado requiere de tiempo de análisis.

Primeramente es requerido determinar el estado de actualización de los establecimientos pertenecientes al directorio de unidades económicas, tomando en cuenta las características de la solicitud.

Se identifican altas, bajas, cambios, problemáticas y movimientos en las unidades económicas que pertenecen al directorio.

Posteriormente se incorporan los directorios externos en caso de que los hubiera. Finalmente se pasa a la actualización del directorio incorporando las actualizaciones antes mencionadas, reconformando las empresas demás de un establecimiento, localizando y formando nuevas empresas de más de un establecimiento en caso que las hubiera y dando de baja las unidades económicas de bajas definitivas.

La propuesta de utilizar un método estadístico basado en el concepto Record Linkage en un sistema automatizado pretende permitir identificar las unidades económicas que al relacionar el DNUE con los directorios externos carentes de

identificador único común que las relacione, incorporando las altas o nacimientos, bajas o muertes, cambios de giro, identificación de problemáticas o movimientos de establecimientos, previniendo la duplicidad y desactualización de ellas, este procedimiento propuesto intenta prevenir que al conformar las empresas no se duplique la información de variables económicas para empresas conformadas por más de un establecimiento y la identificación de unidades económicas que se encuentren bajo la misma denominación de Razón Social y/o Nombre de establecimiento, lo que pretende agilizar la formación o actualización de empresas conformadas por más de un establecimiento.

4.- Características específicas a observar.

La información que se va a integrar en la actualización a las empresas conformadas, será necesario que ya esté validada, sin omisión de información en ninguna de sus variables principales y además debe incluir la homogenización o armonización[5] de nomenclatura de variables propuesta por el área de censos económicos y bajo el cual se rigen las áreas que utilizan información relacionada a unidades económicas.

La conformación de empresas se define como el procedimiento en el cual se agrupan todos los establecimientos que como primera forma de conformación tomen el Folio proveniente del Censo Económico que identifica las empresas con más de un establecimiento.

En base a esta primera agrupación se adherirán a la empresa todas las unidades económicas que se encuentren bajo la misma denominación de Razón Social y/o Nombre del Establecimiento.

[5] Es definida como la actividad en que la nomenclatura de las variable de la información de las unidades económicas sea la misma y conocida por todo el personal involucrado con los concepto, de tal forma que todos hablen un mismo lenguaje al referirse a estas variables.

Se identificará el establecimiento que se denominará matriz que representará a la empresa conformada por más de un establecimiento.
Se realizará la sumatoria de las variables económicas de todos los establecimientos que conforman a la empresa, asignando esto valores a la matriz.
Finalmente se identificará el sector económico a la empresa que mejor la represente, esto es para el caso de que la empresa no tenga identificada el establecimiento matriz de la empresa o que por algún motivo no sea conocido.

En el caso de que si sea claramente identificado el establecimiento matriz en el directorio de empresas se tomará tal cual solo que las variables económicas estarán formadas por la acumulación de las variables de todas sus sucursales incluyendo la información del establecimiento matriz.

El sector de actividad económica se asignará en base al Sistema de Clasificación Industrial de América del Norte (SCIAN)[6]. En México, el clasificador oficial de actividades económicas, construido con Estados Unidos y Canadá, quienes tienen sus propias versiones nacionales de este clasificador. El objetivo del SCIAN MÉXICO es proporcionar un marco único, consistente y actualizado para la recopilación, análisis y presentación de estadísticas de tipo económico, que refleje la estructura de la economía mexicana. El SCIAN 2002 se divide en 20 sectores de actividad en el nivel más general, 95 subsectores, 309 ramas, 631 subramas y, en su nivel más detallado, en 1051 clases de actividad. El catálogo SCIAN, permite a los usuarios localizar la actividad económica con su código y descripción, así como conocer los productos que corresponden a cada actividad en su nivel más detallado.

[6] Sistema de Clasificación Industrial de América del Norte (SCIAN 2002)

5.- Objetivo General.

Generar un prototipo basado en la teoría Record Linkage para la integración de un directorio de empresas, en donde el uso de la teoría permita identificar las unidades económicas duplicadas dentro del directorio para su eliminación, el empate de DNUE contra los directorios externos para detectar las altas al directorio y finalmente identificar las unidades económicas bajo la misma denominación de Nombre del Establecimiento y/o Razón Social para la conformación de empresas de más de un establecimiento.

6.- Objetivos Específicos.

- Identificar las unidades económicas duplicadas con la finalidad de eliminarlas.
- Identificar dentro del directorio de unidades económicas, las unidades que se encuentran bajo la misma denominación de Razón Social y/o Nombre del Establecimiento para la conformación de empresas de más de un establecimiento.
- Permitir el empate entre dos directorios que no tengan un identificador único que los relacione uno a uno, detectando así las altas al directorio e incorporarlas.
- Proponer un algoritmo basado en la teoría Record Linkage para la identificación de unidades económicas bajo la misma denominación de Razón Social y/o Nombre del Establecimiento, que facilite la conformación de empresas de más de un establecimiento.
- Generar el prototipo de un sistema, que utilice la metodología Record Linkage para la conformación de empresas de más de un establecimiento bajo la misma denominación de Razón Social y/o Nombre del Establecimiento.

7.- Preguntas de Caso.

- ¿El prototipo propuesto al utilizar el concepto RECORD LINKAGE permite la identificación de unidades económicas bajo la misma denominación de Razón Social y/o Nombre del Establecimiento?
- ¿El prototipo propuesto permite el empate entre directorios que no cuenten con un identificador único que los relacione uno a uno pero que contengan variables con información similar entre ellos?
- ¿El concepto LINKAGE RECORD permite la identificación de unidades económicas duplicadas?
- ¿La utilización del concepto RECORD LINKAGE facilita la conformación de empresas de más de un establecimiento?
- ¿El prototipo generado permite la identificación de empresas bajo la misma denominación de Nombre del Establecimiento y/o Razón Social?

8.- Proposiciones.

- El prototipo identificará las unidades económicas bajo la misma denominación de Razón Social y/o Nombre del Establecimiento.
- El algoritmo RECORD LINKAGE propuesto permitirá el empate entre ficheros que no tengan un identificador único que permita relacionarlos unos a uno.
- El prototipo permitirá la identificación de unidades económicas duplicadas.
- El concepto Record Linkage facilitará la conformación de empresas formadas por más de una unidad económica.
- La utilización del método RECORD LINKAGE permitirá la incorporación y localización de la mayoría de los establecimientos que conforman una empresa independientemente de las fuentes que recabaron su información para mantener actuales los datos de las empresas.

9.- Marco Teórico.

9.1.- Directorio de Unidades Económicas con fines Estadísticos (Establecimientos y/o Empresas).

9.1.1.- Panorama Internacional.

En esta investigación se hace un pequeño recorrido a las recomendaciones y normativas propuestas o establecidas en la Comunidad Europea, la Comunidad Andina, La CEPAL e INEGI, México, en torno a los directorios de unidades económicas con fines estadísticos.

a.- Comunidad Europea (EUROSTAT) normatividad y recomendaciones.

La Unión Europea (UE)[7] está formada por 27 países soberanos independientes que se conocen como Estados miembros los cuales son: Alemania, Austria, Bélgica, Bulgaria, Chipre, la República Checa, Dinamarca, Estonia, Finlandia, Francia, Grecia, Hungría, Irlanda, Italia, Letonia, Lituania, Luxemburgo, Malta, Países Bajos, Polonia, Portugal, Rumania, Eslovaquia, Eslovenia, España, Suecia y el Reino Unido. Pero comparten su soberanía para ser más fuertes y tener una influencia mundial que ninguno de ellos podría ejercer individualmente.

Compartir la soberanía significa, en la práctica, que los Estados miembros delegan algunos de sus poderes decisorios en las instituciones comunes creadas por ellos para poder tomar democráticamente y a nivel europeo decisiones sobre asuntos específicos de interés conjunto.

En el proceso decisorio de la UE en general, y en el procedimiento de codecisión en particular intervienen tres instituciones principales:

- El Parlamento Europeo (PE), que representa a los ciudadanos de la UE y es elegido directamente por ellos;
- El Consejo de la Unión Europea), que representa a los Estados miembros;
- La Comisión Europea, que defiende los intereses de la Unión en su conjunto

[7] http://europa.eu/institutions/index_es.htm) Instituciones y otros órganos de la Unión Europea

EUROSTAT es la oficina estadística de la Comisión Europea, que produce datos sobre la Unión Europea y promueve la armonización de los métodos estadísticos de los estados miembros.
Dos de sus papeles particularmente importantes son la producción de datos macro-económicos que apoyan las decisiones del Banco Central Europeo en su política monetaria para el euro, y sus datos regionales y clasificación (NUTS[8]) que orientan las políticas estructurales de la Unión Europea.

EUROSTAT es una de las Direcciones Generales de la Unión Europea y tiene su sede en Luxemburgo. Está formado de un Director General y asistido por 7 directores, cada uno con su sector de actividad dentro del EUROSTAT:

- Recursos.
- Métodos estadísticos.
- Cuentas nacionales y europeas.
- Estadísticas económicas y regionales.
- Estadísticas agrícolas y ambientales; Cooperación estadística.
- Estadísticas sociales y Sociedad de la Información.
- Estadísticas de las empresas.

En 1993 los estados miembros de la unión Europea emprendieron programas de armonización y desarrollo de sus registros o directorios nacionales de empresas para fines estadísticos el cual está coordinado por Oficina Estadística de la Comunidad Europea (EUROSTAT), han definido con el transcurso de los años varios documentos en los cuales plasman sus normas y reglamentaciones para el registro de estos directorios y que son publicadas en el Diario Oficial de la Unión Europea (DOUE), ya que en ellos definen las características e importancia del registro y mantenimiento de directorios de empresas para fines estadísticos algunos de los cuales son:

[8] **NUTS** son las siglas en francés de la **Nomenclatura de las Unidades Territoriales Estadísticas** utilizadas por la Unión Europea con fines estadísticos. Fueron creadas por la Oficina Europea de Estadística EUROSTAT para dar uniformidad en las estadísticas regionales europeas y son utilizadas, entre otras cosas, para la redistribución regional de los fondos estructurales de la UE.

- ***Reglamento sobre directorios de empresas a fines estadísticas***[9]

En el Diario Oficial de la Unión Europea el 05 de Agosto de 1993 se publica el Reglamento (CEE) nº 2186/93)[9] del Consejo, de 22 de julio de 1993, relativo a la coordinación comunitaria del desarrollo de los registros de empresas utilizados con fines estadísticos. Estableció y definió algunas consideraciones para sustentar la propuesta de este reglamento y de las cuales solo se enlistan algunas de ellas de forma textual:

- *Considerando que el mercado único crea una mayor necesidad de mejorar la comparabilidad entre las distintas estadísticas elaboradas para responder a las necesidades comunitarias y que, para lograr dicha mejora, es necesario adoptar definiciones y descriptores comunes del campo de las empresas y de las demás unidades cuya actividad es objeto de estadísticas;*
- *Considerando que deben crearse y actualizarse registros de dichas unidades para poder recoger datos sobre ellas;*
- *Considerando que existe una necesidad creciente de información sobre la estructura de las empresas y que esta necesidad no puede ser cubierta por la situación actual de las estadísticas comunitarias;*
- *Considerando que los registros de empresas utilizables con fines estadísticos constituyen un instrumento necesario para el seguimiento de las modificaciones estructurales de la economía resultantes de operaciones como alianzas, asociaciones, compras, fusiones o absorciones.*
- *Considerando que actualmente no se dispone de determinados datos estadísticos, principalmente en sectores, como el de servicios, con abundancia de pequeñas y medianas empresas (PYME), porque no existe un registro de dichas empresas utilizado con fines estadísticos.*
- *Considerando que los registros utilizados con fines estadísticos son uno de los elementos básicos de los sistemas de información sobre empresas y que permiten organizar y coordinar encuestas estadísticas al proporcionar una base para el muestreo, posibilidades de extrapolación e instrumentos de seguimiento de lo exigido a las*

[9] DOUE. Diario Oficial de la Unión Europea, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993R2186:ES:HTML
Reglamento (CEE) nº 2186/93 del Consejo, de 22 de julio de 1993, relativo a la coordinación comunitaria del desarrollo de los registros de empresas utilizados con fines estadísticos. *Diario Oficial n° L 196 de 05/08/1993 p. 0001 – 0005, Edición especial en finés: Capítulo 13 Tomo 24 p. 0150, Edición especial sueca: Capítulo 13 Tomo 24 p. 0150.* «© Comunidades Europeas, http://eur-lex.europa.eu/». «Únicamente se consideran auténticos los textos legislativos de la Unión Europea publicados en la edición impresa del *Diario Oficial de la Unión Europea*».

empresas, especialmente a las contempladas en las Directivas 78/660/CEE (3) y 83/349/CEE (4);

- *Considerando que la puesta en marcha de un nuevo sistema de recogida para las estadísticas de intercambios de bienes y servicios entre los Estados miembros requiere la elaboración de un registro de los que tienen obligación de facilitar información y que es deseable que dicho registro de personas obligadas a facilitar información derive de un registro central de empresas utilizado con fines estadísticos;*
- *Considerando que el nivel de elaboración de los registros utilizados con fines estadísticos varía según los Estados miembros; que el desarrollo prolongado y costoso de dichos registros sólo se puede hacer en dos fases, y que la primera fase debe consistir en la armonización de sus unidades básicas en unos determinados plazos.*

En base a las consideraciones antes definidas se estableció este reglamento (CEE nº 2186/93) el cual en su momento fue constituido por 11 Artículos y su anexo.

En el artículo 2 especifica y hace referencia a un reglamento anterior sobre las definiciones el cual mencionaremos en el punto *"Reglamento de las unidades Estadísticas".*

Siguiendo con este reglamento definimos que a "Efecto de este reglamento tal y como se especifica a continuación."

Artículo 1.

Objetivos: Los Estados miembros crearán, con fines estadísticos, uno o varios registros armonizados con las definiciones y la cobertura contempladas en los artículos siguientes.

Artículo 2 Definiciones.

1. A efectos del presente Reglamento, se entenderá por:

a) « unidad jurídica »: la unidad jurídica a que se refiere el punto A.3 de la sección II del Anexo del Reglamento (CEE) no 696/93 (5);

b) « empresa »: la empresa a que se refiere la letra A de la sección III del Anexo del mismo Reglamento.

El vínculo entre la empresa y la unidad jurídica quedará caracterizado por las expresiones siguientes:

- la empresa va unida a una o varias unidades jurídicas, y la unidad jurídica responde de la empresa;

c) « unidad local »: la unidad local a que se refiere la letra F de la sección III del Anexo del Reglamento (CEE) no 696/93.

El vínculo entre la unidad local y la empresa quedará caracterizado por la expresión siguiente:

- la unidad local depende de una empresa.

2. El presente Reglamento sólo atañe a las unidades que ejercen total o parcialmente una actividad de producción.

Artículo 3 Cobertura.

1. Quedarán registradas, de conformidad con las definiciones contenidas en el artículo 2 y con las restricciones previstas en el presente artículo:

- *todas las empresas que ejerzan una actividad económica que suponga una contribución al producto interior bruto a precios de mercado (PIB),*
- *las unidades jurídicas que respondan de ellas,*
- *las unidades locales que dependan de ellas.*

No obstante, quedarán excluidos los hogares:

- *en la medida en que su producción vaya destinada al autoconsumo,*
- *en la medida en que produzcan servicios de arrendamiento de bienes inmobiliarios propios o alquilados, incluidos en el grupo 70/2 de la nomenclatura estadística de actividades económicas de la Comunidad Europea (NACE rev. 1), establecida por el Reglamento (CEE) no 3037/90 (6). Será optativa la inclusión:*
- *de empresas cuya actividad principal figure en las secciones A, B o L de la NACE rev. 1,*
- *de las unidades jurídicas que respondan de ellas,*
- *de las unidades locales que dependan de ellas.*

Se decidirá, según el procedimiento establecido en el artículo 9, en qué medida deben registrarse las pequeñas empresas que no ofrecen interés estadístico para los Estados miembros.

2. Para el registro de las empresas, unidades jurídicas y locales mencionadas en el apartado 1, se respetarán los plazos establecidos en el Anexo I.

3. El registro separado de las unidades jurídicas será optativo, siempre que toda la información relativa a estas unidades se incluya en los registros relativos a las empresas. Las modalidades de este registro se adoptarán según el procedimiento establecido en el artículo 9.

Artículo 5 Actualización.

1. Se actualizarán una vez al año, como mínimo:

a) las altas y bajas en el registro;

b) las variables indicadas en las letras b) y f) del punto 1 del Anexo II;

c) las variables indicadas en las letras b), c), d), e) y h) del punto 3 del Anexo II para las unidades que sean objeto de encuestas anuales, en la medida en que dichas variables figuren en las encuestas. En general, los datos obtenidos a partir de ficheros administrativos o de encuestas anuales se actualizarán cada año, y el resto cada cuatro años.

2. Al terminar el primer trimestre de cada año civil, los Estados miembros elaborarán una copia del registro, tal y como se encuentre, y la conservarán durante diez años, con el fin de poder realizar análisis.

En los últimos años estos reglamentos han ido actualizándose y derogando reglamentos, como en el año 2008.

Reglamento (CE) no 177/2008 del Parlamento Europeo y del Consejo de 20 de febrero de 2008 que establece un marco común para los registros de empresas utilizados con fines estadísticos y deroga el Reglamento (CEE) no 2186/93 del Consejo. Compuesto por 18 Artículos y su anexo.

De el cual se enlistan algunas de las consideraciones definidas:

- *El Reglamento (CEE) no 2186/93 del Consejo [2] estableció un marco común para la creación de registros de empresas utilizados con fines estadísticos, y armonizó las definiciones, las características, el ámbito de aplicación y los procedimientos de actualización. Para que los registros de empresas sigan desarrollándose en un marco armonizado, es necesario adoptar un nuevo Reglamento.*
- *El Reglamento (CEE) no 696/93 del Consejo, de 15 de marzo de 1993, relativo a las unidades estadísticas de observación y de análisis del sistema de producción en la Comunidad [3], contiene las definiciones de las unidades estadísticas que es preciso utilizar. El mercado interior requiere una mejor comparabilidad de las estadísticas a fin de responder a las necesidades comunitarias. Para conseguir dicha mejora, es preciso adoptar definiciones y descripciones comunes para las empresas y las demás unidades estadísticas pertinentes que es necesario abarcar.*
- *El Reglamento (CE, Euratom) no 58/97 del Consejo, de 20 de diciembre de 1996, relativo a las estadísticas estructurales de las empresas [4], y el Reglamento (CE) no 1165/98 del Consejo, de 19 de mayo de 1998, sobre las*

estadísticas coyunturales [5], establecieron un marco común para la recogida, compilación, transmisión y evaluación de las estadísticas comunitarias relativas a la estructura, la actividad, la competitividad y los resultados de las empresas de la Comunidad. Los registros de empresas utilizables con fines estadísticos constituyen un elemento fundamental de dicho marco común, que permite organizar y coordinar las encuestas estadísticas mediante una armonización del marco de muestreo.

- *Los registros de empresas constituyen uno de los elementos que permiten conciliar las exigencias antagónicas de obtener mayor información comparable sobre las empresas, por una parte, y de aligerar las obligaciones administrativas de estas, por otra, utilizando en particular los datos existentes en expedientes administrativos y jurídicos, especialmente en el caso de las microempresas y las pequeñas y medianas empresas, tal como se definen en la Recomendación 2003/361/CE de la Comisión.*
- *(7) Los registros de empresas con fines estadísticos constituyen también la fuente principal de información sobre la demografía de las empresas, al permitir un seguimiento de la creación y del cierre de empresas, así como de las modificaciones estructurales de la economía por concentración o desconcentración resultantes de operaciones como fusiones, absorciones, disoluciones, escisiones o reestructuraciones de empresas.*
- *(8) Los registros de empresas proporcionan la información básica requerida para satisfacer el gran interés político por el desarrollo rural, no solo en lo que se refiere a la agricultura sino también a su creciente combinación con otras actividades no incluidas en las estadísticas agrícolas basadas en la producción.*
- *(9) Las empresas públicas desempeñan un importante papel en la economía nacional de los Estados miembros. La Directiva 80/723/CEE de la Comisión, de 25 de junio de 1980, relativa a la transparencia de las relaciones financieras entre los Estados miembros y las empresas públicas [9], se aplica a determinados tipos de empresas públicas. En consecuencia, conviene que en los registros de empresas se identifiquen las empresas y sociedades públicas, lo que puede conseguirse mediante la clasificación del sector institucional.*
- *(10) Es necesario conocer las relaciones de control entre unidades jurídicas para la definición de los grupos de empresas, la correcta delimitación de cada empresa, la distinción de unidades complejas y de gran dimensión, y*

para el estudio del nivel de concentración de determinados mercados. La información sobre los grupos de empresas mejora la calidad de los registros de empresas y puede utilizarse para reducir el riesgo de divulgación de datos confidenciales. A menudo, determinados datos financieros son más significativos a nivel de grupo o subgrupo empresarial que a nivel de empresa individual, y pueden estar únicamente disponibles a nivel de grupo o subgrupo. Los registros de datos de grupos de empresas permiten, si es necesario, realizar encuestas directamente al grupo, en vez de a sus empresas constitutivas, lo que puede aligerar significativamente la carga de respuesta. Para registrar los grupos de empresas, es imprescindible una mayor armonización de los registros de empresas.

- *(11) La globalización creciente de la economía representa un desafío para la producción actual de diversas estadísticas. Gracias al registro de los datos de grupos multinacionales de empresas, los registros de empresas constituyen una herramienta básica para mejorar muchas estadísticas sobre globalización: comercio internacional de bienes y servicios, balanza de pagos, inversiones extranjeras directas, filiales extranjeras, investigación, desarrollo e innovación, y mercado laboral internacional. La mayoría de dichas estadísticas abarca el conjunto de la economía, lo que exige que los registros de empresas cubran todos los sectores de la misma.*
- *(13) Para garantizar el cumplimiento de las obligaciones establecidas en el presente Reglamento, los institutos nacionales responsables de la recogida de datos en los Estados miembros pueden necesitar acceder a fuentes de datos administrativos, como por ejemplo registros de las autoridades fiscales y de la seguridad social, bancos centrales, otras instituciones públicas y otras bases de datos que contengan información sobre operaciones y posiciones transfronterizas, si dichos datos son necesarios para elaborar*
- *(17) Por consiguiente, procede derogar el Reglamento (CEE) no 2186/93.*

- **Reglamento de las unidades Estadísticas.**[10]

Reglamento (CEE) No 696/93 del consejo de 15 de marzo de 1993 relativo a las unidades estadísticas de observación y de análisis del sistema de producción en la Comunidad. En el que se definieron 9 Artículos y sus respectivos anexos en donde se especifican:

Las unidades estadísticas del sistema de producción en la Comunidad Europea

Sección I.- La lista de unidades estadísticas del sistema de producción.

Sección II.- Criterios utilizados. Las unidades estadísticas que aparecen en el presente Reglamento se definen a partir de tres criterios cuya importancia relativa varía según las unidades. Criterios jurídico, contable o de organización.

Sección III.- Definiciones de las unidades y notas explicativas específicas de cada unidad.

Sección IV.- Notas explicativas complementarias.

- **Manual de Recomendaciones Sobre el Registro de Empresas.**

(Situación: Primera revisión - marzo de 2003).[11]

Ámbito de aplicación de las recomendaciones

Marco general.

1.- Los Estados miembros de la Unión Europea han emprendido un programa de armonización y desarrollo de sus registros nacionales de empresas para uso estadístico. Dicho programa está coordinado por EUROSTAT. En las reuniones anuales del Grupo de Trabajo sobre Registros de Empresas se toman las decisiones relativas al programa y se informa sobre su progreso. La herramienta más importante para evaluar los progresos alcanzados es el cuestionario anual, que administra Eurostat. Por otra parte, se mantiene un contacto periódico entre los Estados miembros y Eurostat por medios menos formales, como el correo electrónico y el sitio de la red de registros de empresas en Internet.

2.- En general, este programa está abierto a otros países europeos, en particular a los de la AELC[12] y a los países candidatos, la mayor parte de los cuales participan en las reuniones

[10] Reglamento (CEE) nº 696/93 del Consejo, de 15 de marzo de 1993, relativo a las unidades estadísticas de observación y de análisis del sistema de producción en la Comunidad, *Diario Oficial n° L 076 de 30/03/1993 p. 0001 – 0011, Edición especial en finés...: Capítulo 13 Tomo 24 p. 0007, Edición especial sueca...: Capítulo 13 Tomo 24 p. 0007*. . «© Comunidades Europeas, http://eur-lex.europa.eu/». «Únicamente se consideran auténticos los textos legislativos de la Unión Europea publicados en la edición impresa del *Diario Oficial de la Unión Europea*».

[11] Manual de recomendación sobre directorios de empresas de la Unión Europea.

[12] La Asociación Europea de Libre Cambio (AELC) o Asociación Europea de Libre Comercio (también conocida por sus siglas en inglés EFTA - European Free Trade Area) es un bloque comercial creado el 4 de enero de 1960 por la Convención

y los debates. Además se desarrolla en estrecha coordinación con la Comisión Económica para Europa de las Naciones Unidas (CEPE/ONU), con la que se reúne periódicamente. Las principales herramientas de apoyo de que dispone son:

– El Reglamento (CEE) nº 2186/93 del Consejo, de 22 de julio de 1993, relativo a la coordinación comunitaria del desarrollo de los registros de empresas utilizados con fines estadísticos (DO nº L 196, 5.8.93), que constituye su base jurídica.

– El Manual de recomendaciones, que, pese a carecer de base jurídica, aporta las directrices que sirven para interpretar el Reglamento, así como información útil para orientar el desarrollo de los registros de empresas.

Reglamento.

3.- El 22 de julio de 1993, el Consejo de Ministros de la Unión Europea adoptó el Reglamento, que entró en vigor el 25 de agosto del mismo año y que forma parte de una serie de Reglamentos cuyo objetivo es armonizar la infraestructura necesaria para las estadísticas europeas de empresas y que incluye:

– El Reglamento (CEE) nº 3037/90 del Consejo, de 9 de octubre de 1990, relativo a la nomenclatura estadística de actividades económicas en la Comunidad Europea, que aportó la base jurídica de la clasificación de la NACE. Posteriormente, este Reglamento ha sido modificado por el Reglamento (CEE) de la Comisión nº 761/93 de 24 de marzo de 1993 y por el Reglamento (CE) de la Comisión nº 29/2002 de 19 de diciembre de 2001. Este último introdujo la versión más reciente de la NACE, conocida como NACE Rev. 1.1.

– El Reglamento (CEE) nº 696/93 del Consejo, de 15 de marzo de 1993, relativo a las unidades estadísticas de observación y de análisis del sistema de producción en la Comunidad.

4.- El Reglamento sobre los registros de empresas constituía un compromiso entre lo que en los años noventa se podía considerar deseable y lo razonablemente posible, dado que la

de Estocolmo como alternativa a la Comunidad Económica Europea (1957) y por los países Austria, Dinamarca, Gran Bretaña, Noruega, Portugal, Suecia y Suiza. Entró en vigor en junio de 1960. En 1961 entró Finlandia, en 1970 Islandia y en 1991 Liechtenstein.

situación inicial de los registros difería considerablemente según los países: algunos de ellos tenían que desarrollarlos o incluso crearlos, mientras que otros sólo tenían que adaptarlos de modo que satisficiesen los requisitos del Reglamento.

Objetivos del manual.

Explicar el Reglamento

5.- El Reglamento presenta una selección de normas concretas para la armonización de los registros nacionales, pero la lógica de esta última no siempre se trasluce en su redacción final, que es el resultado de complejas negociaciones. El manual intenta explicar el razonamiento que justifica las disposiciones del Reglamento, así como aportar la información adicional necesaria para que la interpretación de éste sea correcta y coherente en todos los países.

Superar el Reglamento y guiar la evolución futura

6.- El manual va más allá de las disposiciones del Reglamento, por las razones que siguen:

- *Durante la elaboración del Reglamento, una de las principales limitaciones venía dada por el poco tiempo de que disponían todos los Estados miembros, por lo que más que la situación ideal se reflejó una de compromiso. Dado que el manual carece de valor jurídico, se puede utilizar como herramienta para buscar soluciones ideales mediante la identificación y la recomendación de las mejores prácticas.*
- *Aunque la aplicación de las disposiciones del Reglamento garantiza cierto nivel de armonización entre los registros de los diferentes Estados miembros, esto no basta para que dichos registros sean plenamente operativos. Para ello es preciso que los Estados miembros añadan ciertos elementos que tengan en cuenta las peculiaridades nacionales, tales como las fuentes administrativas utilizadas y las necesidades de los distintos usuarios de los registros. La libertad de decidir cómo elaborar y mantener el registro es coherente con el principio de subsidiariedad, pero acrecienta la importancia de contar con directrices documentadas e información sobre las experiencias de otros países.*

Estructura del manual.

7.- El manual se presenta en tres partes, que se ocupan de:

- *Los fundamentos: Objetivos, unidades, contenido y acceso (capítulos 1 a10).*
- *La demografía de la unidad: Cambios y continuidad (capítulos 11 a 16).*
- *El contenido: Actualización y desarrollo (capítulos 17 en adelante).*

8.- La estructura del manual no se ha fijado de un modo permanente, pues admite nuevos capítulos que reflejen nuevas evoluciones, así como la revisión de los ya existentes cuando sea necesario.

9.- La última versión del manual está destinada principalmente a una difusión electrónica, a través de Internet. Sin embargo, si se desea se puede imprimir por capítulos.
Por lo tanto, ha de ser posible leer cada capítulo como algo independiente que, sin embargo, formará parte de un conjunto coherente. Debido a ello, no siempre es posible entender plenamente un tema determinado sin leer los otros capítulos relacionados con él.

10.- En cada capítulo se indican claramente la fecha y su situación, así como si se trata de un borrador o bien de un texto adoptado.

Alcance de las recomendaciones.

11.- Todos los Estados miembros de la Comunidad poseen registros estadísticos. De conformidad con el artículo 1 del Reglamento sobre los registros de empresas, «los Estados miembros crearán, con fines estadísticos, uno o varios registros armonizados con las definiciones y la cobertura contempladas en los artículos siguientes». En este manual, la expresión «registros de empresas» se referirá a listas de empresas y otras unidades, tal como se establecen en el Reglamento sobre los registros de empresas o elaboradas voluntariamente, cuyas actividades contribuyan al producto interior bruto del Estado miembro de que se trate. Dichas unidades se pueden definir como las que ejercen un control sobre el uso de los recursos, incluidos la tierra, el trabajo, el capital, los bienes y los servicios, con el fin de producir bienes y servicios para consumo propio o por otras unidades.

12.- El alcance y la naturaleza de los registros nacionales de empresas están determinados por factores propios del país. El primer factor es, naturalmente, el objetivo que se haya asignado al registro, por ejemplo, como herramienta de realización de encuestas, o como fuente de estadísticas por derecho propio. Un segundo factor importante es que las disposiciones legislativas determinan en gran parte tanto los datos que pueden incluirse en un registro estadístico como las condiciones en que pueden almacenarse. En tercer lugar, la información que las empresas necesitan por motivos propios o para cumplir requisitos administrativos condiciona en gran medida las recopilaciones que un instituto estadístico puede realizar para su registro, ya que a éste le resulta prácticamente imposible obligar a las empresas a suministrar unos datos que ellas mismas no necesitan. Por último, los costes de

creación y gestión de un registro son muy elevados y los recursos que a ello se dedican varían de un Estado miembro a otro.

13.- Si actualmente el alcance y naturaleza de los registros estadísticos varían mucho de un Estado a otro, hay que añadir además que, por lo general, no tienen directamente en cuenta las precisiones estadísticas de la Comunidad como tal. Por otra parte, éstas no dejarán de aumentar, dadas las necesidades crecientes de datos oportunos y precisos que presentarán la administración de la unión monetaria y el mercado único.

14.- Las recomendaciones formuladas en el presente documento en cuanto a la armonización de los principios y del contenido de los registros estadísticos de los Estados miembros no recogen simplemente los factores comunes a los registros existentes en los Estados miembros, sino que reflejan prácticas actuales que los estadísticos consideran útiles, mejoras de los registros que consideran posibles y necesidades futuras, en la medida en que se pueden prever. Las recomendaciones tienen totalmente en cuenta la necesidad de coherencia a la hora de usar las clasificaciones en los registros armonizados y, particularmente, la necesidad de compatibilidad con la NACE. Asimismo, tienen en cuenta la necesidad de hallar un equilibrio entre lo que es deseable y lo que es posible, a la vista de los costes que hay que afrontar y de los datos que las entidades pueden estar razonablemente dispuestas a suministrar.

15.- Cuando estas recomendaciones se apliquen, los registros se crearán y gestionarán de forma coherente en todos los Estados miembros. Por lo tanto, aumentará la comparabilidad y, en muchos casos, la calidad de las encuestas estadísticas basadas en ellos. También ayudarán a desarrollar nuevos usos de los registros, por ejemplo como fuente directa de estadísticas sobre la demografía de las empresas.

Relación con el Reglamento sobre los registros de empresas.

16.- Este capítulo es una introducción del Manual de recomendaciones, su marco, sus objetivos y su ámbito de aplicación. Por lo tanto, no se considera una interpretación del Reglamento.

- **Código de buenas prácticas de las estadísticas europeas.**[13] [14]

Definiciones: A efectos del Código de buenas prácticas de las estadísticas europeas:
Por estadísticas europeas se entenderán las estadísticas comunitarias, tal como se definen en el Reglamento (CE) nº 322/97 del Consejo, de 17 de febrero de 1997, sobre la estadística comunitaria, elaboradas y definidas por las autoridades nacionales de estadística y la autoridad estadística de la Comunidad (Eurostat) de conformidad con el artículo 285, apartado 2, del Tratado.

Por autoridad estadística se entenderá, a escala nacional, el Instituto Nacional de Estadística (INE) y los demás organismos estadísticos responsables de elaborar y difundir las estadísticas europeas y, a escala europea, Eurostat.
Por Sistema estadístico europeo, en lo sucesivo denominado SEE, se entenderá la asociación formada por Eurostat, los institutos nacionales de estadística y los demás organismos estadísticos responsables, en cada Estado miembro, de elaborar y difundir las estadísticas europeas.

El Código de buenas prácticas tiene el doble objetivo de:

- *aumentar la confianza en la independencia, la integridad y la responsabilidad de las autoridades nacionales de estadística y de Eurostat, así como en la credibilidad y la calidad de las estadísticas que elaboran y difunden.*
- *promover la aplicación de los mejores principios, métodos y prácticas entre todos aquellos que elaboran las estadísticas europeas a fin de aumentar su calidad.*
- *El documento está estructurado entorno a quince principios, agrupados en tres secciones:*
- *Entorno institucional.*
- *Principios: Independencia profesional, mandato de recogida de datos, adecuación de los recursos, compromiso de calidad, confidencialidad estadística, imparcialidad y objetividad.*
- *Procesos estadísticos.*
- *Principios: Metodología sólida, procedimientos estadísticos adecuados, una carga para los encuestados que no sea excesiva, relación coste-eficacia.*
- *Producción estadística.*
- *Principios: Pertinencia, precisión y fiabilidad, oportunidad y puntualidad, coherencia y comparabilidad, accesibilidad y claridad.*

[13] Código de Buenas Practicas de las Estadísticas Europeas
http://secgen.comunidadandina.org/andestad/adm/upload/file/codigo.pdf
[14] COMISIÓN DE LAS COMUNIDADES EUROPEAS, Bruselas, 25.5.2005, COM(2005) 217 final

El Código está destinado, para su aplicación, a:

- *las autoridades encargadas de la gobernanza (es decir, gobiernos, ministerios,*
- *la Comisión y el Consejo), para proporcionarles orientaciones destinadas a asegurarse de que sus servicios estadísticos están organizados profesionalmente y dotados de recursos para elaborar estadísticas europeas de tal manera que la independencia, la integridad y la responsabilidad queden garantizadas;*
- *las autoridades estadísticas y su personal, para proporcionar una referencia de principios, valores y buenas prácticas en materia de estadística que deberían ayudarlas a elaborar y difundir estadísticas europeas de gran calidad y armonizadas.*

Asimismo, está destinado a informar a:

- *los usuarios, para poner de manifiesto que las autoridades europeas y nacionales en materia de estadística son imparciales y que las estadísticas que elaboran y difunden son objetivas y fidedignas;*
- *los proveedores de datos, para demostrar que la confidencialidad de la información que proporcionan está protegida y que no se les pedirá demasiado.*

El Código de conducta se basa en quince Principios. Las autoridades encargadas de la gobernanza y las autoridades estadísticas de la Unión Europea se comprometen a respetar los principios establecidos en el presente Código y a revisar su aplicación periódicamente utilizando los Indicadores de buenas prácticas correspondientes a cada uno de los quince principios, que deberán utilizarse como referencia. El Comité del programa estadístico establecido por la Decisión 89/382/CEE del Consejo, de 19 de junio de 1989, realizará periódicamente una revisión inter pares para hacer un seguimiento de la aplicación del presente Código.

- *Entorno institucional. (Primeros 6 principios)*

Los factores institucionales y organizativos tienen una influencia considerable en la eficacia y la credibilidad de una autoridad estadística que elabora y difunde estadísticas europeas. Las cuestiones pertinentes son la independencia profesional, el mandato de recogida de datos, la adecuación de los recursos, el compromiso de calidad, la confidencialidad estadística, la imparcialidad y la objetividad.

Principio 1: Independencia profesional – La independencia profesional de las autoridades estadísticas de otros departamentos y organismos políticos, reguladores o administrativos, así como de los operadores del sector privado, garantiza la credibilidad de las estadísticas europeas.

Indicadores.

– *En la legislación se especifica la independencia de la autoridad estadística de las interferencias políticas y de otras interferencias externas a la hora de elaborar y difundir estadísticas oficiales.*

– *El director de la autoridad estadística tiene un nivel jerárquico lo suficientemente elevado como para garantizar un acceso de alto nivel a las autoridades políticas y a los organismos públicos de carácter administrativo. Debe ser una persona de una gran capacidad profesional.*

– *El director de la autoridad estadística y, cuando proceda, los jefes de sus organismos estadísticos tienen la responsabilidad de garantizar que las estadísticas europeas se elaboran y difunden de forma independiente.*

– *El director de la autoridad estadística y, cuando proceda, los jefes de sus organismos estadísticos son los únicos responsables para decidir sobre los métodos, las normas y los procedimientos estadísticos, así como sobre el contenido y el calendario de las comunicaciones estadísticas.*

– *Se publican los programas de trabajo estadístico y se describen los progresos realizados en informes periódicos.*

– *Las comunicaciones estadísticas se distinguen claramente y se emiten al margen de las declaraciones políticas.*

– *Cuando procede, la autoridad estadística realiza comentarios públicos sobre cuestiones estadísticas, en los que incluye críticas y usos inadecuados de las estadísticas oficiales.*

Principio 2: Mandato de recogida de datos - Las autoridades estadísticas deben tener un mandato jurídico claro para recoger información destinada a la elaboración de estadísticas europeas. A petición de las autoridades estadísticas, se podrá obligar por ley a las administraciones, las empresas, los hogares y el público en general a que permitan el acceso a los datos destinados a la elaboración de estadísticas europeas o a que presenten dichos datos.

Indicadores

– *En la legislación se especifica el mandato de recoger información destinada a la elaboración y la difusión de estadísticas oficiales.*

– *La legislación nacional permite a la autoridad estadística la utilización de expedientes administrativos a efectos estadísticos.*

– *Sobre la base de un acto jurídico, la autoridad estadística puede obligar a responder encuestas estadísticas.*

Principio 3: Adecuación de los recursos – Los recursos a disposición de las autoridades estadísticas deben ser suficientes para cumplir los requisitos de las estadísticas europeas.

Indicadores

- *– Se dispone de recursos humanos, financieros e informáticos adecuados tanto en tamaño como en calidad para cumplir las necesidades actuales de las estadísticas europeas.*
- *– El alcance, el detalle y el coste de las estadísticas europeas son proporcionados respecto a las necesidades.*
- *– Existen procedimientos para evaluar y justificar las solicitudes de nuevas estadísticas europeas en relación con su coste.*
- *– Existen procedimientos para evaluar la necesidad continua de todas las estadísticas europeas, para determinar si alguna de ellas puede realizarse de forma discontinua o reducirse y, así, poder liberar recursos.*

Principio 4: Compromiso de calidad – Todos los miembros del Sistema estadístico europeo (SEE) se comprometen a trabajar y cooperar conforme a los principios establecidos en la Declaración sobre la calidad del Sistema estadístico europeo.

Indicadores.

- *– La calidad del producto se controla periódicamente conforme a los componentes de calidad del SEE.*
- *– Existen procesos para controlar calidad de la recogida, el tratamiento y la difusión de estadísticas.*
- *– Existen procesos para abordar consideraciones de calidad, en los que constan compromisos en este ámbito, y para orientar la planificación de las encuestas actuales y futuras.*
- *– Las orientaciones de calidad están documentadas y el personal tiene una formación adecuada. Dichas orientaciones se expresan por escrito y se ponen a disposición del público.*
- *– Existe una revisión periódica y profunda de la producción estadística clave utilizando expertos externos cuando proceda.*

Principio 5: Confidencialidad estadística – Deben garantizarse absolutamente la privacidad de los proveedores de datos (hogares, empresas, administraciones y otros encuestados), la confidencialidad de la información que proporcionan y su uso exclusivo a efectos estadísticos.

Indicadores.

– *En la legislación se garantiza la confidencialidad estadística.*

– *El personal de la autoridad estadística firma un compromiso de confidencialidad jurídico cuando es nombrado.*

– *Se establecen sanciones importantes por cualquier incumplimiento premeditado de la confidencialidad estadística.*

– *Se proporcionan instrucciones y orientaciones sobre la protección de la confidencialidad estadística en los procesos de elaboración y difusión. Dichas orientaciones se expresan por escrito y se ponen a disposición del público.*

– *Existen disposiciones físicas y tecnológicas para proteger la seguridad y la integridad de las bases de datos estadísticas.*

– *Se aplican protocolos estrictos a los usuarios externos que acceden a microdatos a efectos de investigación.*

Principio 6: Imparcialidad y objetividad **–** *Las autoridades estadísticas deben elaborar y difundir estadísticas europeas respetando la independencia científica y hacerlo de forma objetiva, profesional y transparente, de modo que se trate a todos los usuarios por igual.*

Indicadores.

– *Las estadísticas se recopilan sobre una base objetiva determinada por consideraciones estadísticas.*

– *La elección de las fuentes y las técnicas depende de consideraciones estadísticas.*

– *Los errores descubiertos en las estadísticas publicadas se corrigen y se dan a conocer lo antes posible.*

– *La información sobre los métodos y los procedimientos utilizados por la autoridad estadística están a disposición del público.*

– *Se anuncian previamente la fecha y la hora de comunicación de las estadísticas.*

– *Todos los usuarios tienen acceso al mismo tiempo a las comunicaciones estadísticas, y se restringe, se controla y se hace pública toda comunicación previa privilegiada a cualquier usuario externo. En caso de que se produzcan filtraciones, deberían revisarse los acuerdos de comunicación previa para garantizar la imparcialidad.*

– *Las comunicaciones y declaraciones estadísticas realizadas en ruedas de prensa son objetivas e imparciales.*

- *Procesos estadísticos. (Principios 7 al 10).*

Las normas, orientaciones y buenas prácticas, tanto europeas como internacionales, deben cumplirse plenamente en los procesos utilizados por las autoridades estadísticas para

organizar, recoger, elaborar y difundir las estadísticas oficiales. La credibilidad de las estadísticas se ve reforzada por una reputación de buena gestión y eficacia. Los aspectos pertinentes son una metodología sólida, unos procedimientos estadísticos adecuados, una carga para los encuestados que no sea excesiva y la relación coste-eficacia.

Principio 7: Metodología sólida – Unas estadísticas de calidad deben apoyarse en una metodología sólida, lo cual exige herramientas, procedimientos y conocimientos especializados adecuados.

Indicadores.

– El marco metodológico general de la autoridad estadística sigue normas, orientaciones y buenas prácticas tanto europeas como internacionales.

– Existen procedimientos para garantizar que se aplican coherentemente conceptos, definiciones y clasificaciones estándar en toda la autoridad estadística.

– El registro de empresas y el marco de las encuestas de población se evalúan periódicamente y, en caso necesario, se ajustan para garantizar una alta calidad.

– Existe una concordancia detallada entre las clasificaciones nacionales y los sistemas de sectorización y los sistemas europeos correspondientes.

– Se contratan titulados en las disciplinas académicas pertinentes.

– El personal asiste a cursos de formación y conferencias internacionales pertinentes, y se relaciona con colegas especialistas en estadística a nivel internacional para aprender de los mejores profesionales y aumentar sus conocimientos especializados.

– Se establece una cooperación con la comunidad científica para mejorar la metodología, se evalúa, mediante revisiones externas, la calidad y la eficacia de los métodos aplicados y se promueve la adopción de herramientas mejores cuando ello es viable.

Principio 8: Procedimientos estadísticos adecuados – Unas estadísticas de calidad deben apoyarse en unos procedimientos estadísticos adecuados, aplicados desde la recogida de los datos hasta la validación de los mismos.

Indicadores.

– Cuando las estadísticas europeas se basan en datos administrativos, las definiciones y los conceptos utilizados a efectos administrativos deben aproximarse bastante a los requeridos a efectos estadísticos.

– En el caso de las encuestas estadísticas, se prueban sistemáticamente los cuestionarios antes de la recogida de datos.

– El diseño de las encuestas, y la selección y ponderación de las muestras están bien fundamentados y se revisan o actualizan conforme a lo dispuesto.

– El trabajo de campo, la introducción de los datos y la codificación se controlan y revisan de forma rutinaria conforme a lo dispuesto.

– Se utilizan sistemas informáticos de edición y de imputación y se revisan o actualizan periódicamente conforme a lo dispuesto.

– Las revisiones siguen procedimientos normalizados, consolidados y transparentes.

Principio 9: Una carga para los encuestados que no sea excesiva **–** *La carga que supone la notificación debería ser proporcionada respecto a las necesidades de los usuarios y no ser excesiva para los encuestados. La autoridad estadística controla la carga que supone responder a la encuesta y fija objetivos para reducirla progresivamente.*

Indicadores.

– El alcance y el detalle de las demandas de estadísticas europeas se limita a lo estrictamente necesario.

– La carga que supone la notificación se reparte lo más ampliamente posible entre la población sobre la que se efectúa la encuesta mediante técnicas de muestreo apropiadas.

– En la medida de lo posible, se puede acceder fácilmente a la información que se solicita de las empresas a partir de sus cuentas y, cuando es posible, se utilizan medios electrónicos para facilitar su transmisión.

– Se aceptan las estimaciones y aproximaciones más fiables cuando no se dispone inmediatamente de la información exacta.

– Cuando es posible se utilizan fuentes administrativas para evitar que se dupliquen las solicitudes de información.

– Está generalizada la puesta en común de datos entre las autoridades estadísticas a fin de evitar la multiplicación de las encuestas.

Principio 10: Relación coste-eficacia – Los recursos deben utilizarse eficazmente.

Indicadores.

– Se controla la utilización de los recursos de la autoridad estadística a través de medidas internas y externas independientes.

– Las operaciones administrativas rutinarias (por ejemplo, toma, codificación y validación de los datos) están automatizadas en la mayor medida posible.

– *Se está optimizando el potencial de productividad de la tecnología de la información y la comunicación a efectos de recogida, tratamiento y difusión de los datos.*

– *Se están realizando esfuerzos proactivos para mejorar el potencial estadístico de los registros administrativos y evitar encuestas directas costosas.*

- *Producción estadística. (Principios del 11 al 15)*

Las estadísticas disponibles deben satisfacer las necesidades de los usuarios. Las estadísticas cumplen las normas de calidad europeas y responden a las necesidades de las instituciones europeas, los gobiernos, los organismos de investigación, las empresas y el público en general. Las cuestiones importantes atañen a la medida en que las estadísticas son pertinentes, precisas y fiables, oportunas, coherentes, comparables entre regiones y países, y de fácil acceso para los usuarios.

Principio 11: Pertinencia - Las estadísticas europeas deben satisfacer las necesidades de los usuarios.
Indicadores.

– *Existen procesos para consultar a los usuarios, controlar la pertinencia y la utilidad práctica de las estadísticas existentes por lo que se refiere a la satisfacción de las necesidades, así como para asesorar sobre las nuevas necesidades y prioridades.*

– *Se satisfacen las necesidades prioritarias y se reflejan en el programa de trabajo.*

– *Se realizan periódicamente encuestas para conocer el grado de satisfacción de los usuarios.*

Principio 12: Precisión y fiabilidad - Las estadísticas europeas deben reflejar la realidad de forma precisa y fidedigna.
Indicadores.

– *Se evalúan y validan los datos originales, los resultados intermedios y la producción estadística.*

– *Se miden y se documentan sistemáticamente los errores de muestreo y los que no son de muestreo con arreglo al marco de los componentes de calidad del SEE.*

– *Se realizan de forma rutinaria y se utilizan internamente estudios y análisis de revisiones para moldear los procesos estadísticos.*

Principio 13: Oportunidad y puntualidad - Las estadísticas europeas deben difundirse oportuna y puntualmente.

Indicadores.

- *– La oportunidad es conforme a las normas más estrictas de difusión a escala europea e internacional.*
- *– Se establece una hora determinada del día para la comunicación de estadísticas europeas.*
- *– Para establecer la periodicidad de las estadísticas europeas se tienen en cuenta los requisitos de los usuarios en la medida de lo posible.*
- *– En caso de que la comunicación no vaya a producirse a la hora establecida, se notifica por adelantado, se dan explicaciones y se fija un nuevo plazo de comunicación.*
- *– Cuando se considere conveniente, pueden difundirse resultados preliminares de una calidad global aceptable.*

Principio 14: Coherencia y comparabilidad – Las estadísticas europeas deberían ser coherentes a nivel interno, a lo largo del tiempo y comparables entre regiones y países; debería ser posible combinar y hacer un uso conjunto de los datos relacionados a partir de fuentes distintas.

Indicadores.

- *– Las estadísticas son coherentes a nivel interno (por ejemplo, se observan las identidades aritméticas y contables).*
- *– Las estadísticas son coherentes o conciliables durante un período razonable.*
- *– Las estadísticas se recopilan sobre la base de normas comunes respecto al alcance, las definiciones, las unidades y las clasificaciones en las distintas encuestas y fuentes.*
- *– Se comparan y concilian las estadísticas de las distintas encuestas y fuentes.*
- *– Se garantiza la comparabilidad transnacional de los datos mediante intercambios periódicos entre el Sistema estadístico europeo y otros sistemas estadísticos; se efectúan estudios metodológicos en estrecha colaboración entre los Estados miembros y Eurostat.*

Principio 15: Accesibilidad y claridad – Las estadísticas europeas deberían presentarse de forma clara y comprensible, difundirse de forma adecuada y conveniente y estar disponibles, asimismo se debería permitir el acceso a las mismas de forma imparcial, con metadatos y orientación de apoyo.

Indicadores

– Las estadísticas se presentan de tal forma que facilitan una interpretación adecuada y comparaciones significativas.

– Los servicios de difusión utilizan una tecnología moderna de información y comunicación y, si procede, una copia impresa tradicional.

– Cuando sea posible se suministran análisis a medida y se hacen públicos.

b.- Comunidad Andina (CAN) normatividad y recomendaciones.

La Comunidad Andina (CAN) es una comunidad de cuatro países que decidieron unirse voluntariamente con el objetivo común: alcanzar un desarrollo integral, más equilibrado y autónomo, mediante la integración andina, suramericana y latinoamericana. Los países que la forman son: Bolivia, Colombia, Ecuador y Perú, cuenta con 5 países asociados Chile, Argentina, Brasil, Paraguay y Uruguay además de dos países observadores México y Panamá.

Los órganos e instituciones que integran la comunidad Andina son:

- Consejo Presidencial Andino.
- Consejo Andino de Ministros de Relaciones Exteriores.
- Comisión de la Comunidad Andina.
- Secretaría General de la Comunidad Andina.
- Tribunal de Justicia de la Comunidad Andina.
- Parlamento Andino.
- Corporación Andina de Fomento (CAF).
- Fondo Latinoamericano de Reservas (FLAR).
- Consejo Consultivo Empresarial Andino.
- Consejo Consultivo Laboral Andino.
- Consejo Consultivo de Pueblas Indígenas.
- Organismo Andino de Salud, Hipólito Unanue
- Convenio Simón Rodríguez.
- Universidad Andina Simón Bolívar.

Legislación Estadística Comunitaria SG/de 279 30 de septiembre de 2009 E.3.1.[14]

La Comunidad Andina en su sección de Documentos y Publicaciones oficiales publica "La Legislación Estadística Comunitaria SG/de 279[15] 30 de septiembre de 2009 E.3.1" vigente en donde en su parte introductoria declara:

"La Comunidad Andina (CAN), está constituida por los órganos e instituciones del Sistema Andino de Integración (SAI) y por cuatro países que decidieron unirse voluntariamente en un proyecto político con el objetivo de alcanzar un desarrollo equilibrado y autónomo, mediante la integración andina, suramericana y latinoamericana.

[15] LEGISLACION ESTADITICA COMUNITARIA, COMUNIDAD ANDINA, SECRETARIA GENERAL, http://intranet.comunidadandina.org/IDocumentos/c_Newdocs.asp?GruDoc=13

Para el diseño, aplicación, seguimiento y evaluación de gran parte de sus políticas, la CAN debe poder acceder en el momento oportuno a datos estadísticos comparables entre los Países Miembros, actualizados, confiables, pertinentes y producidos con máxima eficacia.

La especificación en la elaboración de las estadísticas comunitarias, basada en los sistemas estadísticos nacionales, requiere de los instrumentos jurídicos necesarios para establecer dichas estadísticas comunitarias. Esta normativa está enmarcada dentro del Ordenamiento Jurídico Comunitario."

En un siguiente apartado define:

Ordenamiento Jurídico de la CAN

El Ordenamiento Jurídico de la Comunidad Andina es el conjunto sistemático de normas jurídicas, las cuales brindan los principios para hacer posible la integración y se aplican en el territorio de los Países Miembros a todos sus habitantes. Dentro de este ordenamiento se encuentran el Acuerdo de Cartagena, las Decisiones, las Resoluciones y algunos convenios de complementación industrial y otros adoptados en el marco de la Comunidad Andina, que consideren necesarios los países para el proceso de integración.

Por otro lado, la Decisión 472 en su artículo cuatro refiere a los países para que adopten las medidas que sean necesarias para asegurar el cumplimiento de las normas y los comprometen a no adoptar ninguna nacional que sea contraria a dichas normas.

¿Qué es una DECISIÓN?

Una Decisión es un instrumento jurídico adoptado de acuerdo a lo establecido por el Tratado del Tribunal de Justicia de la Comunidad Andina, por el Consejo Andino de Ministros de Relaciones Exteriores y/o la Comisión de la Comunidad Andina.

Las Decisiones forman parte del Ordenamiento Jurídico Comunitario y por tanto son de obligatorio cumplimiento en los cuatro Países Miembros, desde su entrada en vigencia. Por lo general, entran en vigencia desde la publicación en la Gaceta Oficial del Acuerdo de Cartagena excepto que la propia Decisión establezca una fecha diferente".

¿Qué es una RESOLUCIÓN?

De conformidad con lo previsto en el ordenamiento jurídico de la Comunidad Andina, la Secretaría General expresa su voluntad a través de Resoluciones. Las Resoluciones de la Secretaría General serán dictadas por el Secretario General y tramitadas de acuerdo al procedimiento activo correspondiente.

Las Resoluciones de la Secretaría General deberán ajustarse a lo establecido en el Acuerdo de Cartagena y en el Tratado de Creación del Tribunal de Justicia de la Comunidad Andina.
Las Resoluciones son de obligatorio cumplimiento para los cuatro Países Miembros desde la fecha de publicación de la Gaceta Oficial del Acuerdo de Cartagena, pues forman parte del Ordenamiento Jurídico Comunitario. Las Resoluciones pueden tener contenido reglamentario"

Posteriormente a estas secciones enlista las decisiones y resoluciones vigentes actualmente, con respecto muy específicamente al registro de directorios de empresa con fines estadísticos señala las siguientes:

- **Decisión 698 (DEC698[16]) Creación y Actualización de Directorios de Empresas, Fuente Cooperante Cooperación UE - Proyecto ANDESTAD.**

  La Comisión de la Comunidad Andina.
  VISTOS: El Capítulo IV del Acuerdo de Cartagena, y los Capítulos 3 y 11 del Anexo 1 de la Decisión 488 relativa al Programa Estadístico Comunitario;

  CONSIDERANDO: Que, para lograr la conformación del mercado común y alcanzar la comparabilidad entre las distintas estadísticas elaboradas para responder a las necesidades comunitarias, es necesario una coordinación con los servicios nacionales de estadística de los Países Miembros y una armonización en los conceptos, nomenclaturas y definiciones utilizadas por las estadísticas económicas;

  - *Que, la utilización por parte de los Países Miembros de definiciones comunes de unidades estadísticas permitirá la producción de información estadística integrada, armonizada y comparable;*
  - *Que, deben crearse y actualizarse registros de dichas unidades para poder recoger datos sobre ellas;*
  - *Que, existe una necesidad creciente de información sobre la estructura del sector productivo y la demografía de las empresas, la cual no puede ser cubierta por la situación actual de las estadísticas comunitarias;*
  - *Que, los directorios de empresas utilizables con fines estadísticos constituyen un instrumento necesario para el seguimiento de las modificaciones estructurales de la*

[16] NORMATIVIDAD ANDINA, DESICIONES, http://www.comunidadandina.org/normativa/dec/decnum.htm, http://www.comunidadandina.org/normativa/dec/D698.htm Publicado en la Gaceta Oficial 1678

economía resultantes de operaciones como: fusiones, absorciones, escisiones y transformaciones de empresas;

- *Que, actualmente no se dispone de información estadística suficiente en sectores donde predominan pequeñas y medianas empresas (PYMES), debido a la ausencia de directorios de empresas adecuados con fines estadísticos;*
- *Que, los directorios de empresas elaborados con fines estadísticos son uno de los elementos básicos de los sistemas de información que permiten organizar y coordinar encuestas estadísticas al proporcionar una base para el muestreo;*

Con base en ellas se definieron 10 Artículos y 2 Anexos,

Artículo 1.- La presente Decisión tiene por objeto establecer las bases de obtención de estadísticas del sector productivo confiable y comparable entre los Países Miembros de la Comunidad Andina, fijando normas para la constitución y la actualización de los directorios de empresas en los Sistemas Estadísticos Nacionales.

Artículo 2.- A efectos de recoger, transmitir, publicar y analizar la información, los Países Miembros utilizarán las unidades estadísticas con las definiciones establecidas en el Anexo I de la presente Decisión.

Artículo 3.- Los Países Miembros, a través de los Servicios Nacionales de Estadística, crearán y actualizarán un directorio nacional de empresas, con fines estadísticos, según las definiciones y la cobertura contempladas en los artículos siguientes.

Artículo 4.- En el directorio quedarán registradas, de conformidad con las definiciones contenidas en el Anexo I y con las restricciones previstas en el presente artículo:

a) Todas las empresas que ejerzan una actividad económica que suponga una contribución al producto interno bruto (PIB), así como las unidades legales que respondan de ellas, y las unidades locales que dependan de ellas.

b) Por Resolución de la Secretaría General de la Comunidad Andina, a propuesta del grupo de expertos y con la aprobación del Comité Andino de Estadística, se establecerán los criterios para la exclusión de empresas en el directorio.

Artículo 5.- Las variables a incluir en los directorios son las que figuran en el Anexo II de la presente Decisión.

Artículo 6.- Los Países Miembros actualizarán anualmente:

c) Las altas, bajas, fusiones y escisiones de las unidades estadísticas en los directorios.

d) Las variables relativas al nombre o razón social, dirección, y la forma jurídica.

e) Las otras variables de acuerdo a la disponibilidad de los datos.

Los Países Miembros guardarán los archivos históricos de los cambios realizados en los directorios.

Artículo 7.- Los Servicios Nacionales de Estadística responsables de la creación y actualización de los directorios establecidos en el artículo 3 de la presente Decisión estarán autorizados para recoger la información necesaria de los registros administrativos mantenidos por las entidades públicas y privadas.

Se requerirá el uso de las fuentes fiscales para la actualización de al menos la información contenida en los literales a) y b) del artículo 6 de la presente Decisión.

Artículo 8.- La Secretaría General de la Comunidad Andina convocará a reuniones de expertos gubernamentales en directorios de empresas, en el marco del artículo 38 de la Decisión 471 de la Comisión, a fin de establecer por Resolución los procedimientos necesarios para la aplicación del Artículo 4, las actualizaciones que requieran los Anexos I y II de la presente Decisión y actualizar las listas de variables, la frecuencia de actualización de los directorios y, en general, los ajustes que se requieran de acuerdo a la evolución económica y técnica del proceso de integración.

Artículo 9.- Cuando sea necesario realizar adaptaciones importantes en los sistemas estadísticos nacionales para cumplir lo dispuesto en los Anexos, la Secretaría General podrá otorgar excepciones por períodos transitorios. A partir del final de dicho período, serán aplicables plenamente las obligaciones derivadas de la presente Decisión.

Artículo 10.- La presente Decisión entrará en vigencia al día siguiente de su publicación en la Gaceta Oficial del Acuerdo de Cartagena.

Dada en la ciudad de Lima, Perú, a los diez días del mes de diciembre del año dos mil ocho.

Anexos de este artículo:

Anexo I Unidades Estadísticas.

- *Lista de las unidades estadísticas.*
- *Definiciones.*

Anexo II Variables a incluir en los directorios.

- *Unidad Legal.*
- *Empresa.*
- *Unidades locales.*

- **Resolución 1218 (RES1218[17]) Cobertura de los Directorios de Empresas.**

Definida en base a consideraciones expuestas resuelve 4 Artículos, solo se presentan los 2 primeros:

Artículo 1.- La cobertura mínima obligatoria del directorio de empresas estará comprendida por todas las empresas que ejerzan una actividad económica que contribuyan al Producto Interno Bruto (PIB).

No se incluirán en los Directorios de Empresas:

- *Las actividades productivas realizadas dentro de los hogares destinadas al autoconsumo o correspondientes al arrendamiento de bienes inmobiliarios no constituyen una unidad productiva de tipo empresarial, por lo tanto se excluyen de los directorios de empresas, y;*

- *Las unidades legales (y las empresas que dependen de ellas) que no se encuentren inscritas en algún registro administrativo, ya sea fiscal, mercantil, de sociedades, del seguro social obligatorio u otro.*

Artículo 2.- La inclusión de las empresas cuya actividad económica principal sea agricultura, pesca o gobierno (secciones A u O de la CIIU Rev. 4) será facultativa.

[17] NORMATIVIDAD ANDINA, RESOLUCIONES, http://www.comunidadandina.org/normativa/res/resoluciones.htm, http://www.comunidadandina.org/normativa/res/R1218sg.htm Publicado en la Gaceta Oficial 1700

- **Resolución 1273 (RES1273[18]) Manual de Recomendaciones sobre los Directorios de Empresas con fines estadísticos en la Comunidad Andina.**

Esta Resolución resuelve 3 artículos en donde el Artículo 1 define:

Artículo 1.- Adoptar el Manual de Recomendaciones sobre Directorios de Empresas con fines estadísticos en la Comunidad Andina, contenido en el Anexo de la presente Resolución.

Y en su único anexo el "Manual de Recomendaciones sobre los Directorios de Empresas con fines estadísticos en la Comunidad Andina" publicando su primera versión en Mayo del 2009, compuesto de 8 capítulos y 1 Anexo.

El Índice general del Manual se presenta a continuación:

Introducción

Capitulo 1 Ámbito de Aplicación
Capitulo 2 Armonización de los Directorios de Empresas
*Capítulo 3 Objetivos y usos de los Directorios de Empresas con fines Estadí*sticos
Capítulo 4 Gestión de Directorios
Capítulo 5 Contenido del Directorio
Capítulo 6 Cobertura del Directorio
Capítulo 7 La Unidad Legal y la Empresa
Capítulo 8 Las Unidades Locales y un solo anexo:
Anexo I Directrices sobre determinadas actividades para las unidades locales.

- **Resolución 1274 (RES1274[19]) Guía para la construcción de los Directorios de Empresas con fines estadísticos en la Comunidad Andina.**

Resuelve dos artículos y un Anexo y en el primero artículo define:

Artículo 1.- Adoptar la Guía para la construcción de los Directorios de Empresas con fines estadísticos en la Comunidad Andina, contenido en el Anexo I de la presente Resolución.

[18] NORMATIVIDAD ANDINA, RESOLUCIONES, http://www.comunidadandina.org/normativa/res/resoluciones.htm, http://www.comunidadandina.org/normativa/res/R1273sg.htm , Publicado en la Gaceta Oficial 1748
[19] NORMATIVIDAD ANDINA, RESOLUCIONES, http://www.comunidadandina.org/normativa/res/resoluciones.htm, http://www.comunidadandina.org/normativa/res/R1274sg.htm, Publicado en la Gaceta Oficial 1749

El Anexo I está integrado de 5 capítulos y del cual solo presentamos el índice general.

Con base en esta normatividad y en la sección del “Proyecto de Cooperación UE-CAN Estadísticas ANDESTAD, Sub sección Áreas Temáticas, Directorios de Empresas[20]

[20] http://www.comunidadandina.org/andestad/areas_tematicas.asp?id=13&m=2

En donde se encuentran todos los documentos, reuniones y e investigaciones en donde se ha plasmado la evolución de este proyecto. Desde el asentamiento de las Reuniones de Grupo de trabajo GT9, que es un grupo de trabajo denominado “Expertos gubernamentales sobre registros administrativos, directorios de Empresas y Marcos Muestrales en la Comunidad ANDIDA” que tuvieron su primera reunión en Quito, Ecuador del 9 al 11 de Noviembre de 2005 hasta la más reciente realizada en Lima, Perú del 19 al 20 de Octubre del 2009.

Registro de los documentos generados como son: Los estudios de diagnostico, Estudios Específicos, Cooperación Horizontal. Decisiones y Resoluciones y los Documentos de Referencia. Dejando la evidencia histórica de su gran trabajo colaborativo, los beneficios y logros que han alcanzan con ellos.

c.- CEPAL y sus recomendaciones.

Naciones Unidas, es una organización internacional fundada en 1945 tras la Segunda Guerra Mundial por 51 países que se comprometieron a mantener la paz y la seguridad internacionales, desarrollar relaciones amistosas entre las naciones y promover el progreso social, mejores niveles de vida y los derechos humanos. Debido a su singular carácter internacional, y las competencias de su Carta fundacional, la Organización puede adoptar una decisión sobre una amplia gama de cuestiones, y proporcionar un foro a sus 192 Estados Miembros para expresar sus opiniones, a través de la Asamblea General, el Consejo de Seguridad, el Consejo Económico y Social y otros órganos y comisiones.

El Consejo Económico y Social (ECOSOC)[21] es el órgano que coordina la labor económica y social de las Naciones Unidas y de las instituciones y organismos especializados que conforman el sistema de las Naciones Unidas. Está formado por 54 miembros elegidos por la Asamblea General, con mandatos de tres años. Cada miembro tiene un voto y las decisiones dentro de este órgano se toman por mayoría simple.

Funciones y poderes.

El Consejo Económico y Social tiene las siguientes prerrogativas:

- *Servir de foro central para el examen de los problemas económicos y sociales y la elaboración de recomendaciones de política dirigidas a los Estados Miembros y al Sistema de las Naciones Unidas*
- *Realizar o iniciar estudios, informes y recomendaciones sobre cuestiones de índole económica, social, cultural educacional, de salud y otros asuntos conexos*
- *Fomentar el respeto y la observancia a los derechos humanos y las libertades fundamentales de todos*
- *Convocar conferencias internacionales y preparar proyectos de convención para someterlos a la consideración de la Asamblea General*
- *Coordinar las actividades de los organismos especializados, mediante consultas y recomendaciones directas, o haciéndole recomendaciones a la Asamblea y a los Estados Miembros*
- *Celebrar consultas con las organizaciones no gubernamentales que se ocupan de asuntos que competen al Consejo*

[21] http://www.cinu.org.mx/onu/estructura/ecosoc.htm

Órganos Subsidiarios.

Estos llevan a cabo la labor permanente del ECOSOC y son una serie de comisiones y comités que se reúnen a intervalos regulares y presentan sus informes al Consejo. Dicho mecanismo subsidiario está formado por:

- ***Nueve comisiones orgánicas**: examinan cuestiones en sus respectivas esferas de responsabilidad y conocimientos y hacen recomendaciones:*
  - *Comisión de Estadística.*
  - *Comisión de Población y Desarrollo.*
  - *Comisión de Desarrollo Social.*
  - *Comisión de la Condición Social y Jurídica de la Mujer.*
  - *Comisión de Estupefacientes.*
  - *Comisión de Prevención del Delito y Justicia Penal.*
  - *Comisión de Ciencia y Tecnología para el Desarrollo.*
  - *Comisión sobre el Desarrollo Sostenible.*
  - *Foro de las Naciones Unidas sobre los Bosque.*
- ***Cinco comisiones regionales**: se agrupan mediante criterios propios de las Naciones Unidas y su mandato principal es el de promover medidas que fomenten el desarrollo económico regional y fortalezcan las relaciones económicas de los países de la región entre sí y con el resto del mundo.*
  - *Comisión Económica para África (sede en Addis Abeba, Etiopía).*
  - *Comisión Económica y Social para Asia y el Pacífico (sede en Bangkok, Tailandia).*
  - *Comisión Económica para Europa (sede en Ginebra, Suiza).*
  - *Comisión Económica para América Latina y el Caribe CEPAL (sede en Santiago, Chile).*
  - *Comisión Económica y Social para Asia Occidental (sede en Beirut, Líbano).*
- ***Tres comités permanentes**:*
  - *Comité del Programa y de la Coordinación*
  - *Comité Encargado de las Organizaciones no Gubernamentales.*
  - *Comité Encargado de las negociaciones con las organizaciones Intergubernamentales.*
- ***Organismos permanentes de expertos**: que tratan temas tales como la planificación del desarrollo, los recursos naturales y los derechos económicos, sociales y culturales*

La Comisión Económica para América Latina (CEPAL)[22] *fue establecida por la resolución 106(VI) del Consejo Económico y Social, del 25 de febrero de 1948, y comenzó a funcionar ese mismo año. En su resolución 1984/67, del 27 de julio de 1984, el Consejo decidió que la Comisión pasara a llamarse Comisión Económica para América Latina y el Caribe. Se fundó para contribuir al desarrollo económico de América Latina, coordinar las acciones encaminadas a su promoción y reforzar las relaciones económicas de los países entre sí y con las demás naciones del mundo. Posteriormente, su labor se amplió a los países del Caribe y se incorporó el objetivo de promover el desarrollo social.*

La CEPAL tiene dos sedes subregionales, una para la subregión de América Central, ubicada en México, D.F. y la otra para la subregión del Caribe, situada en Puerto España, que se establecieron en junio de 1951 y en diciembre de 1966, respectivamente. Además, tiene oficinas nacionales en Buenos Aires, Brasilia, Montevideo y Bogotá y una oficina de enlace en Washington, D.C.

Los 33 países de América Latina y el Caribe son miembros de la CEPAL, junto con algunas naciones de América del Norte, Europa y Asia que mantienen vínculos históricos, económicos y culturales con la región. En total, los Estados miembros son 44, y 9 los miembros asociados, condición jurídica acordada para algunos territorios no independientes del Caribe.

Tabla 1[23]

Los Estados miembros de la CEPAL y los Enlaces a los sitios web de las Oficinas Nacionales de Estadística

Alemania Federal Statistical Office	Haití. Institut Haitien de Statistique et d'Informatique
Antigua y Barbuda. Ministry of Finance. Statistics Division.	Honduras Instituto Nacional de Estadística
Argentina Instituto Nacional de Estadística y Censos (INDEC)	Italia L'Istituto Nazionale di Statistica
Bahamas. Department of Statistics	Jamaica Statistical Institute of Jamaica
Barbados. Statistical Service	Japón Statistics Bureau
Belice Central Statistical Office	México Instituto Nacional de Estadística y Geografía (INEGI)
Bolivia Instituto Nacional de Estadística	Nicaragua Instituto Nacional de Estadísticas y Censos (INEC)
Brasil Instituto Brasilero de Geografía y Estadística	

[22] http://www.eclac.org/cgi-bin/getprod.asp?xml=/noticias/paginas/4/21324/P21324.xml&xsl=/tpl/p18f-st.xsl&base=/tpl/top-bottom_acerca.xsl

[23] Estados miembros y miembros asociados a la CEPAL. http://www.eclac.org/cgi-bin/getprod.asp?xml=/noticias/paginas/3/21493/P21493.xml&xsl=/tpl/p18f-st.xsl&base=/tpl/top-bottom_acerca.xsl

- *(IBGE)*
- *Canadá, Statistics Canada*
- *Chile, Instituto Nacional de Estadísticas*
- *Colombia, Departamento Administrativo Nacional de Estadística (DANE)*
- *Costa Rica,, Instituto Nacional de Estadística y Censos (INEC)*
- *Cuba, Oficina Nacional de Estadística*
- *Dominica. Central Statistical Office, Ministry of Finance. csoda@cwdom.dm*
- *Ecuador, Instituto Nacional de Estadística y Censos (INEC)*
- *El Salvador, Dirección General de Estadística y Censos (DIGESTYC)*
- *España, Instituto Nacional de Estadística*
- *Estados Unidos, Bureau of Labor Statistics*
- *Francia, Institut National de la Statistique et des Études Économiques (INSEE)*
- *Granada. Central Statistical Office, Ministry of Finance gogstats@hotmail.com*
- *Guatemala, Instituto Nacional de Estadística*
- *Guyana. The Bureau of Statistics, sisbos@networksgy.com*
- Países Bajos Statistics Netherlands
- Panamá Dirección de Estadística y Censos, Contraloría General de la República
- Paraguay. Dirección General de Estadística, Encuestas y Censos del Paraguay
- Perú, Instituto Nacional de Estadística e Informática (INEI)
- Portugal, Instituto Nacional de Estatística
- Reino Unido de Gran Bretaña e Irlanda del Norte, National Statistics
- República Dominicana, Oficina Nacional de Estadística
- Saint Kitts y Nevis. Statistics Division, Ministry of Finance, Planning and Development planningstk@caribsurf.com
- San Vicente y Las Granadinas. Statistical Office, Central Planning Division, Ministry of Finance & Planning , statssvg@vincysurf.com
- Santa Lucía, Government Statistics Department
- Suriname. General Bureau of Statistics, statistics@cq-link.sr
- Trinidad y Tobago, Central Statistical Office
- Uruguay, Instituto Nacional de Estadística
- Venezuela, Instituto Nacional de Estadística

Los países miembros asociados de la CEPAL

- *Anguila. Statistics Unit*
- *Antillas Neerlandesas. Central Bureau of Statistics*
- *Aruba. Central Bureau of Statistics*
- *IslasTurcos y Caicos. Department of Economic Planning and Statistics*
- *Islas Vírgenes Británicas. Central Administration Complex, Ministry of Finance [dpu@bvigovernment.org]*
- *Islas Vírgenes de los Estados Unidos*
- *Monserrat. Statistics Department, Ministry of Finance and Economic Development. [devunit@candw.ag]*
- *Puerto Rico*

Mandato y misión.[24]

La secretaría de la Comisión Económica para América Latina y el Caribe (CEPAL):

- *Presta servicios sustantivos de secretaría y documentación a la Comisión y a sus órganos subsidiarios;*
- *Realiza estudios, investigaciones y otras actividades de apoyo de conformidad con el mandato de la Comisión;*
- *Promueve el desarrollo económico y social mediante la cooperación y la integración a nivel regional y subregional;*
- *Recoge, organiza, interpreta y difunde información y datos relativos al desarrollo económico y social de la región;*
- *Presta servicios de asesoramiento a los gobiernos a petición de éstos y planifica, organiza y ejecuta programas de cooperación técnica;*
- *Planifica y promueve actividades y proyectos de cooperación técnica de alcance regional y subregional teniendo en cuenta las necesidades y prioridades de la región y cumple la función de organismo de ejecución de esos proyectos;*
- *Organiza conferencias y reuniones de grupos intergubernamentales y de expertos y patrocina cursos de capacitación, simposios y seminarios;*
- *Contribuye a que se tenga en cuenta la perspectiva regional, respecto de los problemas mundiales y en los foros internacionales y plantea en los planos regional y subregional cuestiones de interés mundial;*
- *Coordina las actividades de la CEPAL con los de los principales departamentos y oficinas de la Sede de las Naciones Unidas, los organismos especializados y las organizaciones intergubernamentales a fin de evitar la duplicación y lograr la complementariedad en el intercambio de información.*

La ECOSOC en su Resolución 2000/7[25] *del 25 de julio de 2000 aprobó el establecimiento de la Conferencia Estadística de las Américas de la Comisión Económica para América Latina y el Caribe, como un órgano subsidiario de la CEPAL, que contribuye al progreso de las políticas y actividades estadísticas de la región*.

En las cuatro reuniones de la Conferencia Estadística de las Américas de la Comisión Económica para América Latina y el Caribe las cuales se han llevado a cabo desde el 2001 al 2007, se han

[24] http://www.eclac.org/cgi-bin/getprod.asp?xml=/noticias/paginas/9/21469/P21469.xml&xsl=/tpl/p18f-st.xsl&base=/tpl/top-bottom_acerca.xsl
[25] http://www.un.org/spanish/ecosoc/docs/index.shtml

tratado temas sustantivos relacionados con Elaboración de directorios y utilización de registros administrativos como fuente primaria de información.

En el informe de la primera reunión se destaco el concepto como parte del temario, ya en la segunda reunión en su informe declaran haber llevado a cabo una sesión denominada "sesión sobre "Elaboración de directorios y utilización de registros administrativos como fuente primaria de información, así como marco de referencia de encuestas y de otras investigaciones estadísticas" la cual fue moderada por el delegado de Panamá. La delegación de Brasil presentó el documento de trabajo "Directorios estadísticos de empresas elaborados a partir de registros administrativos" (LC/L.1892(CEA.2003/7)[26]).

En su tercera reunión el INE de España presentó el código de buenas prácticas del sistema estadístico europeo, *cuyo propósito es promover la credibilidad, calidad y la aplicación de los mejores principios en materia estadística, entre los que se mencionaron la confiabilidad, imparcialidad e independencia de las oficinas productoras de estadísticas.*

Para el informe de la cuarta reunión destacan en uno de los puntos del informe los delegados señalaron especialmente la importancia de analizar el tema del cambio del año base. *La Secretaría recordó que los problemas financieros obstan en algunos casos al logro de este objetivo, ya que los países no incluyen en sus presupuestos recursos para modernizar las cuentas nacionales ni para llevar a cabo estos cambios. Los asistentes coincidieron en que el Sistema de Cuentas Nacionales de 1993 (SCN93) debe implementarse de manera gradual, según las posibilidades y necesidades de cada país. Se propuso a la Conferencia elaborar un plan regional de implementación del SCN93 por etapas y apoyar en primer lugar a los países que necesiten mejores estadísticas básicas; asimismo, se solicitó a la Conferencia que recomendara a los países e instituciones donantes que los recursos se distribuyeran de acuerdo con esta prioridad. Por último se propuso a la CEPAL fortalecer su trabajo como secretaría del grupo, para lograr una coordinación eficiente, y se invitó a los diferentes países a participar activamente en el trabajo que desarrolla. Varios delegados apoyaron esta propuesta y la hicieron extensiva a los grupos sobre censos, objetivos de desarrollo del Milenio, tecnologías de la información y de las comunicaciones y género. Se destacó la importancia de los directorios de empresas como requisito básico para la elaboración de buenas estadísticas económicas y la necesidad de fortalecer la cooperación entre las oficinas de estadísticas y las oficinas de impuestos. Por otra parte, también se destacó la cooperación que brindaba el Departamento de Estadística del Fondo Monetario Internacional a*

[26] http://www.eclac.cl/ceacepal/documentos/lcl1892p.pdf

este grupo de trabajo, así como el apoyo ofrecido al grupo sobre capacitación.[27] *En este informe en la sección de resoluciones declara.*

"Acoge la iniciativa de varios países de promover en la región el conocimiento y la discusión para la adaptación del Código de buenas prácticas estadísticas europeas, de acuerdo con la realidad de cada país, y solicita a la Oficina de Estadística de las Comunidades Europeas (EUROSTAT) y a la Comisión Económica para América Latina y el Caribe la preparación de un programa de acción para llevar a cabo esta iniciativa"

- **Taller "Directorio de Empresas y Establecimientos: Desarrollos recientes y desafíos actuales y futuros en América Latina"**

En Santiago de Chile, Sede de la CEPAL, el 22 y 23 de septiembre de 2008 se realizó el *Taller "Directorios de empresas y establecimientos: Desarrollos recientes y desafíos actuales y futuros en América Latina"*[28]

En el cual varios países miembros de la CEPAL expusieron sus situaciones y sus experiencias actuales:

En este taller se expusieron los temas siguientes:

- *Tema 1: Marco Conceptual y metodológico*
- *Tema 2: Estrategias de mejoramiento del Directorio*
- *Tema 3: Usos del Directorio*
- *Tema 4: Resultados y Análisis de la Encuesta aplicada a países seleccionados de América Latina.*

Para ver una lista más detallada de las presentaciones expuestas en cada tema y los documentos complementarios ver:

ANEXO I.- "*Taller "Directorios de empresas y establecimientos: desarrollos recientes y desafíos actuales y futuros de América Latina"*

[27] http://www.eclac.cl/publicaciones/xml/8/30028/LCL2795e.pdf
[28] http://www.eclac.org/scaeclac/taller_directorios_empresas_2008.htm

Este talle además de las experiencias expuestas por los países participantes conto con una investigación que expusieron dos consultores *Vicenta Mardones y Mauricio Ponce que fue plasmado en un documento llamado "Directorios de Empresas y Establecimientos: Revisión de la experiencia internacional y situación, en algunos países de América Latina"* el cual contó con la colaboración financiera del Banco Interamericano del Desarrollo [29] la cual cuenta con una sección de los resultados de un cuestionario llamado "*Encuesta de caracterización, generación y utilización de directorios en países seleccionados de América Latina", en dicho cuestionario se especifica que el objetivo de la encuesta es el de "conocer el estado actual y los principales desafíos en lo que respecta al desarrollo de Directorios de Empresas y sus Establecimientos locales" aplicado a 7 países los cuales fueron Brasil, Chile, El Salvador, México, Nicaragua, Perú y República Dominicana con la finalidad de conocer información relevante sobre el uso y desarrollo de Directorios de Empresas y/o Establecimientos a nivel de América Latina, además de que destacan la importancia de esté insumo para la generación de información básica para enfrentar las necesidades de información con las que se enfrentan los países para la toma de decisiones en sus complejas estructuras económicas además de las necesarias para enfrentar los cambios políticos y económicos y los intercambios de información con sus países vecinos o con el resto del mundo.*

En este paper define al Directorio como:*" Un Directorio de empresas y/o establecimientos es un registro censal de las empresas y otras unidades con actividad económica existentes en un país o en una zona geográfica o administrativa de éste, públicas y privadas, sean éstas personas jurídicas o personas físicas, que considera variables de identificación, localización, actividad, estado de funcionamiento, nivel de actividad económica y empleo, así como vinculaciones con otras empresas*."

Además lo consideran como un instrumento fundamental y especifican:
"*El Directorio es un instrumento fundamental para la generación de muestras y factores de expansión en el proceso de producción de encuestas, para el levantamiento de censos económicos, para la realización de estudios sobre la dinámica demográfica de las empresas y para generar estudios de la evolución del empleo y sus características y líneas base para los estudios de seguimiento y evaluación de impacto de las políticas públicas. De la misma forma apoya la elaboración de las Cuentas Nacionales, entre otros ámbitos. "*

[29] http://www.eclac.org/scaeclac/documentos/2008_09_CEA_tallerDirectorios_MARDONESPAPER.pdf

Destacando la importancia de la cobertura, actualización y mejora continua.

A través de estas páginas he revisado la importancia de los directorios de empresas y/o establecimientos con fines estadísticos pero debemos destacar que no todos los países hablando de su composición estructural económica están formados además de las diferentes formas en como recogen la información económica de su país puede venir de diferentes fuentes, organismos, instituciones, Bancos Centrales o Censos llevados a cabo por instituciones creadas para este fin o bien los registros administrativos de las instituciones antes mencionadas. Otro punto que debemos aclara es que cada país lleva el registro de su información económica de acuerdo con sus capacidades, limitaciones y recursos y lineamientos jurídicos con los que cuente y conforme a la composición estructural económica que tiene establecido en su país aunque sea miembro de los organismos o instituciones antes mencionados.

Otro punto muy importante para los directorios de cada país son las fuentes de entrada o alimentadoras de los directorios de empresas y/o establecimientos así como las unidades estadísticas que lo compondrán y las variables económicas con las que se cuente disponibles para conformar esos directorios,

Una característica también muy peculiar de los directorios es la composición de la estructura económica del país, nos referimos a esta, como la relación existente entre las unidades estadísticas económicas. Existen países en donde el registro se hace hacia la micro unidad económica, que puede ser el establecimiento o la unidad local, otros países en donde a partir de la identificación de la micro unidad económica identifican empresas o viceversas la identificación de la macro unidad económica en algunos casos llamada empres y a partir de ella identifica todas las micro unidades que están relacionadas con ella, esto es de acuerdo con el país.

Otra particularidad de cada país con respecto de los directorios de empresas y/o establecimientos son las clasificadoras actividades económicas que utilizan,

Como se ha hecho mención en algún otro punto la calidad de los directorios también va a depender de sus fuentes, procesos y estándares que acoja en cada uno de ellos.

El siguiente extracto de la tabla No. 2, que es presentada a continuación y la cual está tomada del paper "*Directorios de Empresas y Establecimientos: Revisión de la experiencia internacional y situación en algunos países de América Latina"* de los consultores *Vicenta Mardones y Mauricio Ponce* y cuya fuente es la ***"****Encuesta de caracterización, generación y*

utilización de directorios en países seleccionados de América Latina", nos permite dar un pequeño panorama en síntesis de algunos de los puntos mencionados anteriormente de los países que fueron elegidos para responder el cuestionario.

Tabla 2

Descriptores de caracterización general y del contexto en que opera cada directorio, según país.[30]

	BRASIL	MÉXICO	PERÚ	CHILE	EL SALVADOR	NICARAGUA	REPÚBLICA DOMINICANA
Año de origen	1994	1998	2004	2004	1998	2002	2008
Institución responsable	IBGE	INEGI	INEI	INE	DIGESTYC	Banco Central de Nicaragua	Oficina Nacional de Estadística
Nombre del Directorio	Cadastro Central de Empresas	Directorio Nacional de Unidades Económicas	Sistemas de Registros de Empresas	Directorio nacional de Empresas y establecimientos	Directorio de Establecimientos y Empresas	Directorio Económico Urbano Nacional	Directorio de Empresas y Establecimientos
Entidad Administradora	Gerencia de Cadastro Central de Empresas	Dirección de Diseños y Marcos Estadísticos de la DGE	No existe. Forma parte de las tareas de la Dirección	Proyecto Directorio	División Censos y Encuestas Económicas	Coordinación de Estadísticas Económicas BCN	División de directorio de Empresas
Sujeto de Estudio	Empresas y establecimientos	Empresas y establecimientos	Empresas y personas físicas	Empresas y personas físicas	Empresas y establecimientos	Establecimientos	Empresas y personas físicas
Cobertura	Nacional, excluye servicio doméstico	Nacional Área Urbana y MyGEs rurales, muestra para MyPEs rurales	Nacional Área Urbana	Nacional	Nacional	Nacional Urbano	Nacional
Fuente principal de los datos	Registros catastro de empleo y desempleo y de informaciones sociales	Hasta 2008 Censos económicos Desde 2008 registros administrativos de Seguridad Social, Aduanas, Construcción	Registros Tributarios y Aduaneros SUNAT	Registros Tributarios SII Registros Aduaneros y Síndico de Quiebras	Empresas que tramitan solvencia Estadística para Matrícula Empresa y establecimiento	Superintendencia de bancos y Corporación de Zonas Francas Encuesta económica Anual BCN	Registros Tributarios Aduaneros y de Seguridad Social
Clasificación Activ. Econ.	CIIU REV 4	SCIAN2002	CIIU REV 3	CIIU REV 3	CLAEES	AD HOC	CIIU REV 3

"*Directorios de Empresas y Establecimientos: Revisión de la experiencia internacional y situación en algunos países de América Latina"* de los consultores ***Vicenta Mardones y Mauricio Ponce***

[30] "*Directorios de Empresas y Establecimientos: Revisión de la experiencia internacional y situación en algunos países de América Latina"* de los consultores ***Vicenta Mardones y Mauricio Ponce***

9.1.2.- Panorama Nacional.

Situación actual caso INEGI (México)

- **La LIEG, Ley del Sistema Nacional de Información Estadística y Geográfica.**

México cuenta con una gran experiencia en la generación de información estadística del país. En la actualidad México tiene como instituto generador de información estadística al el organismo autónomo denominado Instituto Nacional de Estadística y Geografía conocido por sus siglas INEGI, quien tiene un reconocimiento internacional por su gran experiencia en la materia, pertenece o es país observador en varios organismos internacionales. En el año 2009 público la información histórica de la generación de información estadística oficial generada a través de los años en una publicación nombrada "Cronología de la estadística en México (1521-2008)".[31]

Esta publicación presenta como en al año 1983 ocurrieron dos acontecimientos importantes para nuestro país los cuales fueron:

- *Surge el Instituto Nacional de Estadística, Geografía e Informática (heredero de la CGSNEGI, también adscrito a la SPP), integrado por las direcciones generales de Integración y Análisis de la Información, Política Informática, Estadística y Geografía. Su titular es el doctor en Economía, Pedro Aspe Armella.*
- *Se publican, en el Diario Oficial de la Federación, las reformas y adiciones a la LIEG, las cuales establecen que el INEGI asume la función coordinadora de los Sistemas Nacionales Estadístico y de Información Geográfica.*

La LIEG que es la Ley del Sistema Nacional de Información Estadística y Geográfica, que en su TEXTO VIGENTE, la Nueva Ley publicada en el Diario Oficial de la Federación el 16 de abril de 2008, entrando en vigor el 15 de Julio de este mismo año, establece:

En su CAPÍTULO IV De los Subsistemas Nacionales de Información, establece:

Artículo 17.- El Sistema contará con los siguientes Subsistemas Nacionales de Información:

I. Demográfica y Social;

II. Económica, y

III. Geográfica y del Medio Ambiente.

[31] DR © 2009, Instituto Nacional de Estadística y Geografía, http://www.inegi.org.mx/prod_serv/contenidos/espanol/biblioteca/default.asp?accion=2&upc=702825460761&seccionB=bd

Cada Subsistema tendrá como objetivo producir, integrar y difundir Información demográfica y social; económica y financiera, y geográfica y del medio ambiente, según corresponda.
El Instituto deberá emitir las disposiciones generales para regular el funcionamiento de los Subsistemas Nacionales de Información.
La Junta de Gobierno, previa opinión favorable del Consejo, podrá crear otros Subsistemas que sean necesarios para el adecuado funcionamiento del Sistema.

SECCIÓN II Del Subsistema Nacional de Información Económica
Artículo 23.- El Subsistema Nacional de Información Económica, contará con una infraestructura de información que contenga como mínimo, un marco geoestadístico y un Directorio Nacional de Unidades Económicas.
El Directorio a que se hace referencia en el párrafo anterior, así como las clasificaciones económicas que formen parte del mismo, son de uso obligatorio para la organización de los registros administrativos de los que se pueda obtener Información de Interés Nacional.

Artículo 24.- El Subsistema Nacional de Información Económica deberá generar un conjunto de indicadores clave, relacionados como mínimo con lo siguiente: sistema de cuentas nacionales; ciencia y tecnología; información financiera; precios y trabajo.

Artículo 25.- El Instituto elaborará, con la colaboración de las Unidades, los indicadores a que se refiere el artículo anterior a partir de la información básica proveniente de:

> *Los censos nacionales económicos y agropecuarios, o los esquemas alternativos que pudieran adoptarse en el futuro para sustituirlos total o parcialmente; un sistema integrado de encuestas en unidades económicas, y los registros administrativos que permitan obtener Información en la materia."*

Además de la LIEG el INEGI cuenta con un reglamente interior que establece las facultades y atribuciones de sus direcciones generales y direcciones generales adjuntas del Instituto.
Es con base en esta ley que el INEGI como organismo nacional autónomo, tiene la difícil tarea de mantener un directorio nacional de unidades económicas que cumpla con los requerimientos de información estadística nacional para llevar acabo la toma de decisiones necesarias para el país y las políticas económica, las planeaciones de inversión tanto para el sector publico como para el privado en las cuales las secretarias o instituciones nacionales se apoyen.

En la actualidad el DNUE esta alimentado con la información recabada en los Censos Económicos[32] y las fuentes internas generadas, aunque existen en nuestro país fuentes oficiales que podrían alimentar al DNUE como son la SE, SHCP, el IMSS, algunas cámaras como CMIC, etc.. El clasificador oficial de los censos económicos es el "SCIAN el Sistema de Clasificación Industrial de América del Norte" y las unidades de observación son el establecimiento y la empresa.

En esta investigación se detecta algunas barreras para alimentar al DNUE con directorios externos como son las políticas, normas o legislación de cada institución al respecto de la compartición de información, el registro de la información , las unidades de observación, los clasificadores de actividades económicas, la desagregación o concentración de información, las variables recabadas de cada unidad de observación, entre otras la nula o carencia de la existencia de un identificador único que relacione al establecimiento entre todas las fuentes, por lo que la utilización de metodologías alternativas como son el concepto Record Linkage podrían ser una opción viable.

[32] Los Censos Económicos se llevan en México a cabo cada 5 años, (los primeros 4 censos en años con terminación 0 y 5, en 1950 no hubo levantamiento censal sino hasta 1951 y hasta 1986 en donde no espero 5 años sino que el siguiente levantamiento censal fue en 1989 y a la fecha que se levanta cada 5 años en años con terminación 4 y 9, aunque la historia de los censos económicos data del año 1930 este primer levantamiento tuvo objetivos muy modestos y solamente en el sector manufacturero. Actualmente el censo económico cubre los 9 grandes sectores de actividades económicas y la cobertura es nacional.

9.2.- Teoría Record Linkage.

9.2.1.- Teoría base y conceptos generales.

Record Linkage (Definición). *Se define como el uso de técnicas estadísticas con el objetivo de identificar pares de registros procedentes de ficheros diferentes de longitud n, o en un mismo fichero, cuando no se disponga de un identificador único que los relacione uno a uno*.

La expresión Record Linkage (vinculación de registros) *se refiere precisamente al uso de técnicas algorítmicas para encontrar registros que, aunque no identifiquen exactamente de la misma forma a una entidad, si se refieren a la misma.*

En este documento definiremos Record Linkage como la metodología que usaremos para EMPATAR dos ficheros mediante campos que contengan información similar y que carecen de un identificador único que los relacione.

El concepto Record Linkage se puede clasificar en dos conceptualizaciones metodológicas:

- El método determinístico. En el cual se busca que el EMPATE de una o más variables sea exacto, sin embargo no es muy conveniente adherirse mucho a esta definición debido a que los errores producidos en la codificación o en la captación de la información de las variables en los ficheros ya sean errores tipográficos o bien la omisión, adición o traslape de uno o más caracteres podrían provocar que esta definición se torne inflexible, la visión debe tomarse un poco flexible al utilizar métodos para comparación de cadenas aproximadas y no exactas. Pero si limitando el número de variables que vayan a ser comparadas.

- El método probabilístico. Permite la utilización de un mayor número de variables que se comparen en cuanto a la información que se utilice para el EMPATE o NO EMPATE de los ficheros y para tomar la decisión deben tomarse en cuenta todas las variables, aunque ellas deberán tener relacionado un peso que permitirá que en el momento de la evaluación para la determinación de si es o no un EMPATE el peso jugará un papel importante en el poder decisivo del método.

Presentación del modelo Teórico base.

Primera parte del Modelo Teórico Fellegi y Sunter según el artículo. "A theory for Record Linkage." 1969.

Los métodos probabilísticos de fusión de registros llevan a cabo la comparación de registros de dos ficheros distintos. Se denotan por A y B los ficheros a ser comparados y por a y b sus elementos, respectivamente.
Se supone que ambos archivos tienen elementos comunes, y por tanto, el objetivo de la fusión es reconocer, de entre todos los pares de AxB que podrían formarse, aquellos que se refieran a la misma persona, objeto o entidad. Es decir, el objetivo es dividir el conjunto

$$AxB = \{(a, b) \mid a \in A, b \in B\}$$

es la unión de los conjuntos disjuntos

$$M = \{(a, b) \mid a = b, a \in A, b \in B\} \qquad (1)$$

y

$$U = \{(a, b) \mid a \neq b, a \in A, b \in B\} \qquad (2)$$

a los que se llaman conjunto de matches y no-matches respectivamente.

Cada unidad de la población tiene unas características asociadas, tales como, nombre, apellidos, edad o dirección. Se han de identificar aquellos registros que se refieren a una misma persona, objeto o entidad. Sin embargo, el proceso de creación de los ficheros podría introducir errores o imprecisiones (errores de codificación, trascripción y tecleo, variaciones tipográficas o fonéticas, pérdida de datos, etc.) en los registros generados. Como resultado de estos errores, dos miembros de A y B que no se refieren al mismo individuo podrían generan registros idénticos y, más frecuentemente, dos miembros idénticos de A y B podrían producir registros distintos. Se denotan los registros correspondientes a los miembros de A y B por α(a) y β(b) respectivamente.

El primer paso al intentar emparejar registros de dos archivos es compararlos. El resultado de la comparación es un conjunto de códigos, que son codificados en afirmaciones del tipo: .el nombre coincide en ambos registros., .el nombre coincide y es Pedro., .el nombre no coincide., .el nombre está en blanco en uno de los registros. o .existe acuerdo en la zona de la ciudad de la dirección pero no en la calle.. Formalmente, se define el vector comparación como un vector función de los registros α(a) y β(b) en la forma:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^{1}[\alpha(a), \beta(b)], ., \gamma^{k}[\alpha(a), \beta(b)]\} \qquad (3)$$

Se ve que γ es una función definida sobre AxB. Se puede escribir γ(a, b), γ(α, β) o simplemente γ. El conjunto de todas las posibles realizaciones de γ se llama espacio de comparación y se denota por Γ.

Durante el proceso de la operación de fusión, se observa γ(a, b) y se tiene que decidir si (a, b) es un emparejamiento, (a, b) Є M (se llama a esta decisión link y se denota por A1) o si es un no-emparejamiento, (a, b) Є U (se llama a esta decisión no-link y se denota por A3). Sin embargo, pueden existir situaciones para las cuales sea imposible tomar una de estas dos decisiones para niveles

específicos de error, por lo que se permite una tercera decisión, denotada por A2, a la que se llama posible link.

En estas condiciones, se define una regla de fusión L como una aplicación del espacio de comparación Γ sobre el conjunto de funciones de decisión aleatorias D = {d(γ)} donde

$$d(\gamma) = \{P(A1|\gamma), P(A2|\gamma), P(A3|\gamma)\};\ \gamma \in \Gamma \qquad (4)$$

y

$$\sum_{i=1}^{3} P(A_i \setminus Y) = 1 \qquad (5)$$

En otras palabras, para cada valor observado de γ, la regla de fusión asigna las probabilidades de tomar cada una de las tres posibles decisiones.

Lo anterior se puede ejemplificar de tal forma que al EMPATAR dos ficheros gráficamente se visualizaría de la siguiente figura:

Figura 2.-Esquematización grafica de posibles resultados obtenidos de la aplicación de la teoría Record Linkage.

Etapas a cubrir bajo una metodología determínistica Record Linkage.

La teoría Record Linkage ha dado apertura a generar diferentes métodos de alternativa para el empate de archivos.

El concepto base de Record Linkage muestra que es posible la comparación entre de ficheros de longitud n, aunque debemos tener cuidado ya que al empatar un archivo de longitud "n" contra otro de longitud "y" el número de comparaciones nos daría (n * y) posibles pares a revisar. Para disminuir el universo de comparaciones es necesario aplicar metodología de Filtering o Blocking, aunque previamente se incorpora una primera etapa al HOMOGENIZAR o ESTANDARIZAR la información contenida en los ficheros y que es parte de la información similar entre ellos.

9.2.2.- El papel de la estandarización de variables similares entre dos ficheros a empatar.

En Instituciones u Organismos Internacionales se ha hecho evidente la necesidad de tomar en cuenta y darle la importancia debida al proceso de homogenizar, estandarizar y corregir el contenido de la información almacenada en sus ficheros, archivos o bases de datos ya que algunas de ellas han aprendido que muchas veces el contener información almacenada errónea puede representar un alto costo.

Según el Instituto Cántabro de Estadística (ICANE[33]) en el cual a la etapa de homogenización o estandarización lo denominan "Normalización" y lo consideran

[33] XVI JECAS, XVI Jornadas de Estadística de las Comunidades Autónomas, SANTANDER del 15 al 18 de Octubre del 2008. Ponencias sobre Registros Administrativos, Directorios e Infraestructura Estadística

como "*un proceso de importancia capital, previo a cualquier tarea de cruce de ficheros de micro datos y en donde se elimina el ruido que contengan los campos.*

En el Cuarto congreso Colombiano de Computación 4CCC[34], en el 2009 en el artículo "Hacia una Metodología para la Selección de Técnicas de Depuración de Datos" del tema "Bases y bodegas de datos, minería de datos" Iván Amón y Claudia Jiméneze, hace referencia a la mención hecha por Andreas Bitterer[35], vicepresidente de investigación de Gartner en 2007 "*No existe una compañía en el planeta que no tenga un problema de calidad de datos y aquellas compañías que reconocen tenerlo, a menudo subestiman el tamaño de éste"*

En este mismo artículo especifican que "*Errores de digitación, datos inconsistentes, valores ausentes o duplicados, son algunos de los problemas que pueden presentar los datos almacenados en las bases y bodegas de datos, deteriorando su calidad y en consecuencia, la calidad de las decisiones que se tomen con base en el nuevo conocimiento obtenido a partir de ellos."*

Dadas las evidencias anteriores no debemos dejar de lado la importancia que tiene la estandarización de variables en nuestro proceso.

Por lo cual, la estandarización de variables similares entre los ficheros que intervendrán en el proceso de EMPATE son uno de los puntos importantes a cubrir y se enfocarán a que una vez realizada esta, permita un primer EMPATE meramente determínistico, es decir información idéntica en el contenido de las variables similares comparadas, así obtener en el primer paso de EMPATE el mayor número de registros posibles con la misma información, disminuyendo el número de registros a empatar en los pasos subsecuentes y permitiendo la

[34] 4CCC. 4 Congreso Colombiano de Computación. http://serverlab.unab.edu.co:8080/wikimedia/memorias/full_papers.html
[35] http://www.gartner.com/it/page.jsp?id=501733

optimización de las consultas o empates directos que se dan en bases de datos relacionales.

La estandarización la dividiremos en tres etapas:

- Homogenización de caracteres dentro de las cadenas. La homogenización de caracteres se define como la sustitución de caracteres no válidos, por válidos, por ejemplo vocales que en algunas ocasiones por error no son capturadas correctamente, o como consecuencia de un cambio del fichero de una plataforma a otra se crean caracteres diferentes al original. Como por ejemplo:

  Á, Ä, Â, À que al ser estandarizado sería A.
  #, @, ¥ que al ser estandarizado sería Ñ.

- Eliminación de caracteres basura.- Son aquellos caracteres que no tienen ninguna relación directa con el contenido de la cadena y que fueron ingresados por error a la cadena, ejemplo:
  , !, ¬, ª,?,!,$,&, % al ser estandarizado debe ser sustituido por un espacio en blanco o ser eliminado de la cadena.

- Eliminación de Cadenas que provoquen “problemas” en el empate. Son parte de la cadena original que al ser muy repetitiva dentro de las cadenas a comparar pueden provocar que la metodología de comparación aproximada generen falsos positivos , es decir posibles empates cuando en realidad no lo son tal es el caso de:

  “CONSTRUCTORA DANA” VS “CONSTRUCTORAS LANAS”

Son cadenas muy similares que al ser comparadas sin eliminar la palabra CONSTRUCTORA tiene mayor posibilidad de considerarse como empate cuando en realidad no lo es.

9.2.3.- El papel que juegan las metodologías de Blocking y Filtering en la disminución de universos a compara entre dos ficheros

Existen técnicas conocidas como Blocking o Filtering, que permiten la reducción de universos de empate entre dos ficheros.

Estas técnicas se implementan teniendo un conocimiento amplio de la información contenida en los ficheros, ya que el éxito de su aplicación en gran parte dependerá de la decisión de cual o cuáles son las variables que intervendrán en esta etapa.

Según Rohan Baxter, Peter Christen and Tim Churchesen su artículo A comparison of fast blocking Methods for record linkage[36], en donde menciona que los métodos de blocking son utilizados en los sistemas Record Linkage para reducir el número de registros pares candidatos a comparar, en este artículo exponen algunos métodos de blocking, filtering y clustering y los resultados de sus experimentos, presentándonos varias opciones que podemos utilizar en nuestra investigación.

El Instituto Vasco de Estadística, denomina al proceso de blocking a la "*Utilización de un esquema que extraiga únicamente pares de registros que sean razonablemente susceptibles de corresponder con un match.*

[36] http://datamining.anu.edu.au/publications/

De tal manera, que se reduce el número de comparaciones que deba llevar a cabo el programa repartiendo los ficheros en bloques mutuamente exclusivos y exhaustivos diseñados para aumentar la proporción de pares de registros a comparar.

Generalmente este se implementa mediante clasificaciones de los dos ficheros sobre una o varias variables."

Aunque los autores no precisan la diferencia entre las metodologías Blocking o Filtering con exactitud, nosotros la diferenciamos como:

Técnicas Filtering o Blocking "A las técnicas que se utilizará para reducir el espacio de comparación entre los ficheros basado en una o más variables similares entre ellos."

- Técnica Blocking a las grandes sub-agrupaciones del universo como un filtro grueso tomando alguna o algunas de las variables de gran escala, por ejemplo ENTIDAD FEDERATIVA, puede considerarse como un filtro grueso aunque muy extenso.
- Técnica Filtering a las agrupaciones o subgrupos de información más finos que viene a complementar a la técnica blocking como por ejemplo el CODIGO POSTAL O UNA PALABRA CLAVE dentro de la cadena a empatar, tal es el caso en la cadena "Restaurantes ELEGANTES" la palabra podría ser ELEGANTES.

En la revisión de la literatura encontramos diferentes técnicas blocking y filtering algunas son:

a- Standard Blocking (SB).

Consiste en la partición de ficheros en subconjuntos de registros mutuamente exclusivos, en donde la llave para la formación de los bloques se hace a partir de uno de los atributos similares dentro de los ficheros que se van a empatar, o bien mediante la combinación de dos o más de ellos, pudiendo ser por ejemplo los primeros n-caracteres de un atributo o la combinación de un atributo más los primeros n-caracteres de otro.

La eficiencia de esta técnica de blocking o filtering va directamente relacionada con la homogenización y estandarización del contenido de cada uno de los atributos de los ficheros sea idéntica en cada uno de ellos ya que de no ser así vamos a caer en los problemas como puede ser el caso de que si en ambos ficheros tenemos el nombre, domicilio, código postal y decidimos para el primer caso de tomar los primeros n-caracteres de el atributo nombre y en el fichero A el contenido es "*Carlos Alvaez Ortha"* y en el fichero B el contenido del atributo nombre es "*Alvares Horta Carlos"* al tomar los 5 primeros caracteres dejamos completamente fuera del bloque de búsqueda a este empate debido a la inconsistencia del contenido de sus atributos.

b.- Q-gram.

En donde la Q permite determinar el tamaño del gran a utilizar para el caso en donde Q=2, se denomina Bi-garm. Consiste en obtener sub-cadenas de 2 caracteres de la cadena original traslapándose estas.

Aunque el q-gram proporciona una forma de filtrado dependiendo del número de sub-cadenas formada de tamaño Q a partir de la cadena original y cada bloque está formado por todos los registros que cumplan con la expresión de

La Sumatoria de la matriz de distancias es menor a C_min .

De manera general, la ejemplificación sería que si tenemos dos cadenas cadena1 que denotaremos C1 y cadena2 que denotaremos C2 cada una de longitud L1 yL2 respectivamente. Aplicando a cada cadena la formación de q-gram en donde para ejemplificar definimos a q=2 sin importar el orden del contenido de cada una de las cadenas.

Para determinar si la cadena C1 es posible empate con la cadena C2 es necesario los respectivos subconjuntos de sub-cadenas de cada una de ellas, estos dos subconjuntos B_{C1} y B_{C2} se incorporan en una matriz de distancias, en donde las dimensiones la matriz por un lado es igual al número de sub-cadenas obtenidas de B_{C1} (filas) y por el otro (columnas) el número de sub-cadenas obtenidas en B_{C2}.

Para obtener la distancia entre las cadenas se comparará todas las sub-cadenas obtenidas en C1 contra todas las sub-cadenas obtenidas en C2, asignando un 1 a las sub-cadenas iguales y 0 en cualquier otro caso. Sumando al final de la comparación la distancia entre las cadenas.

La distancia obtenida en la matriz de distancias es comparada con la expresión

$$C\_min = \max(Longitud\_de\_C1, Longitud\_de\_C2) + q - q * k - 1$$

La determinación del valor *k* depende del conocimiento empírico de los datos. Un valor muy elevado de *k* dejará que muchas parejas de registros que no son empates se filtren, un valor muy bajo de *k* eliminará muchos candidatos a ser empates.

Se busca determinar si las cadenas están dentro de una distancia *k* bajo el siguiente criterio. Si las cadenas están dentro de una distancia *k*, la cardinalidad de $B_{c_1} \cap B_{c_2}$,

Esto significa que se cuenta el número de sub-cadenas comunes en las dos cadenas y éste se compara con el umbral $C_{\min}$ para un *k* particular. Si el número de sub-cadenas comunes en las dos cadenas es más pequeño que el umbral, este par no se considera candidato a ser empate.

Ejemplificando como se construye la matriz de distancias y como se obtiene la distancia entre las cadenas:

C1=" *Carlos Alvaez Ortha"*

C2=" Alvares Horta Carlos

B_{C1} = CA AR RL LO OS / AL LV VA AE EZ / OR RT TH HA = 14

B_{C2} = AL LV VA AE ES / HO OR RT TA / CA AR RL LO OS = 14

Matriz Distancias= Distancia

	AL	LV	VA	AE	ES	HO	OR	RT	TA	CA	AR	RL	LO	OS
CA	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0
AR	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
RL	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
LO	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0
OS	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
AL	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0
LV	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0
VA	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0
AE	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
EZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OR	0	0	0	0	0	0	1	0	0	0	0	0	0	0
RT	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0
TH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HA	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Distancia=10

c.- Clustering.

La clusterización es una técnica que trata de que la o las variables seleccionadas para definir el block o clúster detecten el uso con mayor frecuencia de cierta cadena o parte de la cadena y en base a ellos se formen los grandes bloques, tal vez en n-block agrupar las cadenas que no tiene un peso significativo para formar un gran bloque, esta técnica pretende la división de un n número de blocks pero no deben ser muchos es por eso que se le llama técnica de filtro grueso, con esto se divide el universo de búsqueda para acotar el universo. Se menciona que esta técnica tiene bajos costos de formación del clúster, pero en la aplicación de la técnica de empate puede elevar consideradamente el costo de comparación de las cadenas al ser el universo de búsqueda muy grande y generar un gran número de pares a comparar.

9.2.4.- Metodología ASM para la comparación de cadenas no exactas pero si aproximadas.

Anteriormente se definió que el empate determinístico, es aquel en el cual la comparación entre dos cadenas es idéntico, pero también se especificó que en algunas ocasiones aunque se realice un procedimiento de estandarización de variables por las características de la información esta no es suficiente para realizar un EMPATE totalmente determinístico por lo cual se recurre a una metodología de comparación de cadenas aproximadas tal es el caso de ASM (Aproximate String Matching).

La metodología ASM está basada en el cálculo de la distancia entre una cadena con relación a otra, por el número de errores detectado, entre una para ser igual a la otra, bajo un umbral k de errores permitidos clasificarlo en POSIBLE EMPATE, EMPATE, NO EMPATE. Esta comparación aproximada de cadenas de caracteres está clasificada dentro de las metodologías determinísticas, gira alrededor de medir la distancia entre una cadena la más larga tomada como texto y una segunda cadena llamada patrón la más corta, en longitud y compararlas

determinando la diferencia de una para ser la otra, una metodología mejorada es la implementada por SAMHSA[37] [El Centro de Tratamiento para Abuso de Sustancias (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA), dentro del departamento de Salud y Servicios Humanos de los EEUU (HHS)], el cual además mejoró la metodología adoptando un sistema de penalización de vocales, espacios y caracteres diferentes, concentradas en una matriz de coincidencias, al calificar las comparaciones entre cadenas disminuyendo o aumentando la posibilidad de comparación aproximada. SAMHSA presenta en una sección de su página el trabajo realizado con su Base de Datos Integradora (BDI) Record Linking y sus programas en SAS para la realización del Record Linking[38]. En esta página destaca los programas y la información desarrollada en la creación del BID, sobre la vinculación de datos de múltiples fuentes. Los materiales expuestos incluyen presentaciones, un informe técnico, código de programación en SAS y documentación de apoyo. La metodología utilizada es la de ASM (Aproximate String Matching) y su desarrollo fue realizado en SAS.

9.2.5.- Relación de la Teoría Record Linkage y la Conformación de Empresas de más de un establecimiento bajo la misma denominación de Razón Social y/o Nombre del Establecimiento.

En esta investigación se han analizado dos conceptos principales, la de los directorios de unidades económicas con fines estadísticos en donde las unidades de observación son o pueden ser los establecimientos y/o las empresas y por otro

[37] Heil, S. K. R., Leeper, T. E., Nalty, D., & Campbell, K. (2007). *Integrating State administrative records to manage substance abuse treatment system performance* (DHHS Publication No. [SMA] 07-4268). Technical Assistance Publication (TAP) Series 29. Rockville, MD: Center for Substance Abuse Treatment, Substance Abuse and Mental Health Services Administration.
Whalen D, Pepitone A, Graver L, Busch J.D. Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, July 2000.

[38] http://www.csat.samhsa.gov/IDBSE/idb/modules/linking/recordlink.aspx

lado la teorías Record Linkage, dos conceptos que tal vez a simple vista no se relacionen sino hasta el momento en donde los directorios de unidades económicas se ven en la necesidad de empatar dos o más ficheros que contengan información de las unidades económicas y aunque estos ficheros pueden estar formados con el almacenamiento de información similar en ellos y pueda haber la posibilidad de que la información de un fichero pueda estar en su totalidad o parte de el en otro fichero implicaría la necesidad de empatar ficheros y al presentarse la situación de no contar con un identificador únicos que los relacione uno a uno, la Teoría Record Linkage puede representar una opción viable y muy interesante. Formando con ambos ficheros un solo directorio y a partir de el, la utilización de la teoría en donde busque dentro de el mismo e identifique las agrupaciones de establecimientos que se encuentren bajo una misma denominación de Razón Social o Nombre del establecimiento, detectando así las empresas conformadas por más de un establecimiento bajo la misma denominación.

Se mencionó con anterioridad que nuestra investigación detectó que los directorios de unidades económicas que se registran dentro del INEGI son la de el directorios de unidades económicas a nivel establecimiento que lo definimos como DNUE y a partir de el se formará el Directorio de Empresa, este último conformado por dos partes la de unidades económicas o empresas denominadas únicos que son un solo establecimiento y las empresas que se conforman por más de un establecimiento, y es en esta parte en donde proponemos la utilización del concepto Record Linkage en donde no funcione solo como una metodología para detectar o relacionar ficheros sino como opción para la conformación de empresas formadas por más de un establecimiento bajo la misma denominación de Razón Social y/o Nombre del Establecimiento.

10.- Metodología de desarrollo adaptada a INEGI del Proceso Unificado de Software RUP[39].

El INEGI, tiene amplia experiencia en el desarrollo sistemas de software para el procesamiento de la información estadística y geográfica que capta la Institución. Esta actividad, llevada a cabo durante años por el personal del Instituto, ha utilizado diferentes métodos para el desarrollo de estos programas.

Al paso del tiempo y con el desarrollo de las Tecnologías de Información y Comunicación, se han presentado nuevas metodologías que, con diferentes propuestas de solución y control de los procesos de desarrollo de software, han sido utilizadas para resolver la sistematización del procesamiento de la información.

En la presente sección se describe la adaptación al INEGI de la metodología Proceso de Desarrollo Unificado de Software o RUP.

Se crearon y diseñaron en el Instituto 17 formatos con la finalidad de que se estandarice el concepto del desarrollo, desde definir los requerimientos del sistema, análisis y diseño hasta su implementación.

Pero que es RUP. (*Rational Unified Process)* es un modelo propuesto por Booch, Jacobson y Rumbaugh, está basado en componentes y utiliza como modelado el Lenguaje Unificado de Modelado (UML). Se basa sobre tres ideas básicas que son: casos de uso, arquitectura y desarrollo interactivo e incremental.

El modelo interactivo debe complementar el ciclo completo o sea una versión del software y cada ciclo se lleva a cabo en cuatro fases que son:

[39] Guía para Desarrollo y Documentación de Software. Plantillas INEGI, Metodología RUP adaptada al Instituto. Junio 2005.

- Gestación o Concepción.- Crea una visión del Software.
- Elaboración.- Se definen la mayoría de los casos de uso, así como la arquitectura del sistema.
- Construcción.- Se construye el software.
- Transición.- El software se mueve de una versión Beta a una de producción.

Puede haber múltiples iteraciones por fase.

La metodología adaptada por el INEGI contempla estas 4 fases y asocia a cada una de ellas las **17 plantillas (Ver Anexo II Ingeniería de Software)** de formatos a desarrollar, a continuación se describe brevemente cada una de estas plantillas y las ubicamos dentro de cada una de las fases de la metodología.

10.1.- Fase de Gestión.

La fase de Gestión contempla las plantillas 1 a la 8.

Su meta principal es establecer de forma conjunta entre los usuarios y el equipo de desarrollo, los objetivos, alcances y términos del proyecto. Para el caso de sistemas ya existentes, la fase de gestación deberá ser muy breve. Esta fase incluirá los formatos:

Plantilla 1.- *Análisis del Negocio o Visión. Definir en alto nivel los requerimientos del sistema en términos de las necesidades del usuario. Descripción breve de lo que se pretende lograr con el sistema a fin de lograr las soluciones del negocio.*

Plantilla 2.- *De Requerimientos. Obtener los requerimientos funcionales y no funcionales a través de entrevistas con el usuario o grupo de usuarios involucrados en el proceso. Los requerimientos funcionales son aquellos que*

representan una serie de acciones en el sistema, mientras que los no funcionales determinan las cualidades del sistema, es decir cómo responde el sistema en la interacción del usuario.

Plantilla 3.- *Lista de Riesgos. Deberá especificar el tipo de riesgo y la magnitud de este. Crear una lista de riesgos a partir de los requerimientos, con las estrategias de mitigación planeadas y con los riesgos priorizados.*

Plantilla 4.- *De Casos de Uso. Identificar y crear el Modelo de Casos de Uso inicial del sistema, a partir de los requerimientos obtenidos. El modelo de casos de uso resultante debe ser validado por el usuario o grupo de usuarios principales.*

Plantilla 5.- *Conformación del equipo de trabajo.- En donde se hace la asignación de roles y determinación de la forma de trabajo*.

Plantilla 6.- *Plan del Proyecto. Basados en la Visión, se construye un plan preliminar del proyecto, tomando en cuenta los recursos disponibles y la experiencia del equipo de desarrollo. Este plan se construye mediante la calendarización de actividades.*

Plantilla 7.- Modelo *de Despliegue.- Este modelo es un tipo de diagrama que sirve para modelar los aspectos físicos de un sistema, tales como son servidores, bases de datos, servidores de aplicación, etc. Esto incluye actividades como son:*

- *Seleccionar el tipo de arquitectura para el sistema.*
- *Crear un diagrama de despliegue detallado, tomando en cuenta los casos de uso más significativos, para la arquitectura, incluyendo componentes de software.*

- *Detallar la arquitectura para satisfacer los requerimientos no funcionales.*
- *Crear el prototipo de arquitectura que demuestre la viabilidad de dicha arquitectura.*
- *Documentar las selecciones tecnológicas.*

Plantilla 8.- *Plan de Interacciones. Esta plantilla tiene la finalidad de determinar la estrategia de desarrollo que permita responder a los riesgos y requerimientos más relevantes de esta interacción.*

10.2.- Fase de Elaboración.

Esta fase abarca de las plantillas 9 a la 11.

El objetivo de esta fase es obtener una arquitectura central del sistema y una definición detallada de los casos de uso más significativos, que provean una base fija, principalmente para la parte en la que se realizan el diseño y la implementación, en la Fase de Construcción.

Plantilla 9.- *Identificación de Clases. Identificar las clases participantes en cada caso de uso. Tomando en cuenta las clases identificadas, hacer el ajuste necesario a la arquitectura propuesta*.

Plantilla 10.- *Diagramas de Clases. Los Diagramas de Clases son el pilar básico del modelado con UML, ya que es utilizado tanto para el análisis como para el diseño. Estos diagramas permiten modelar la estructura estática de los sistemas, no se incluyen flujos de mensajes entre datos.*

Plantilla 11.- *Modelo de Datos. El cual se deriva del Modelo de clases. Este modelo también es conocido como el Diagrama de Entidad–Relación y representa gráficamente la realidad del almacenamiento de la información.*

10.3.- Fase de Construcción.

La fase de construcción esta definida dentro de las plantillas 12 a la 15.

Plantilla 12.- *Arquitectura de Software. Es una vista del sistema, que incluye los componentes principales del mismo, su conducta desde la percepción del resto del sistema y la forma en cómo estos interactúan y se coordinan para lograr el objetivo del sistema. Constituye un puente entre el requerimiento y el código.*

Plantilla 13.- *Plantilla de Pruebas. Una forma de asegurar la calidad del software es la implementación de las pruebas al sistema. Se recomienda que en este documento plasme la información relacionada con el tipo de prueba y los resultados que se esperan obtener.*

En esta metodología se recomiendan los pasos siguientes para construir una prueba:

- *Identificar las operaciones del sistema a partir de los Diagramas de Clases del Sistema.*
- *Para cada operación del sistema construir una prueba.*
- *Se sugiere utilizar la Plantilla de Pruebas.*
- *Elaborar un documento mediante el cual se lleve un control de las pruebas realizadas.*

Plantilla 14. *Seguimiento de Pruebas. Una forma de asegurar la calidad del software y que la información que arrojan las pruebas sea implementada si son necesarios o corregidos los errores encontrados. Visualice la realidad en la que se encuentra el sistema en el momento de la prueba y el objetivo que se persigue pudiendo dar un panorama general, y el alcance logrado.*

Plantilla 15.- *Solicitud de Servicio para Centro de Pruebas. El usuario deberá solicitar con tiempo el servicio requerido, esto con la intención de brindar un servicio de calidad a todos los usuarios y tener la oportunidad de instalar los componentes y software requeridos para el servicio. Esto redituará en que al liberarse el sistema además de su correcta funcionalidad ya habrá sido sometido a pruebas de estrés y accesibilidad.*

10.4.- Fase de Transición.

Las últimas dos plantillas la 16 y 17 son para la fase de transición.

Esta fase se centra en implantar el producto en su entorno de operación. Los objetivos básicos de esta fase consisten en cumplir los requisitos establecidos en las fases anteriores, hasta la satisfacción de todos los usuarios, así como gestionar todos los aspectos relativos a la operación en el entorno del usuario, incluyendo la corrección de los defectos remitidos por los usuarios de la versión beta o por los encargados de las pruebas de aceptación.

Plantilla 16.- *De la Liberación y Aceptación del Sistema. Para la aceptación o liberación del proyecto, se requiere que se hayan llevado a cabo las pruebas del software por las partes implicadas en el proyecto y que se genere la documentación final necesaria para utilizarlo en producción. De esta manera el cliente comprobará si tiene todo lo necesario para efectuar el pase a producción. Si el proyecto ha quedado concluido, se sugiere la elaboración de un documento de Liberación del Sistema, en donde el responsable del mismo, firmará de aceptación de ello.*

Plantilla 17.- *Control de Cambios. Una vez aprobado y aceptado el sistema en su totalidad por parte del usuario, seguramente surgirá la necesidad de nuevos elementos que implicarán nuevos desarrollos para el mismo sistema, por lo que puede servir de gran ayuda, la elaboración de un documento de Control de Cambios, en el cual se especificarán esos cambios, modificaciones o adecuaciones del sistema, de acuerdo a sus nuevas necesidades del sistema.*

11. Estudio de Casos Similares.

11.1 Instituto Vasco de Estadísticas.

A lo largo de esta investigación nos dimos cuenta que esta teoría a servido de referencia a instituciones u organismos internacionales para resolver problemas sobre todo de empate o relaciones entre ficheros o bien para la localización rápida de registros, aunque en esta investigación se detectó que la mayoría de la utilización se hace en archivos en donde se relaciona la mayoría de las veces a directorios de personas, expedientes clínicos, registros postales, censos de población, en muy pocos artículos se hace mención de casos relacionados con directorios de unidades económicas.

El caso del Instituto Vasco de Estadísticas **(**Eustat - Euskal Estatistika Erakundea)**,** en Marzo de 2007 genera un documento al que llamo *"Métodos Automáticos de fusión de registros y su utilización en EUSTAT*" en donde nos presenta un trabajo realizado en el período 2005-2008 bajo el Plan Vasco de Estadística, dirigido a mejora continua y excelencia de los procesos.

Este Instituto menciona que la aplicación de los métodos automáticos de fusión de registros inicio en EUSTAT en los años 90's, con la utilización de métodos determinísticos y posteriormente en el año 2002 se desarrollaron métodos probabilísticos.
En este documento se hace la referencia y descripción de la metodología utilizada que la base de su investigación, fundamentándose en el artículo "A theory for Record Linkage de Iván P. Fellegi y Alan B. Sunter".

Este Instituto desarrolló sus aplicaciones en macros de lenguaje SAS.

Dentro de la sección de metodología nos hace referencia a características específicas de importancia relevante como la creación de listas estandarizadas, en donde definieron criterios de estandarización, se menciona algunos criterios que utilizaron para la creación de los Blocking, cálculos de pesos para la determinación de los match y no match.

En la última sección las aplicaciones que se hicieron en *"Ficheros de Matrimonios y Registro Estadísticos de Población",* con el objeto de fusionar el fichero que contiene la información de los datos personales de los titulares de los matrimonios y el que contiene los datos de cada Unidad Poblacional Básica, es decir, el "*Registro Estadístico de Población, Registros Mercantiles y Directorio de Actividades Económicas"*, el objetivo de esta fusión era poder utilizar la información económica que aporta el Registro Mercantil como fuente de actualización del Directorio de Actividades Económicas, y de esta manera, aumentar la cobertura de la operación, al ser el Directorio de Actividades Económicas el marco posterior de elevación de dicha información, "*Estadística de la Renta Personal y Familiar, Censo de Población y Encuestas de Población"* en relación con la actividad*, "Padrón Municipal de Habitantes de Vitoria-Gasteiz y Registro Estadístico de Población"*, aquí el objetivo es localizar todos los individuos registrados en el Padrón Municipal de Habitantes de Vitoria-Gasteiz en el Registro Estadístico de Población.

Por último en la parte de conclusiones remarca puntos importantes de la utilización de estas metodologías en donde especifica:

"*La conclusión general es que el método de fusión probabilístico es de gran utilidad a la hora de fusionar ficheros que no tienen como variable común una clave identificadora de registro.*

La principal ventaja del método probabilístico respecto al determinista reside en una mayor eficacia en los casos más difíciles, frente a la simplicidad y mayor

rapidez del segundo. Ello implica la eliminación del tratamiento manual en algunos casos y por tanto una reducción de los costos.

Una exigencia del método probabilista es una alta capacidad computacional cosa que hoy en día cada vez es menos importante.

EUSTAT se plantea no sólo seguir trabajando con este método para la fusión de individuos sino también ampliarlo para poder considerar la fusión de ficheros de empresas.

Para ficheros de empresas, se observa que la principal diferencia entre fusionar individuos y fusionar empresas es el tipo de variables de las que se dispone en cada caso. En los registros de empresas, la variable Nombre de la empresa, no puede ser tratada de la misma forma que se hace cuando se trata del nombre de un individuo, dado que la casuística es totalmente distinta. Existe más diversidad en los nombres de empresa y además, la probabilidad de que ocurran errores es mayor, debido a que la complejidad es también mayor. Otra diferencia está en que una variable que aparece con gran asiduidad en los ficheros de empresas y que no se ha tratado en el caso de individuos, es la variable dirección. Esta variable puede aparecer registrada de muchas maneras distintas y su estandarización exigirá un trabajo adicional.

Por otra parte, el hecho de que se haya observado una gran utilidad en los métodos probabilísticos no implica que se dejen de utilizar en el Instituto los métodos deterministas. Debido a que los resultados de fusión dependen en gran parte de la calidad de los ficheros, habrá casos en los cuales sea mejor utilizar otro tipo de métodos en lugar de los probabilísticos. Es mas, puede que para algunos ficheros en concreto la opción ideal sea una combinación de varios tipos de métodos.

Actualmente, EUSTAT está trabajando en la construcción de un módulo de fusión que optimice y facilite su utilización computacional de todos estos métodos para aplicar en cada caso el método más adecuado."

11.2.- División de Investigación de Estadísticas de la Oficina de Censo de Estados Unidos. Investigador William E. Yancey.

Para el caso de la División de Investigación de Estadísticas de la Oficina de Censo de Estados Unidos, el investigador William E. Yancey, estadístico-matemático. Ha desarrollado el programa Bigmath, utilizado para la fusión de ficheros a gran escala, utilizado actualmente para la deduplicación del censo de 2010. Durante los últimos años ha trabajado en investigación y desarrollo de software para la fusión de registros. Es licenciado en matemáticas por el Oberlin College y doctor en teoría de números por la Universidad de Maryland. Es experto en optimización. Es miembro de la Sociedad de Estadística Americana. Este investigador participó en el "XXI Seminario Internacional de Estadística" FUSIÓN DE REGISTROS exponiendo algunas de sus investigaciones como han sido:

- Yancey William, 2007,Big match: a program for extracting probable matches from a large file. Research Report Series. U.S. Census Bureau.
- Yancey William, 2005. Evaluating string comparator performance for record linkage. Research Report Series. U.S. Census Bureau.
- Yancey William, 2004. Improving em algorithm estimates for record linkage parameters. Research Report Series. U.S. Census Bureau.
- Yancey William, 2004. An adaptive string comparator for record linkage. Research Report Series. U.S. Census Bureau.
- Yancey William, 2000. Frequencey-dependent probability measures for record linkage. Research Report Series. U.S. Census Bureau.

12.- Propuesta de Desarrollo.

Esta investigación nació con base en las necesidad de la Dirección Diseño y Marcos Estadísticos más específicamente en Subdirección de Diseño Muestral de Unidades Económicas, en donde una de sus funciones como ya se mencionó con anterioridad, además del diseño muestral de encuestas tradicionales y especiales en unidades económicas, es el de mantener actualizados los marcos de unidades económicas de los cuales se generaban los marcos muestrales de referencia para las investigaciones.

Estos marcos tienen dos unidades de observación establecimientos y empresas, su alimentación de información primaria es en base a la información generada por los Censos Económicos las cuales tiene cobertura geográfica Nacional.

A partir de ellos, especificamos en los sectores de actividades económicas de construcción y de transportes la unidad de observación son las empresas, para el resto de los sectores de actividades económicas la unidad de observación son los establecimientos.

Es preciso aclarar que los sectores de actividades económicos en donde la unidad de observación son los establecimientos, pueden agruparse en más de un establecimiento, aquí es donde se define la necesidad de esta investigación.

Se mencionó anteriormente que en algunas ocasiones no se identifican plenamente todas las unidades económicas que pertenecen a una empresa formada bajo la misma denominación de Razón Social y/o Nombre del Establecimiento, sobre todo cuando al paso de los años se deben actualizar los directorios de empresas formadas por más de un establecimiento con directorios externos de instituciones oficiales o bien por las investigaciones internas del

instituto o resultados de levantamientos de encuestas especiales o tradicionales del mismo, cuando la información de unidades económicas no trae un identificador único que relacione los directorios existentes con algunos registros que contengan la información de registros a incorporar ya sea por mantenimiento o actualización entre ellos es necesario identificarlos por medio de variables similares, a esto se le conoce como empates o bien mach's entre ficheros.

Un procedimiento similar se realiza cuando es necesario identificar las unidades económicas que se encuentran asociadas por encontrarse bajo la misma denominación Razón Social o Nombre del Establecimiento para identificar las empresas formadas por más de un establecimiento.

El realizar esta revisión puede llevar meses de trabajo si se realiza el procedimiento de forma semiautomática o manual sobre todo si los ficheros de los que se hablan son de millones de registros, pues identificar todas y cada una de las empresas que se encuentran bajo la misma denominación de Razón Social y/o Nombre del Establecimiento conllevaba revisiones de pares de registros, donde se revisa una unidad económica contra el resto de las información de unidades económicas para identificar aquellas que se agrupen por la misma denominación.

A ello se aúnan los problemas derivados del almacenamiento de la información como pueden ser:

- El orden de la cadena ALMACENADA EN UN CAMPO O VARIABLE no es igual en un fichero y en otro. Ejemplo si un campo contiene el nombre de la persona en un fichero puede contener: RAUL PEREZ PEÑA, y en otro PEREZ PEÑA RAUL.

- Contenido de caracteres basura por errores de dedo a la captura o por las plataformas en donde se realiza la captura de información. Ejemplo: el

nombre de una compañía en un fichero es: Diseño de Textiles Ramìrez y en otro Dise&o y testiles Ramírez, en apariencia es la misma cadena, pero sabemos que si comparamos estas dos cadenas como caracteres idénticas no se consideran la misma.

- Falta de homogenización en los campos en donde se almacena la información ejemplo: un fichero que almacena información sobre personal lo hace en los campos siguientes: NOMBRES, APELLIDO_PATERNO, APELLIDO_MATERNO siendo que en otro fichero el contenido de la información se encuentra almacenada en un solo campo llamado NOMBRE.

- Diferentes reglas o decisiones del contenido del campo, ejemplo: un fichero contiene el nombre de un negocio “Abarrotes la Perla” mientras que en otro solo es “La Perla”.

Estas características se deben tener presentes cuando se realiza la comparación entre cadenas ya que provoca dificultad o no identificación en la comparación entre ellas, un ejemplo concreto para nuestra investigación sería:

Una primer Empresa identificada con más de un establecimiento tiene en Nombre del Establecimiento: “Hernández Ruiz José Antonio” y en Razón Social: “Dise&o y Construcciones del Norte” y uno de los establecimientos no identificados dentro de la empresa tiene Nombre del Establecimiento: “José Antonio Hernández Ruiz” y en Razón Social: “Diseño y Construcciones del Norte”, a simple vista es la misma pero si no se conoce el comportamiento de estas dos variables o bien no se ha analizado su contenido, si se realiza una comparación idéntica entre campos no se identifica como que esta unidad económica pertenece o se cobija bajo la misma

denominación de Razón Social y/o Nombre del Establecimiento de la empresa y se descarta como parte de la empresa.
Entonces se observó que la identificación de unidades económicas que se encuentra bajo la denominación de Razón Social y/o Nombre del Establecimiento es en su concepto base el empate de registros para identificar las empresas que se agrupan bajo la misma denominación.

La Teoría Record Linkage esta basada precisamente en el empate de registros y de ella se derivan varias metodologías las cuales se implementan formas estadísticas comparativas entre cadenas similares evaluándolas de tal forma que se define si son o no la misma, o bien si tiene un cierto grado de similitud o parecido entre ellas, es por esto que en esta investigación se tomó como base esta teoría.

Se han escrito un sinfín de artículos sobre investigaciones que giran alrededor de esta teoría, asimismo investigaciones y conceptos básicos que ayudan a preparar un entorno más óptimo en su aplicación como son la estandarización u homogenización en el contenido de almacenamiento de la información, en la eliminación de basura o corrección de caracteres dentro de las cadenas de comparación, el Blocking o Filtering llamada así a los filtros gruesos y finos para disminuir los universos de comparación de cadenas, enfocados a minimizar el tiempo de comparación y la disminución de los registros a comparar creando subuniversos del universo total, sobre todo cuando los ficheros que se comparan son muy grandes y por último la utilización de alguna metodología derivada de investigaciones realizada sobre la teoría Record Linkage.

En esta investigación se toman como base teórica el artículo “A theory for Record Linkage de Iván P. Fellegi y Alan B. Sunter” y el documento generado por el Instituto Vasco de Estadística *“Métodos automáticos de fusión de registros y su utilización en EUSTAT*”.

El prototipo del sistema se basa en las investigaciones, programas en SAS y presentaciones de SAMHSA[40] aunque esta no va dirigida a la conformación de un directorio de empresas, si nos da el panorama real de la teoría y la implementación de ella.

Se mencionó anteriormente que los directorios de empresas con fines estadísticos son estudiados e investigados a nivel internacional ya que se reconoce la importancia de ellos, ya sea para la toma de decisiones económicas de una país, establecer marcos muéstrales para encuestas de unidades económica, establecer demografía de unidades económicas basadas en sistemas nacionales de información histórica que pueden ser alimentados a través de los años por censos y/o registros administrativos o bien para la comparabilidad de información económica entre países ya sea que pertenezcan a la misma región geográfica o no, es por ello que no debemos perder de vista las recomendaciones internacionales y los conceptos legales de nuestro país en lo que se refiere a estos conceptos.

Si bien se menciona en las recomendaciones internacionales que es recomendable realizar por lo menos anualmente la actualización de directorios de empresas con fines estadísticos, con las nuevas tecnologías de información y plataformas podría ser posible generar sistemas que permitan la actualización en tiempos más cortos.

El nacimiento del prototipo surge en su inicio con investigaciones enfocadas a resolver el problema de empates y la conformación de un directorio de empresas,

[40] Whalen D, Pepitone A, Graver L, Busch J.D. Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, July 2000.The Center for ubstance Abuse Treatment (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA), http://www.csat.samhsa.gov/IDBSE/idb/modules/linking/recordlink.aspx

buscando metodologías e investigaciones actuales apoyadas en nuevas tecnologías de la información que pudieran servir en la resolución de estos.

Se encontraron artículos e investigaciones que giraban entorno a la resolución de problemas similares, se formó un equipo de trabajo e investigación con profesionistas de las áreas de estadística e informática, en este equipo de trabajo se identificó la página de SAMHSA en donde presenta toda su investigación plasmada en documentos, presentaciones y rutinas en SAS, las que presentan su caso real el desarrollo y parámetros de utilización, a partir de esta investigación y en conjunto con otros artículos se interpretó las rutinas en SAS reprogramándolas e interpretándose en R, donde se hacen adecuaciones y calibraciones, posteriormente a plataforma .NET y SQL server 2005, en donde las adecuaciones y calibraciones siguen mejorándose

En paralelo a las interpretaciones y calibraciones de los algoritmos q-gram y ASM, para crear la primer parte del prototipo, se realiza análisis de información generando tablas de todos los caracteres detectados en las variables candidatas con las cuales sería posible hacer empates dentro de un directorio de unidades económicas las variables analizadas fueron: Razón Social, Nombre del Establecimiento, Calle, Colonia y Teléfono. Partiendo de este análisis se detectan los caracteres contenidos en cada variable, una vez detectados se analizan sus comportamientos dentro de cada una de las cadenas definiendo las primeras reglas de estandarización, esta tabla se presenta en el *Anexo de Estandarización de Caracteres*.

Una vertiente más del análisis se realiza a partir de la detección de todas las palabras contenidas en cada cadena que componen los campos seleccionados, se localiza la existencia de palabras que pueden ser omitidas o eliminadas de las cadenas que pueden provocar que la comparación de cadenas no sea muy óptima debido a su comportamiento, por ejemplo, en la variable Razón Social, las siglas o

abreviaciones SA, S.A., S A y las palabras Sociedad Anónima se eliminarían ya que puede contenerse o no en la cadena, o palabras en la variable calle como: Blv. Boulevar, Bulevar, Blvar. Etc.

Este análisis detecta que para futuras investigaciones se pruebe la eliminación de cadenas analizando estas por sector de actividad económica según su comportamiento, por ejemplo en el sector de la construcción la palabra CONSTRUCTORA, CONSTRUCCIÓN e INMOBILIARIA; ABARROTES, CREMERIA, etc. en el sector comercio, este tipo de palabras pudieran provocar que una comparación entre cadenas no sea tan acertada debido a que la puede traer o no. Como ejemplo: la Razón Social “Constructora La Moderna” y “La Moderna” desde la longitud de las cadenas, al ser comparadas directamente no tiene mucha posibilidad de similitud, aun cuando para el empate se consideren más de una variables con diferentes pesos en las comparaciones de las variables como puede ser la calle y colonia, ya que se vería afectado disminuyendo el porcentaje de similitud entre ellas. En esta investigación se eliminan ciertas palabras similares al los ejemplos antes mencionados, aunque solo se tomo de referencia para su eliminación la frecuencia de aparición en las cadenas y en algunas ocasiones eliminación de abreviaturas como se presentó en el ejemplo de la calle, mencionado con anterioridad.

Otra parte importante que arrojó el análisis de la estandarización fue que el contenido de las variables de Nombre del Establecimiento o Razón Social pudiera estar cambiado ya sea en el orden del contenido de la palabra o bien el contenido de un campo en otro ejemplo:

Nombre del Establecimiento	**Razón Social**
José Rodríguez Urrutia	La Perla
La Perla	Rodríguez Urrutia José
Rodríguez Urrutia José	La Perla
La Perla	José Rodríguez Urrutia

Los ejemplos anteriores nos muestran las diferentes combinaciones que se pueden presentar en el orden del contenido de la cadena por lo que en la parte de herramientas del prototipo se implementa una herramienta para la ordenación alfabética en el contenido de las cadenas, de tal forma que las cadenas en el momento de ser empatadas el grado de similitud sea lo más parecido a la realidad, al mismo tiempo generar cadenas más estandarizadas.

Una herramienta más propuesta en el prototipo, es la combinación de campos de tal forma que genere uno solo. Si se aplicara al ejemplo anterior las dos herramientas propuestas se generaría un campo que en su contenido final sea:

José La Perla Rodríguez Urrutia

Esta cadena para los casos anteriores puede darnos empates más aproximados que sí se realizará con la información inicial.

Aunque estas herramientas se han probado se debe tener cuidado dependiendo del conocimiento de la información, si es o no necesario aplicarlas y en que situaciones.

13. Pantallas Propuestas para el Prototipo del Sistema.

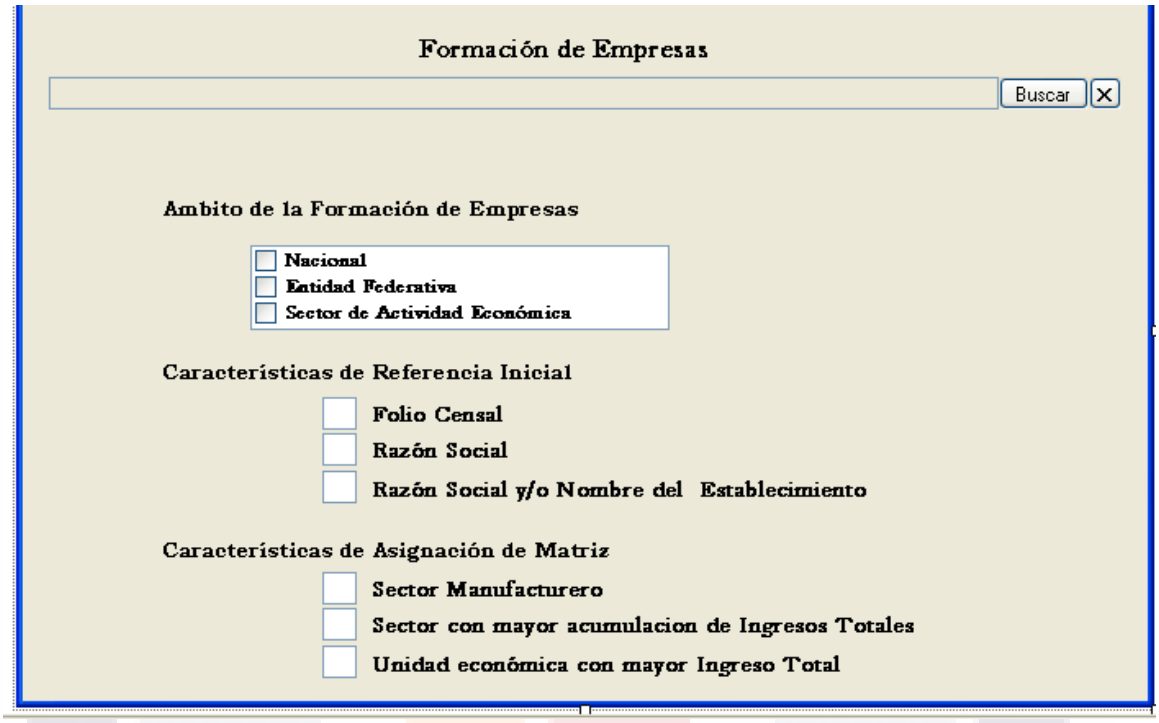


Figura 3. Pantalla Formación de Empresas.

El módulo FORMACIÓN DE EMPRESAS requiere de un primer enlace con el prototipo en el cual se cargar el fichero del directorio de unidades económicas, se especifica que este fichero ya debió de haber pasado antes por la parte de empates, en donde se detectará si existen unidades económicas duplicadas.

El directorio al ser cargado deberá ya contener la estandarización y homologación de variables de acuerdo a la propuesta por censos económicos. En el caso que el directorio que se cargue, no cuente con nombre del establecimiento el prototipo creará una variable en donde el contenido se repetirá a partir del campo de Razón Social.

Después de ser cargado el fichero en procedimientos almacenados se realizaran las siguientes validaciones:

El fichero deberá contener como mínimo los siguientes campos: folio censal, código de la entidad federativa, clave del municipio, razón social, mínimo de codificación del sector de la actividad económica, ingresos totales y total de personal ocupado.

- Ámbito de la formación de la empresa.

Si la información mínima esta completa el usuario podrá definir el ámbito de la formación de la empresa, el cual puede ser:

- Nacional: en donde las empresas se forman no importando su ubicación física dentro del territorio nacional.
- Entidad Federativa, en donde las empresas se forman identificando empresas dentro de una misma entidad federativa creando subdirectorios de empresas por entidad federativa.
- Sector de Actividad Económica. Los subdirectorios o subuniversos de formación son en base a los grandes sectores económicos basados en el SCIAN: que son Sector Construcción, Sector Manufacturero, Sector Comercio y Sector Servicios. En este ámbito, si se desea que algún sector de unidades económicas no deba entrar, deberá eliminarse previamente del directorio antes de ser cargado al prototipo.

- Características iniciales de formación.

Estas características determinaran el orden y prioridad de la formación de la empresa, podrá seleccionar uno de los tres criterios, o bien combinar dos criterios que pueden ser Folio Censal y Razón Social o bien Folio Censal y Razón Social y/o Nombre del Establecimiento.

- Folio Censal: Identificación primaria de la empresa, la variable dentro del directorio deberá primamente agrupar a los establecimientos que

tengan folio censal y que realicen agrupaciones con más de una unidad económica que contengan el mismo folio.

- Razón Social, Este criterio agrupará las empresas que se encuentren bajo la denominación de Razón Social.

- Razón Social y/o Nombre del Establecimiento. Este criterio permitirá agrupar las empresas primeramente bajo la denominación de Razón Social y después de Nombre del Establecimiento

❖ Características de criterios para definir la matriz de la empresa.

Esta opción permite definir las características del establecimiento que fungirá como establecimiento matriz de la empresa. Siempre tomando como primer criterio, la variable de tipo de establecimiento en donde se especifica con una M si el establecimiento se declara MATRIZ, S si el establecimiento se declarará SUCURSAL y una U, si el establecimiento se declara ÚNICO. Si esta variable no existiera o contuviera caracteres diferentes, se tomarán los criterios definidos por el usuario, en caso de existir y el usuario declare alguna característica específica con el número de 1 a 3, en donde 1 es la mayor prioridad, se identificarán las matrices, o bien se resolverán casos en donde no exista una matriz única por empresa o bien múltiples matrices.

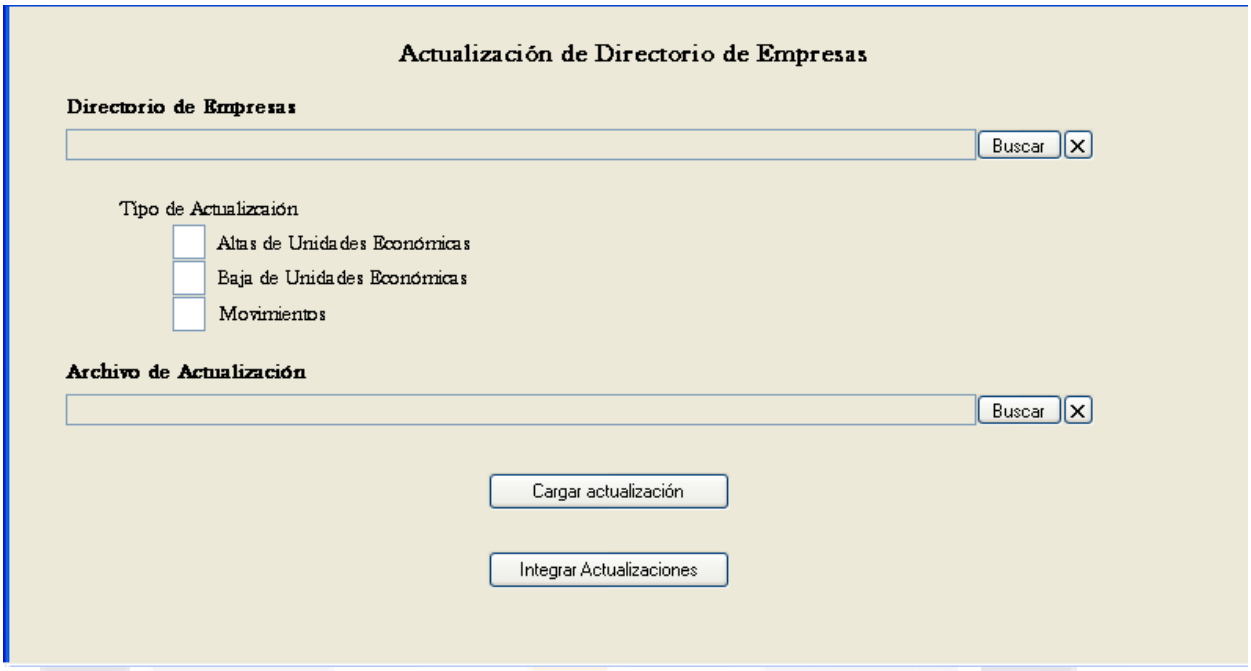
- Sector Manufacturero. Las empresas que en su contenido tengan unidades económicas de sectores diferentes, pero que al menos una de ellas pertenezca al sector manufacturero. Serán analizados los casos siguientes: si el sector manufacturero acumula el mayor ingreso total, la matriz se definirá en ese sector y será la matriz a la unidad económica de mayores ingresos totales. Las empresas que se conformen de diferentes sectores, que contenga el sector manufacturero y no tengan el mayor

ingreso total en ese sector se analizarán por separado por razón social, y se identificará la unidad económica matriz.

- Sector con la mayor proporción de acumulación de ingresos totales. Este criterio tomará la matriz del sector que acumule mayores ingresos totales, y que la empresa en el comportamiento de sus establecimientos contengan más de un sector de actividad económica, pero que no tenga ninguna unidad económica del sector manufacturero.

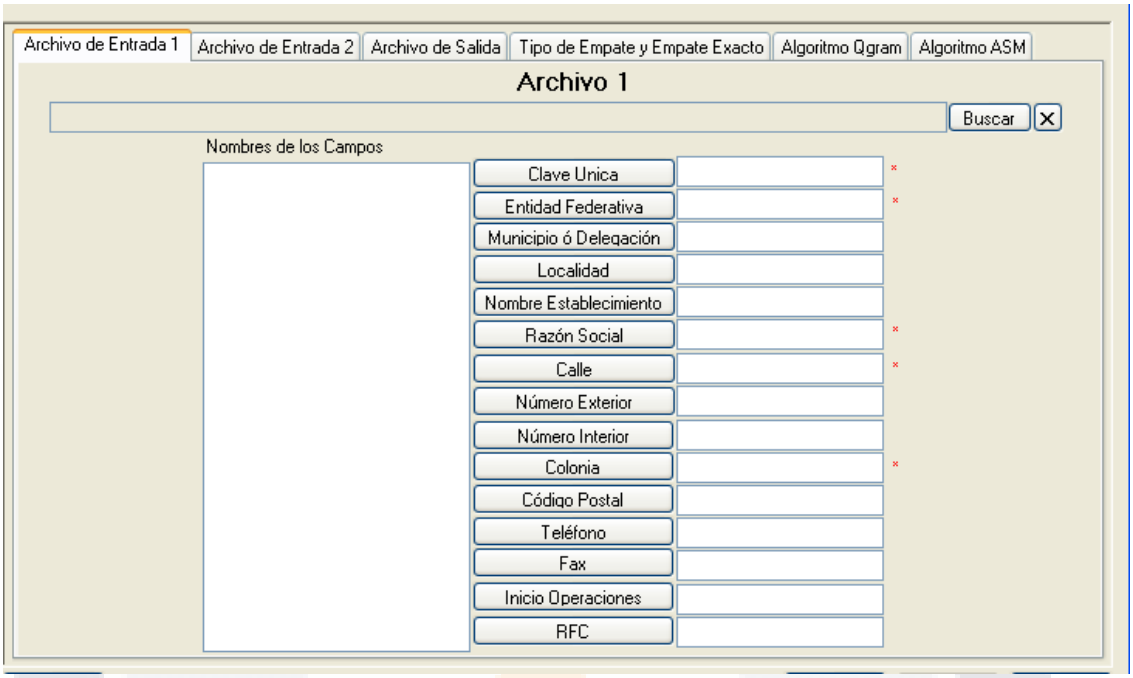
- Establecimiento que dentro de la empresa tenga el mayor ingreso total. La matriz se define dependiendo de la unidad económica con mayores ingresos totales.

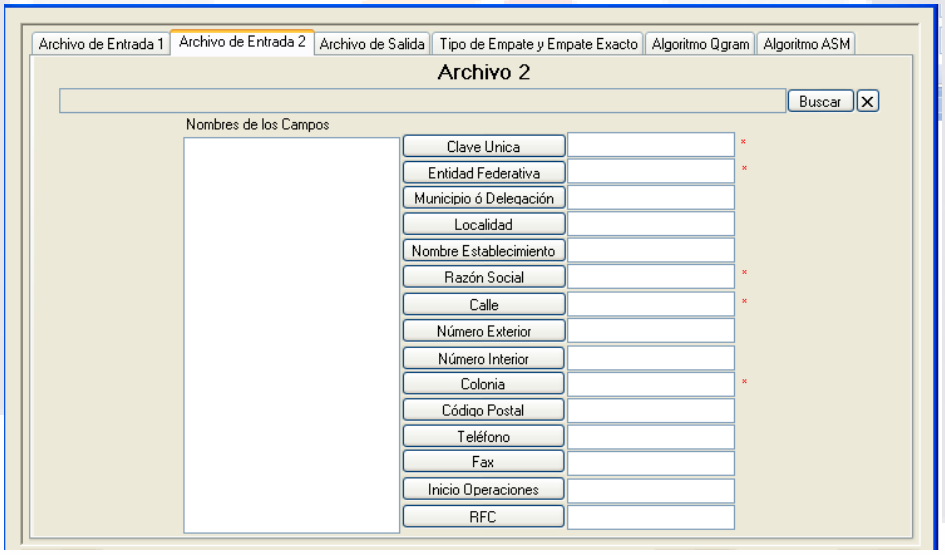
Terminado este proceso deberán de existir tres ficheros, uno de empresas formadas por más de un establecimiento que son los que forman las empresas identificadas por folio, un directorio de empresas únicos y un directorio de establecimientos de las empresas formadas por más de un establecimiento.



La ACTUALIZACIÓN de empresas, esta alimentada por las tablas, altas, bajas, cambios por movimiento y cambios a cualquier campo.

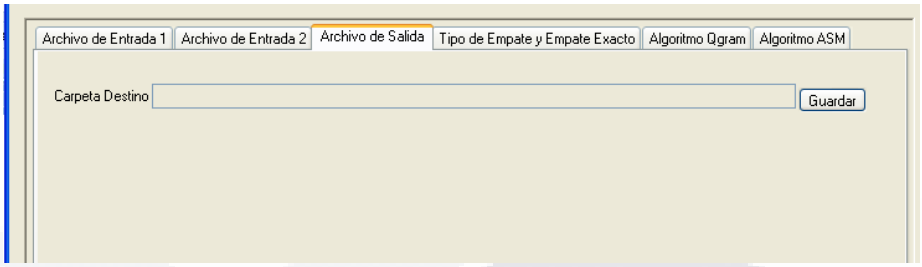
El usuario cargará primeramente el directorio de empresas que se desea actualizar. Especificará que tipo de actualización desea realizar, una vez definida la actualización cargara el archivo de actualización, oprimirá el botón CARGAR ACTUALIZACIÓN para que realice las validaciones correspondientes de el fichero y tipo de actualización, por último si no se presenta ningún error con las validaciones deberá oprimir el botón INTEGRAR ACTUALIZACIÓN, para hacer efectivas las actualizaciones en el directorio de empresas.



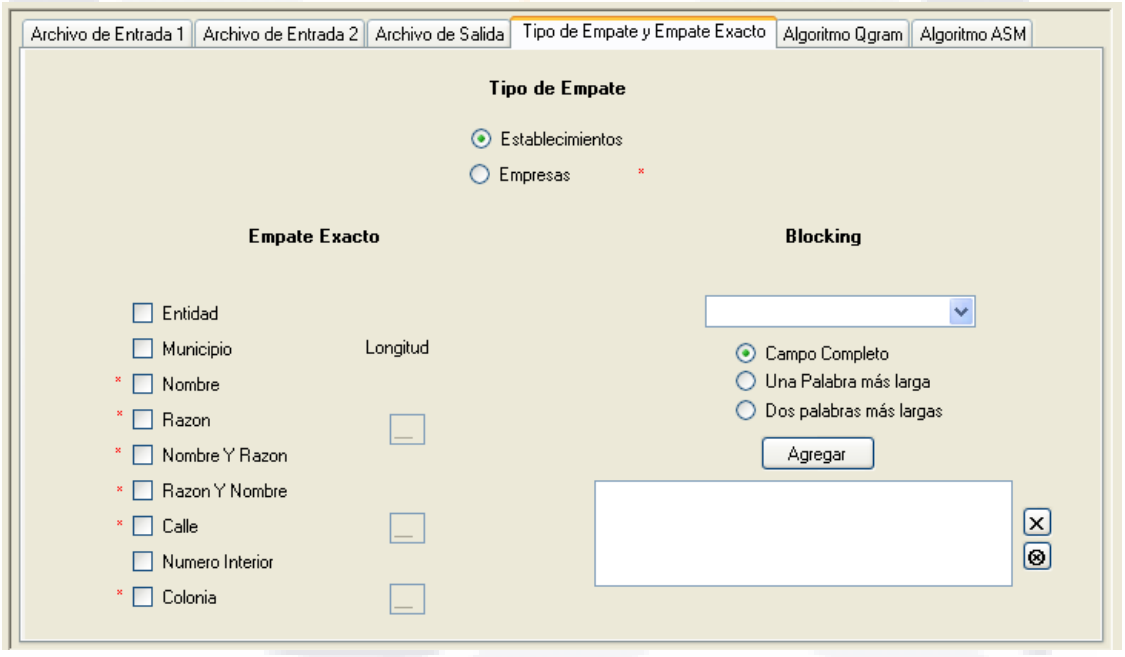


La primera y segunda pantalla de este módulo de HERRAMIENTAS EMPATES permitirá cargar los ficheros que deseemos que se empaten, al cargar estos dos ficheros se identificarán y relacionarán las variables que contengan, con las variables necesarias de contenido similar para la realización del empate. Las

variables que en su lado izquierdo están marcadas con un asterisco * rojo son variables mínimas necesarias para la realización del empate.



La tercera pantalla permitirá al usuario definir una ubicación de almacenamiento del fichero de salida.



En esta pantalla se definen los parámetros del Blocking para el empate, definición del primer subuniverso.

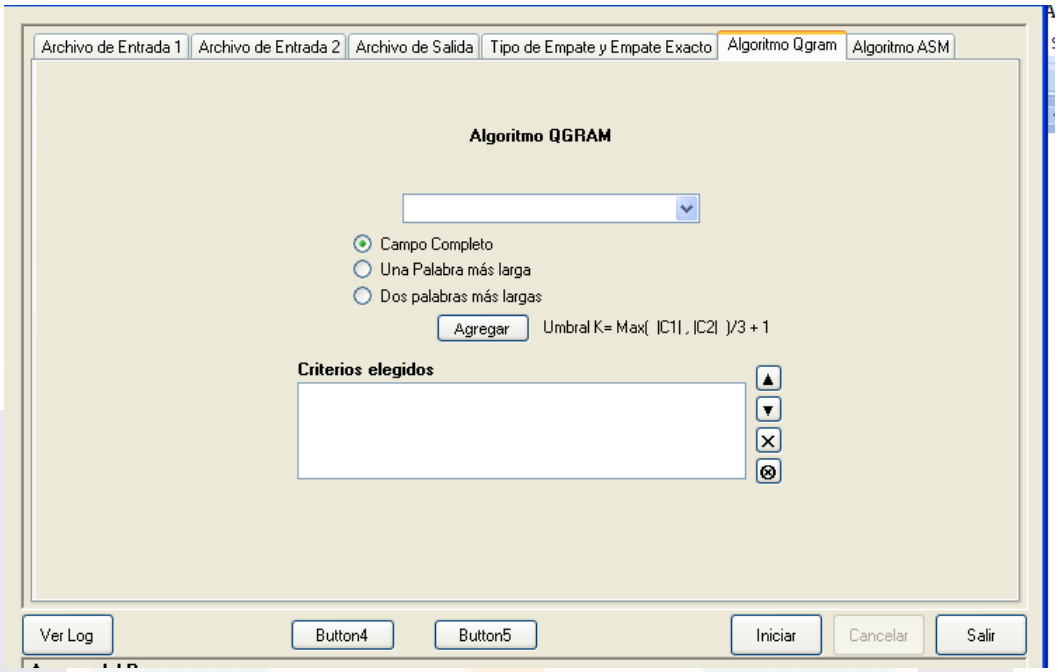
Primeramente definimos si el empate se desea hacer por unidad de observación, establecimiento o empresa, definido este parámetro, definimos las características de un empate exacto, esto es que si existen unidades

económicas exactamente iguales en su contenido se identifique y no sea necesario ser identificadas mediante los algoritmos generados del q-gram y el ASM, se seleccionan las variables para este empate exacto, permitiendo definir en el caso de la Razón Social, de la Calle y de la Colonia, la longitud de las cadenas que se considerarán como exactas.

Se define la variable del Blocking, la cual podrá ser cualquiera que se encuentre en ambos ficheros.

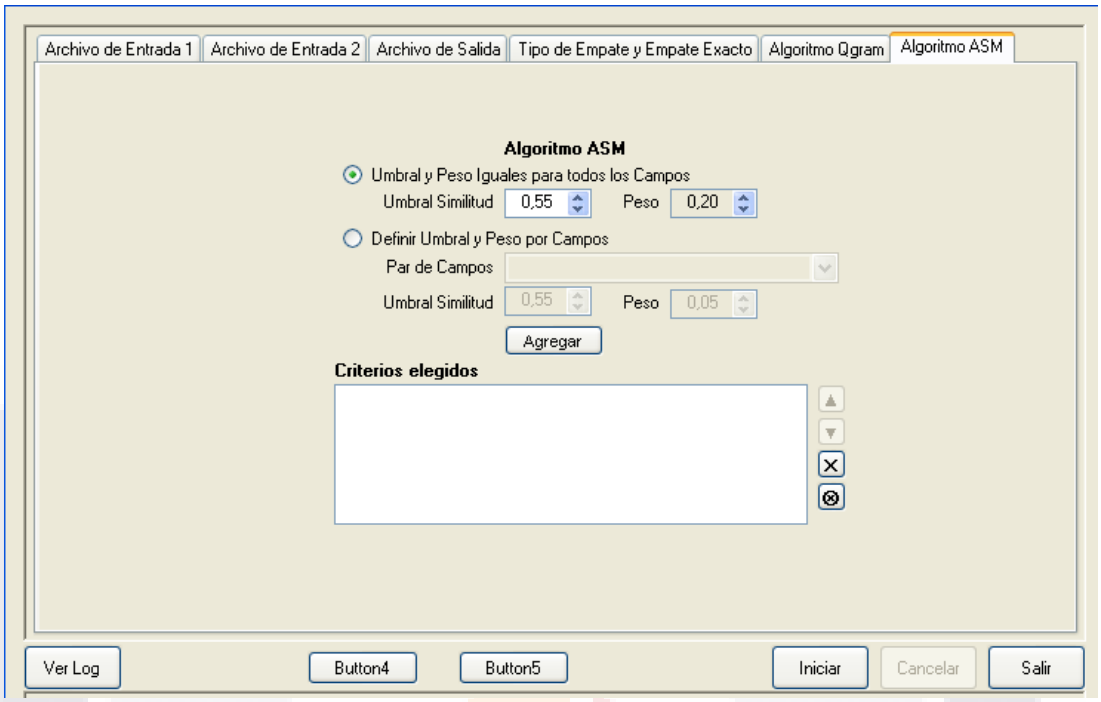
Una vez seleccionada se definirá que parte de este campo se tomará como Blocking, puede ser el contenido del campo completo, la palabra más largas de la cadena o bien las dos palabras más largas de la cadena, se agregan, una vez agregadas el algoritmo irá primeramente al módulo de estandarización, eliminará y corregirá los caracteres considerados como raros, eliminará palabras definidas como problemáticas, y ordenará alfabéticamente el contenido de las cadenas definidas para el empate.

Con las variables y especificaciones del Blocking creará los primeros subuniversos de empate. Y pasará a la siguiente pantalla.



La pantalla de los parámetros del q-gram primeramente definirá el campo o los campos para realizar el filtro fino del empate es decir disminuir más los subuniversos de comparación o empate entre los ficheros.

Se agregan las especificaciones y se pasa a la siguiente pantalla.



En la pantalla del ASM se definirán los parámetros de comparación entre cadenas, la primer parte define un umbral de similitud y peso iguales para todos los campos seleccionados para el empate, dependiendo del número de campos seleccionados para el empate el peso se distribuirá equitativamente para todos de tal forma que sumados los pesos sea de 1.

La segunda opción permite definir el umbral de similitud de las cadenas a comparar, podrá ser diferente para cada una de ellas y los pesos que tendrán las similitudes en las cadenas, estos pesos podrán ser diferentes, pero la suma de todos ellos deberá ser 1.

Se agregan todos los parámetros y se iniciará el proceso de empate, con el botón Iniciar.

Una vez terminado este proceso el prototipo arrojará un histograma con la distribución de frecuencias de los pesos acumulados de las variables empatadas

en donde el usuario tendrá la oportunidad que de acuerdo a esa distribución de frecuencias tomar la decisión de descartar los que considere empates y no empates.

Definido este parámetro se generará el archivo de empates en donde podrá verificar los resultados de ellos.

14.- Respuesta a las preguntas de caso.

- ¿El prototipo propuesto al utilizar el concepto RECORD LINKAGE permite la identificación de unidades económicas bajo la misma denominación de Nombre del Establecimiento y/o Razón Social?

R. Si, el módulo de EMPATE identifica a unidades económicas que se encuentran bajo la denominación de Razón Social y/o Nombre del Establecimiento, aunque la precisión de este dependerá de los parámetros que el usuario defina.

- ¿El prototipo propuesto permite el empate entre directorios que no cuenten con un identificador único que los relacione uno a uno pero que contengan variables con información similar entre ellos?

R. Si, el prototipo realiza mediante variables similares el empate entre dos ficheros aunque no tengan un identificador único, aunque la precisión este dependerá de los parámetros que el usuario defina.

- ¿El concepto RECORD LINKAGE permite la identificación de unidades económicas duplicadas?

R. Por el momento este prototipo solo puede detectar duplicidad de unidades económicas realizando un empate entre el mismo archivo, teniendo que crear una copia idéntica del archivo y empatándolo con el mismo.

- ¿Al utilizar del concepto RECORD LINKAGE facilita la conformación de empresas de más de un establecimiento?

R. En este momento si se considera que facilitaría la conformación de empresas ya que si identifica unidades económicas bajo la denominación de Nombre o Razón Social, pero el algoritmo esta en el proceso de calibración y

con las pruebas realizadas identifica algunas no todas las unidades económicas bajo las denominaciones mencionadas.

- ¿El prototipo generado permite la identificación de empresas bajo la misma denominación de Nombre del Establecimiento y/o Razón Social?

R.- El prototipo esta en proceso de calibración de los parámetros que mejor identifiquen las unidades económicas bajo las denominaciones de Nombre del Establecimiento y/o Razón Social. Por lo pronto cubre la primera etapa de conformarlas bajo los parámetros definidos en la pantalla de “Formación de Empresas”.

15.- Respuestas a las proposiciones.

- ¿El prototipo identificará las unidades económicas bajo la misma denominación de Nombre del Establecimiento y/o Razón Social?

R.- El prototipo si identificará las unidades económicas bajo la denominación de Nombre del Establecimiento y/o Razón Social, aunque esta exactitud dependerá de los parámetros que se definan en la calibración final del algoritmo.

- ¿El algoritmo RECORD LINKAGE propuesto permitirá el empate entre ficheros que no tengan un identificador único que permita relacionarlos uno a uno?

R.- Si, el algoritmo Record Linkage propuesto utilizando de base las investigaciones de SAHMSA, permitirá el empate entre dos ficheros que no tengan un identificador único que les permita relacionarlos uno a uno, este empate se dará en la pantalla de Herramientas. Empates.

- ¿El prototipo permitirá la identificación de unidades económicas duplicadas?

R.- El prototipo en este momento pudiera identificar unidades duplicadas si se hace una copia exacta del fichero, en el que se desea identificar duplicados y empatarlo con el mismo, de esta forma, aunque como se ha mencionado anteriormente los parámetros que se den como entrada en el módulo de empate definirán la exactitud de él, además de que las calibraciones del algoritmo en este momento no han sido totalmente definidas.

- ¿El concepto Record Linkage facilitará la conformación de empresas formadas por más de una unidad económica?

R.- Si, siempre y cuando se calibre el algoritmo para encontrar los parámetros más ad-hoc que permitan identificar sino en 100% de las unidades económicas si, el mayor número de unidades económicas bajo las denominaciones de Nombre del Establecimiento y/o Razón Social.

- ¿La utilización del método RECORD LINKAGE permitirá la incorporación y localización de la mayoría de los establecimientos que conforman una empresa independientemente de las fuentes de las que se recabaron su información para mantener actuales los datos de las empresas?

R.- En esta investigación se han realizado pruebas con diversos parámetros y hasta este momento se han obtenido resultados positivos, pero no se ha logrado calibrar el algoritmo al 100% y por el momento solo se han realizado pruebas con ficheros provenientes de una sola fuente, queda para futuras investigaciones realizar pruebas para verificar si las calibraciones serían las mismas al utilizar ficheros de diferentes fuentes.

16.-Alcance de los objetivos.

El objetivo General se planteó como el de generar un prototipo de sistema que tomaría como base la Teoría Record Linkage y a partir de la implementación de ella en el prototipo se detectarían duplicados dentro de un directorio, se realizarían empates entre el DNUE y directorios externos e identificarían unidades económicas bajo la misma denominación de Razón Social o Nombre del establecimiento, este objetivo general fue alcanzado mediante la generación del prototipo desarrollado en .NET, Visual Basic 2005 conjuntamente el almacenamiento de las bases de datos en SQL Server 2005. El prototipo esta compuesto de 3 módulos principales:

- Formación de Empresa
- Actualización de Empresas
- Herramientas (Empates)

La aplicación de la Teoría Record Linkage, se desarrollo en dos ensamblados el primero con la implementación del algoritmo q-gram y el segundo el algoritmo ASM.

En cuanto a los objetivos específicos, hasta este momento las unidades económicas solo es posible identificarlas mediante el módulo de Herramientas, Empates, en donde se empatara el directorio que se desea analizar.

El objetivo de identificar unidades económicas bajo la misma denominación de Razón Social y/o Nombre del Establecimiento, en este momento estamos en la etapa de calibración si fue alcanzado, aunque no podemos todavía afirmar en que porcentaje logra identificar el total de las unidades económicas

El prototipo en este momento si permite el empate entre dos directorios que no tienen un identificador único que los relacione mediante variables o campos con información similar, debido a que la teoría es flexible permitiendo definir los parámetros que el usuario desea establecer para la realización del empate, la identificación de las unidades económicas dependerá del conocimiento de la información del usuario del prototipo.

El prototipo se implementa con el uso de un algoritmo basado en la teoría Record Linkage los cuales son el Q-gram, para este prototipo el bi-gram y el ASM (Aproximate String Match), se especifica que en este momento el prototipo se encuentra en la etapa de calibración de los parámetros para encontrar los más óptimos en la identificación de unidades económicas bajo la misma denominación de Razón Social y/o Nombre.

17.- Conclusiones.

Esta investigación muestra la importancia que tiene los ficheros que guardan información de directorios de empresas y/o establecimientos con fines estadísticos, además muestra un panorama muy general de la importancia de incursionar en metodologías que permitan relacionar estos ficheros de tal forma que no sea una limitante para la actualización y mantenimiento de ellos. Que no importe que la alimentación de información no provenga de una sola fuente y que no importe que no se tenga un identificador único que relacione los ficheros.

Se inicio en el estudio del concepto Record Linkage que aunque muy amplio, ha sido bien adoptado por múltiples instituciones y organismos internacionales, que le han ido dando diferentes usos y fines, mostrándonos que vale la pena incursionar y profundizar en él, aunque también encontramos que esta relacionado con otros conceptos que permiten lograr mejores resultados como la investigación de la corrección y estandarización de cadenas.

Al realizar revisiones de artículos o documentos en donde algunas instituciones nos plantean sus investigaciones, se observó que el concepto no es tan simple como a primera vista se podría apreciar, algunas de estas instituciones o investigadores del concepto han escrito más de un artículo sobre ello y con el crecimiento y mejor conocimiento del concepto van proponiendo mejoras y nuevas metodologías.

Esta investigación ha concluido con un prototipo de un sistema en donde se adopta una de las variantes de la metodología basada en el concepto Record Linkage, para la conformación de empresas, siendo un prototipo, el algoritmo ya se encuentra desarrollado en su parte fundamental, queda para el futuro próximo calibrar la metodología de tal forma que pueda culminar en un sistema que se

alimente del directorio de unidades económicas y nos arroje las empresas ya conformadas.

Áreas de conocimiento vistas en la maestría que se utilizaron en la propuesta de este prototipo:

- Análisis y Diseño de Sistemas en conjunto con Ingeniería de Software.- en la recolección de requerimientos, desarrollo y planteamiento del prototipo, en el área de la ingeniería de software se adoptó el proceso de desarrollo RUP adaptado por INEGI.

- Bases de Datos.- Para el análisis de las variables contenidas en ficheros, así como la manipulación de ellos, Se adopta para este prototipo el almacenamiento de la información en SQL Server 2005.

- Control de la Función Informática.- Para la visualización de el planteamiento de las estrategias a seguir para alcanzar el objetivo de generar el prototipo, planeando los diferentes escenarios que se conocían en base a los objetivos que se deseaban alcanzar o que cubriera el prototipo de tal forma que en el desarrollo del prototipo se tuviera un mínimo de contratiempos posibles.

- Seminario de Tesis I.- Base para la documentación, control y administración de la investigación en la que se basó el desarrollo y propuesta de este prototipo.

Lecciones aprendidas profesional y personalmente en la realización de este proyecto.

Profesionalmente considero que el crecimiento fue a grandes pasos, debido al área en la cual me desarrollo la cual es 100% estadística, no se tiene acceso directo a las investigaciones relacionadas con las actuales tecnologías de la información o bien con las que el instituto esta trabajando o implementando, y considero que independientemente del área, la función o especialización en la que nos desarrollemos profesionalmente, la necesidad de utilizar y conocer herramientas informáticas o tecnologías de la información que apoyen y soporten la implementación o propuesta de nuevos proyectos, mejoras en procesos ya establecidos siempre debe ser una de las prioridades en cualquier área del Instituto.

Personalmente considero que esta maestría da una visión amplia de cómo podemos impulsar y mejorar nuestro desarrollo profesional aplicándolo a las áreas no solo personales sino profesional y laborales, además que el crecimiento como seres humanos que siempre da el obtener nuevos y actuales conocimientos que vayan de la mano con las situaciones modernas y globales. Sobre todo en los tiempos en los que vivimos ya que las generaciones actuales y futuras basan sus conocimientos en las tecnologías de la información.

La Maestría en Informática y Tecnologías Computacionales, fue de gran valor en el desarrollo de este prototipo debido a que me mostro el mejor camino para su logro.

18.- Recomendaciones y Futuras Investigaciones.

Se ha mencionado con anterioridad que ésta investigación finalizó en un prototipo de un sistema, no concluyendo con la conformación satisfactoria al 100% de las empresas utilizando el concepto record linkage, para futuras investigaciones y para la culminación del prototipo en un sistema, se recomienda profundizar más sobre conceptos de corrección de cadenas o limpieza de datos, tal vez por alguna metodología para detección de duplicados, outliers o valores atípicos, un breve panorama sobre este concepto lo podemos encontrar en el paper "Towards a methodology for selection of data cleansing techniques", de Ivón Amón Uribe & Claudia Jiménez Ramírez, Grupo de Investigación y Desarrollo en Inteligencia Artificial – Universidad Nacional de Colombia, en donde mencionan algunas metodologías, propuestas y recomendaciones al respecto.

En específico a la metodología utilizada, es necesario realizar las pruebas suficientes con un mismo directorio y pruebas entre ficheros de diferentes fuentes de tal forma que nos permita calibrar lo más correctamente posible el algoritmo.

Se hace del conocimiento que como propuesta para la creación de tablas de palabras posibles a eliminar es recomendable se realice una revisión de palabras por sector de actividad económica de tal forma que junto con las calibraciones se pueda determinar si las palabras que se eliminan estar relacionadas directamente con el sector de actividad económica.

Un punto importante de las calibraciones es que se recomienda que se realicen pruebas con las letras que se penalizan, el algoritmo en estos momentos solo penaliza las vocales, espacios en blanco, y consonantes, es recomendable realizar pruebas con letras que debido al origen de nuestra lengua son utilizadas

de muy diversas formas como son las letras c,s,z,x / v,b,w / y,ll / k,c / j,x, ya que en algunas ocasiones se utilizan como similares.

Palabras clave: Conformación de Empresas, Método Record Linkage, Ingeniería de Software, Prototipo, Directorio de Unidades Económicas.

19.- Glosario de Términos

DIRECTORIO DE UNIDADES ECONOMICAS. Conjunto de registros que guarda información de directorio y de las principales variables económicas a nivel de unidad económica.

EMPATAR: Proceso de relacionar dos ficheros mediante variables similares en contenido cuando no cuentan con un identificador único que las relacione.

EMPRESA: es la unidad económica y jurídica que bajo el control de una sola entidad propietaria o controladora se dedica a la realización de actividades económicas y que puede estar constituida por un solo establecimiento, o por varios establecimientos que operan bajo la misma denominación o razón social.

ESTABLECIMIENTO: es la unidad económica que en una sola ubicación física, asentada en un lugar de manera permanente y delimitada por construcciones e instalaciones fijas, combina acciones y recursos bajo el control de una sola entidad propietaria o controladora, para realizar actividades de producción de bienes, maquila total o parcial de uno o varios productos, la compra-venta de mercancías o prestación de servicios, sea con fines mercantiles o no.

ESTABLECIMIENTO MATRIZ. Es el establecimiento de la empresa en que se centran las decisiones acerca de la utilización de los recursos financieros, de la planeación y del control de las diferentes actividades por llevarse a cabo en los distintos establecimientos que de él dependen. En este tipo de establecimientos pueden coincidir o no las actividades administrativas o contables.

ESTABLECIMIENTO SUCURSAL. Es el establecimiento que pertenece a una empresa, sin ser éste la oficina matriz u oficina central

ESTABLECIMIENTO ÚNICO. La empresa cuenta con un solo establecimiento para el desarrollo de sus actividades.

TIPO DE UNIDAD ECONÓMICA. Puede ser un establecimiento Sucursal. Matriz o Único.

UNIDADES ECONOMICAS[41]. Son las unidades estadísticas sobre las cuales se recopilan datos, se dedican principalmente a un tipo de actividad de manera permanente en construcciones e instalaciones fijas, combinando acciones y recursos bajo el control de una sola entidad propietaria o controladora, para llevar a cabo producción de bienes y servicios, sea con fines mercantiles o no. Se definen por sector de acuerdo con la disponibilidad de registros contables y la necesidad de obtener información con el mayor nivel de precisión analítica.

VARIABLES ECONOMICAS.- Variables que guardan información característica de una unidad económica y es propia de ella. Como es el Personal ocupado, ingresos, valor de la producción, valor de activos fijos, etc.

[41] Metodología de los Censos Económicos 2004

ANEXO I

TALLER "DIRECTORIOS DE EMPRESAS Y ESTABLECIMIENTOS: DESARROLLOS RECIENTES Y DESAFÍOS ACTUALES Y FUTUROS EN AMÉRICA LATINA"

Tema 1: Marco Conceptual y metodológico

El Directorio Central de Empresas del INE (DIRCE); marco conceptual y principales características. Esteban Barbado, Subdirector General Adjunto, INE de España.

TEMA 2: Estrategias de mejoramiento del Directorio

- *Mantención y mejoramiento de la calidad del Directorio*
- *La experiencia de los Estados. Eddie Salyers, Assitant Division Chief y Eli Serrano, Small Business Ombudsman, United States Census Bureau.*
- *Business register.*
- *La experiencia de México. Ana María Landeros, Directora de Diseño y Marcos Estadístico, INEGI, México*
- *La experiencia de Chile. Oriana Villegas, Coordinadora Proyecto Directorio, INE, Chile.*
- *Proyecto directorio nacional de empresas y establecimientos documentación marco muestral anual.*
- *La experiencia de Brasil. Ana Rosa Ribeiro, Gerente do Cadastro de Empresas, IBGE, Brasil.*
- *Cadastro central de empresas do IBGE – CEMPRE. Relatório metodológico.*

TEMA 3: Usos del Directorio

- *Marcos muestrales, demografía empresarial e indicadores coyunturales. Hugues Picard, INSEE, Francia*
- *La experiencia de los Bancos Centrales en el desarrollo y utilización de los Directorios:*
- *Cesar Guerrero, Coordinador de la Unidad de Análisis y Procesamiento, Banco Central de Nicaragua.*
- *Marlen Uraña, Jefe de División, Banco Central de República Dominicana.*
- *René Luengo, Jefe de Grupo, Cuentas Nacionales, Banco Central de Chile*

TEMA 4: Resultados y Análisis de la Encuesta aplicada a países seleccionados de América Latina.

- *Caracterización en Países de la Región. Análisis de resultados.*
- *Presentación de los resultados de la Encuesta. Vicenta Mardones y Mauricio Ponce Consultores, CEPAL.*

Documentos Complementarios:

- *ARGENTINA (INDEC): DINUE: Directorio nacional de unidades económicas*

- *BOLIVIA (INE): Directorio central de empresas de Bolivia-DIRCEMBOL (primera fase)*
- *COLOMBIA (DANE): Directorio de Empresas.*
- *COSTA RICA (INEC): Proyecto Directorio de Unidades Institucionales y Establecimientos Resumen Ejecutivo*
- *ECUADOR (INEC): El Directorio de Empresas y Establecimientos de Ecuador.*
- *EL SALVADOR (DIGESTYC): Directorio de empresas y establecimientos de El Salvador.*
- *PARAGUAY (DGGEEC): Directorio de empresas industriales.*
- *REPÚBLICA BOLIVARIANA DE VENEZUELA (BANCO CENTRAL): Actualización del marco estadístico de empresas y locales*
- *REPÚBLICA BOLIVARIANA DE VENEZUELA (INE): Directorio de Empresas y Establecimientos.*
- *REPUBLICA DOMINICANA (ONE): Directorio de Empresas y Establecimientos (DEE) - Construcción.*

ANEXO II

INGENIERIA DE SOFTWARE, RUP INEGI

El presente anexo muestra en la primer parte Figura 3, el diagrama de contexto del prototipo propuesto para el sistema de Conformación de Empresa a partir de un Directorio de unidades Económicas.

En la segunda parte muestra las 17 plantilla propuestas por el INEGI para el desarrollo de sistemas basado en la metodología RUP, adaptada.

Primera Parte

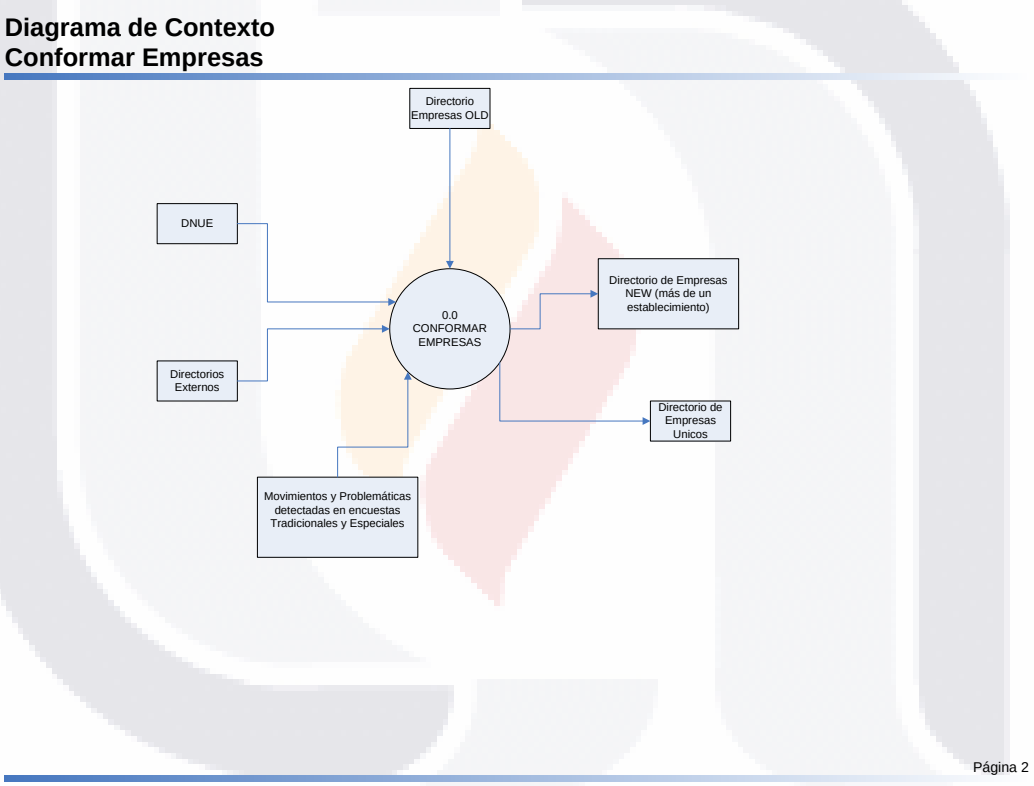
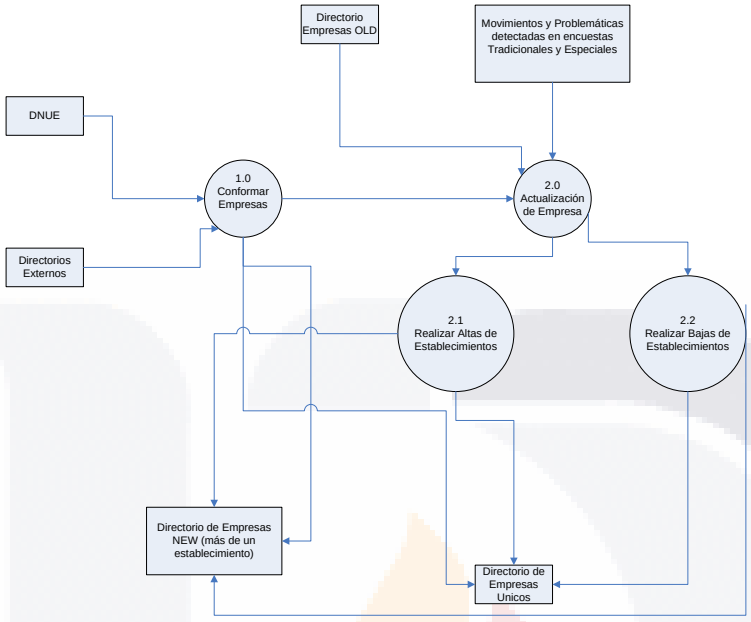


Figura 3. Diagrama de contexto, Conformar Empresas

Nivel 0 Conformar Empresas



Página 3

Figura 4. Diagrama Nivel 0, Conformar Empresas

Nivel 1 Conformar Empresas
Proceso 1.0

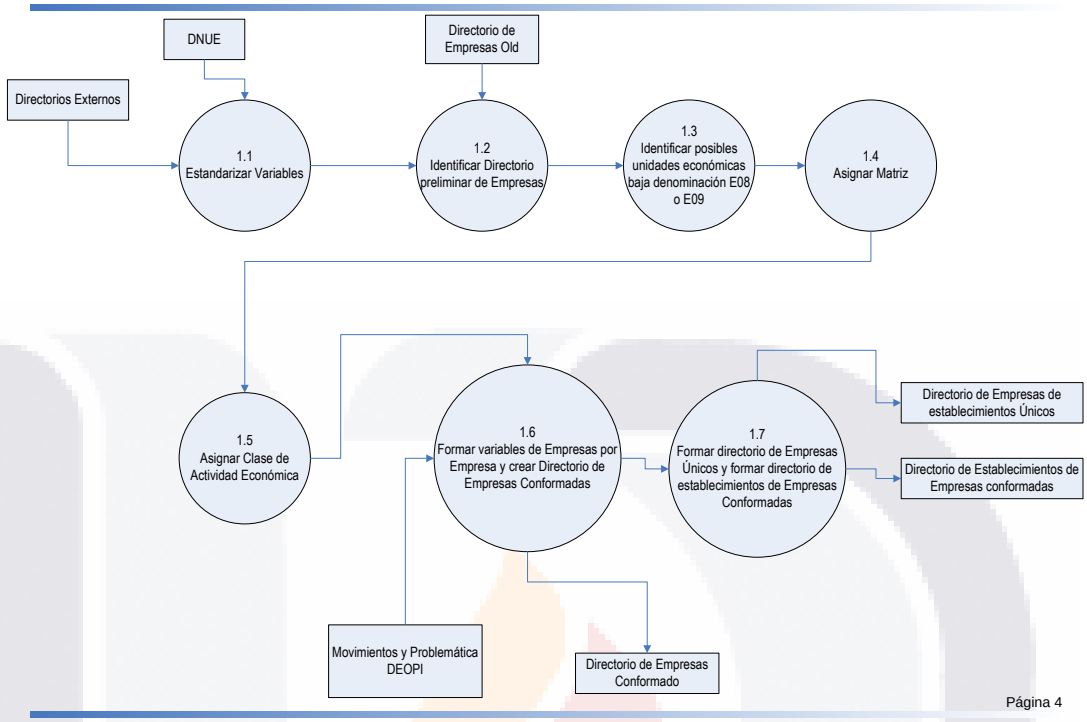


Figura 5. Diagrama Nivel 1, Conformar Empresas

Segunda Parte

Plantillas Metodología RUP INEGI

Plantilla 1.-

<Prototipo del Sistema SDNE>

1. ANÁLISIS DEL NEGOCIO O VISIÓN

Versión <1.0 Beta>

Fecha de última actualización <16/03/10>

1. INTRODUCCIÓN

Propósito

El presente prototipo propone un sistema el cual, a partir de un directorio de unidades económicas conforme empresas bajo la denominación de Razón Social y/o Nombre del Establecimiento tomando como primer agrupación en caso de existir, el Folio Censal y posteriormente la denominación de Nombre del Establecimiento y/o Razón Social.

Tratando de que se conformen en su mayoría de forma automática minimizando la intervención de las revisiones manuales por parte del usuario

1.2 Alcance

El prototipo propuesto propone la utilización de la teoría Record Linkage (Empate de Registros) para la conformación de Empresas bajo la denominación de Razón Social y/o Nombre del Establecimiento.

Esta Teoría se basa en el empate de registro perteneciente a uno o varios ficheros mediante variables por lo general de tipo carácter que no contiene un identificador único que los relacione uno a uno, sino que el empate lo hace por medio de las variables similares entre los ficheros.

La idea es que en base a esta teoría se detecte la mayoría de las unidades económicas que se puedan agrupar bajo la denominación de Razón Social y/o Nombre del establecimiento, por lo que propongo la adopción de un algoritmo automatizado que minimice las revisiones manuales.

Este prototipo solo se limitará a la parte probabilística de la teoría. Y toma como base y referencia el uso de la teoría en el desarrollo del sistema de SAHMSA[42] tomando como base su experiencia y algoritmos propuestos.

2. SITUACIÓN ACTUAL

Descripción del problema

Actualmente la formación del Directorio Nacional de Empresa, se realiza en base a los siguientes pasos del procedimiento:

2.1.- Se agrupan la empresa por Folio Censal, el cual es detectado y registrado en el Censo Económico dando como referencia las unidades económicas que pertenecen a una empresa conformada por más de un establecimiento, es decir identifica la unidad económica matriz y las unidades económicas sucursales, las relaciona e identifica por medio del Folio Censal.

[42] Whalen D, Pepitone A, Graver L, Busch J.D. Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, July 2000.

2.2.- Una vez detectadas las empresas se toma la Razón Social y el Nombre del Establecimiento de la unidad económica matriz para que bajo estas variables se busquen todos los establecimientos que puedan ser cobijadas bajo estas denominaciones, esta parte se hace de forma manual o semiautomática, se apoya en aplicaciones generadas por los usuarios que han realizado de forma intuitiva y con los conocimientos del proceso de esta actividad apoyados en las funciones que contenga la plataforma en la que se esté trabajando.

2.3.-Una vez localizadas todas las unidades económicas pertenecientes a esa empresa bajo las denominaciones de la unidad económica matriz, se procede a identificar asociaciones de unidades económicas que no formaron parte de la empresas ya identificadas y que además no se agruparon por Folio Censal, pero que se pueden agrupar como empresas conformadas por más de un establecimiento ya que entre ellas, es posible la asociación bajo las denominaciones de Razón Social y/o Nombre del establecimiento. Las asociaciones por denominación deben ser primeramente bajo la denominación de Razón Social y después Nombre del Establecimiento.

2.4.-Iidentificadas todas las empresas conformadas por más de un establecimiento, se revisa la variable Tipo de Establecimiento. Identificación de las actividades de la unidad económica en donde define Matriz o Sucursal. En teoría debería estar identificada una matriz por cada una de las empresas conformadas por más de un establecimiento.

Las empresas que tienen solo una matriz, será esta la que se queda en representación de la empresa y pasa a formar parte del Directorio de Empresas, solo que sus variables económicas deberán reflejar la sumatoria de todos los establecimientos que la conforman.

2.5.- Del resto de las empresas conformadas por más de un establecimiento se identifican aquellas que tienen más de un establecimiento como matriz y en otro grupo las empresas que no cuentan con ninguna unidad económica identificada como matriz.

2.6.- Para el grupo de empresas conformadas por más de un establecimiento y con más de una unidad económica identificada como matriz. Se analizan las unidades económicas matriz y una de las opciones para designar la unidad matriz que identifique la empresa puedes ser en base a aquella unidad económica que contenga la mayor cantidad de Ingresos Totales, es posible que la identificación de la unidad económica matriz pueda establecerse en base a diferentes criterios, necesidades o lineamientos.

2.7.- Del grupo de empresas conformadas en las calles no existe identificada ninguna unidad económica como matriz, se realiza un análisis y se identificarán los lineamientos para la asignación de una unidad económica que funja como matriz, puede ser por el ingreso total mayor de las unidades económicas o por alguna otra especificación de clasificación de actividad económica.

2.8.- Un aspecto que se debe cuidar sobre una empresa conformada es la de la clase ya que es la clasificación de la actividad económica que definirá la empresa en el directorio.

2.9.- Una vez identificadas todas las unidades económicas que serán la matriz en cada empresa conformada por más de un establecimiento pasarán a formar el directorio de empresas. Las variables de directorio serán las de la unidad económica que funge como empresa matriz y las variables económicas serán conformadas por la sumatoria de las variables económicas de todos los establecimientos incluyendo las cantidades de las variables de la unidad económica matriz.

Actualmente, se identifica que el mayor gasto de recursos en la conformación de empresas, es la identificación de las agrupaciones de las empresas conformadas por más de un establecimiento bajo denominación de Razón Social y/o Nombre del Establecimiento, pues estamos jugando con ficheros de millones de registro.

Aunque se debe visualizar que si existieran ficheros externos de fuentes oficiales no solo es necesario conformar las empresas con la identificación de las unidades económicas bajo la denominación de Razón Social o Nombre del Establecimiento, sino que antes de tratar de incorporar los ficheros externos es necesario

identificar los registros de las unidades económicas que están en ambos ficheros con la finalidad de no duplicar información de una misma unidad económica en más de un fichero, creándonos conflicto al inflar la información acumulada en las variables económicas. Este empate permitirá no duplicar información e identificar aquellas unidades económicas que sí deban ser agregadas a los directorios de las empresas.

Una vez identificadas las unidades económicas que se duplican se elimina una de ellas y el resto será incorporado con previa identificación de la empresa a la que pertenezca.

3. PROPUESTA DE SOLUCIÓN

Este prototipo de sistema propone una solución de mejora en los siguientes aspectos:

La propuesta del prototipo es la de tomar una de las metodologías basadas en la Teoría Record Linkage (Empate de Registros) primeramente para la identificación de unidades económicas que se encuentren en dos o más ficheros eliminando los posibles duplicados. Posteriormente identificar las unidades con el uso de la misma teoría identificar las unidades económicas que se cobijen bajo la denominación de Razón Social o Nombre del Establecimiento para identificar las empresas conformadas por más de una unidad económica

La Teoría Record Linkage es una teoría basada en el concepto en empate de registros de uno o más ficheros cada uno de longitud igual o diferente en número de registros, a través de comparar variables similares o aproximadamente iguales que no cuenten con un identificador único que las relacione uno a uno.

Esta Teoría a avanzado mucho después de sus concepciones iniciales en donde se identifica que la teoría puede ser dividida en dos posibles etapas la primera que se basa en comparaciones determinísticas, es decir, tratar que la comparación sea uno a uno por el contenido idéntico de las variables que lo componen y la segunda en donde el uso de conceptos probabilísticos el que permite la comparación de cadenas en donde se pueden establecer reglas para hacer comparaciones aproximadas o similares de contenido. Fellegi and Sunter (1969) presentan su artículo llamado "A Theory for Record Linkage" en donde ya hacen una referencia de una metodología probabilística para el reconocimiento de personas en más de un fichero definiéndolo "said to be match" presentando en su investigación el panorama de que al realizar los empates se presentan tres casos posibles, loa empates, los no empates, y los que caen en un grupo de indecisos en donde puede ser que se consideraron match y no lo son o viceversa se consideran no match y si son match.

Por las definiciones y bases de esta teoría de empates, es posible que su utilización adaptándola para que en base a la Razón Social y/o Nombre del Establecimiento se realice la identificación o el reconocimiento de las unidades económicas que se encuentren bajo esas denominaciones. Además se propone la teoría para el empate de registros que permita identificar las unidades económicas que puedan estar duplicadas.

La propuesta de utilización de la teoría Record Linkage está relacionada con aspectos necesarios que optimicen la utilización de esta como son las teorías de estandarización y armonización de la información contenida en las variables contenidas en los ficheros que alimentaran el sistema.

3.1 Descripción del Prototipo del Sistema Propuesto

El prototipo del sistema que se propone esta compuesto de 3 módulos principales:

Formación de empresas.- Este módulo deberá cargar un directorio de unidades económicas a partir del cuales le definirán los parámetros sobre los cuales se conformará la empresa, este directorio deberá estar por lo menos estandarizadas las variables.

Actualización de empresas.- El módulo de Actualización de Empresas tendrá como requisito cargar primeramente un directorio de empresas conformado con un folio de empresa ya definido como único en cada una de las empresas y un directorio de unidades económicas el cual contenga una variable llamada MOV que defina o especifique altas o bajas de establecimientos que afectaran el directorio de empresas.

Herramientas

Empates de registros.- deberá cargar dos directorios en donde como requisito las variables que intervendrán en el empate deberán estar estandarizadas

(Estandarización de variables, combinación de campos y ordenamiento del contenido de variables). El módulo de herramientas permitirá estandarizar las variables, combinar variables si fuera necesario en un solo campo u ordenar alfabéticamente las palabras que compongan un campo o variable. Estas herramientas tiene la finalidad de crear un medio ambiente más adecuados para que el empate de registros y conformación de empresas tenga un mayor éxito mejorando las condiciones y disminuyendo algunos posibles errores o basura en el contenido de las cadenas que pueda dar parámetros menos exactos

Un cuarto módulo. Administración del Sistema. Módulo no analizado en este prototipo**.**

3.2 Rasgos Esenciales

Los rasgos esenciales de este prototipo de sistema son la creación de ensamblados construidos en Visual Basic uno que contenga el algoritmo q-gram para crear el filtro fino de comparación y un segundo ensamblado que contenga el algoritmo ASM.

Procedimientos almacenados para las validaciones.

Migración de las bases de datos a SQL Server 2005.

3.3 Rasgos de Alto Valor

Rasgos de alto valor están contenidos en los ensamblados que contienen los algoritmos q-gram y ASM así como los módulos de estandarización de variables, combinación de variables y eliminación de palabras que pueden crear conflictos en los empates. Las estandarizaciones, combinaciones y eliminación de palabras generan un medio ambiente más óptimo a los ensamblados para los empates.

3.4 Rasgos para Próximas Versiones

Los rasgos para próximas versiones deben enfocarse en calibrar de forma más eficiente los algoritmos creados en los ensamblados el q-gram y el ASM, así como la definición de estandarización de las variables y el análisis de palabras propuestas a eliminar por sector de unidad económica, mediante un análisis previo del comportamiento de las palabras, análisis más profundo sobre estandarización o corrección en cadenas de caracteres, armonización de variables entre directorios y análisis de eliminación de palabras en cadenas basado en estudio de comportamiento de cadenas por sector de actividad económica.

Mejorarse la parte donde se hace la penalización de pesos en letras diferentes en el algoritmo ASM.

Definir un metodología más had-oc y bajo normas, lineamientos o recomendaciones de organismos internacionales para la asignación de matriz y clasificación de actividades económicas que definan mejor a las empresas de más de una establecimiento tal vez con la utilización de redes neurales.

4. DESCRIPCIONES DEL USUARIO

Para proporcionar efectivamente los productos y servicios que resuelvan las necesidades del solicitante del sistema y de los usuarios, es necesario identificar a estos últimos, como parte de los requisitos que modelarán el proceso.

Para ello, a continuación se muestra una tabla que incluye un perfil de los solicitantes del sistema y de los usuarios implicados en el proyecto, así como los problemas dominantes que se perciben, para ser tratados por la solución propuesta.

CUADRO 1: Resumen del usuario

Nombre	Descripción	Responsabilidades
Responsable del Marco Nacional de Unidades Económicas	Este usuario determinará los parámetros sobre los cuales es sistema trabajará a nivel nacional y la participación de todos los sectores de actividades económicas a nivel nacional	Módulo de creación de empresas: Se encargará de cargar el directorio de unidades económicas de su respectivo sector el cual utilizará para la formación de empresas Determinará el marco de creación de empresas • Nacional • Por Entidad Federativa • Por Sector de Actividad Económica Define las prioridades a tomar cuando las empresas se encuentran conformadas por establecimientos de diferentes sectores de actividad económica, cuando exista un conflicto de más de una matriz en una misma empresa o bien en empresas con ausencia de unidad económica definida como matriz
Responsable del Marco de unidades económicas de alguno de los sectores económicos	Este usuario determinará los parámetros sobre los cuales el sistema trabajará	Módulo de creación de empresas: Se encargará de cargar el directorio de unidades económicas de su respectivo sector el cual utilizará para la formación de empresas Determinará el marco de creación de empresas • Nacional • Por Entidad Federativa Definir los parámetros de asignación de unidad económica matriz cuando exista más de una matriz en una empresa o bien cuando exista ausencia de matriz en una empresa.

5. RESTRICCIONES

Una de las principales restricciones del sistema es la carencia o falta de información en las variables que forman el directorio de unidades económicas, como puede ser falta de Nombre de Establecimiento o Razón Social, falta de la clasificación de la actividad económica, falta de información de variables cuantitativas como son Total de Personal Ocupado, Total de Ingresos y Totales.

Una restricción más es la estandarización u homologación de contenido o forma de las variables que contiene el directorio. Basura contenida en él, sobre todo en el caso de directorios de unidades económicas recibido de instituciones externas.

5.1 Del Proceso de Desarrollo

Una restricción del proceso es que se desconoce el almacenamiento y plataforma de almacenamiento de los directorios externos que podrían alimentar el sistema.

El desarrollo de este prototipo esta diseñado en base a una investigación y percepción de los conceptos hechos en instancia de ella, y es una propuesta de mejora.

5.2 De Ambiente y Tecnología

Una de las restricciones importantes del desarrollo sería la necesidad de un servidor SQL server 2005 y un servidor de aplicación para pruebas del sistema y estrés de este. Posteriormente de las pruebas y aceptación al pasar de un prototipo a un sistema liberado los servidores de producción y los clientes para los usuarios.

La propuesta de este prototipo tiene como principal restricción que es desarrollada en un equipo que funge como servidor lo recomendable es almacenar la aplicación a un servidor en donde el espacio en disco y la capacidad de la RAM permita que los componentes trabajen en condiciones más óptimas, el equipo sobre el que trabaja solo tiene un GB en RAM, hasta este momento el prototipo no se ha probado en ningún otro ambiente ni servidor de prueba, el prototipo esta diseñado pensando en un servidor SQL Server 2005 y ambiente .NET específicamente Visual Basic, esta aplicación esta pensada modularmente para que en un futuro pueda ser alimentada o sea utilizada en algún otro sistema. La licencia utilizada para su desarrollo es precisamente de desarrollo no de explotación o producción.

5.3 De Entrega e Instalación

La entrega e instalación de este sistema por el momento no esta contemplada ya que dependerá de la asignación de recursos técnicos y la aprobación de este prototipo por las áreas y usuarios involucrados.

PLANTILLA 2.-

2. REQUERIMIENTOS
Versión <1.0 Beta>

Área:	Prototipo de Sistema dirigido la Dirección de Marcos y Muestreo
Fuente:	Protocolo de Tesis para la Maestría de Informática y Tecnologías Computacionales
Analista(s):	Sara Josefina Palacio Gámez
No. Solicitud:	
Fecha de Captura del Requerimiento:	Enero 2009

Número	Descripción	Material de Apoyo	Caso de Uso	Clasificación (Subclasificación)
1	Módulo que permitirá conformar empresas que se encuentren bajo la misma denominación de Razón Social y/o Nombre del Establecimiento basado en un directorio de unidades económicas en donde como primer agrupación o reconocimiento de empresa sea en base al folio censal.	Proceso de conformación de empresas Documento de descripción de homogenización de variables propuesta por censos económicos Proceso de Asignación de Matriz	Formación de Empresas	Determinará características de conformación de la empresa. (Requerimiento no funcional) Determinar características de asignación de matriz. (Requerimiento no funcional) Determinar características de

		Teoría Record Linkage, Estandarización de variables		asignación de sector. (Requerimiento no funcional) Realizar formación de empresas (Requerimiento funcional)
2	Actualización de un directorio de empresas con información validas provenientes de fuentes internas o externas	Documento de descripción de homogenización de variables propuesta por censos económicos Ficheros o archivos de actualización. Documento de actualización del directorio de empresas	Actualización de empresas	Cargar tablas de actualización, (Requerimiento no funcional) Empate entre ficheros. (Requerimiento funcional) Procesos de validación de actualización de directorios. (Requerimiento funcional)
3	Herramienta basada en la teoría Record Linkage que sirve de base para todo el sistema, para tratar de evitar duplicado, empatar archivos e intentar identificar los registros de un fichero que se encuentren en uno de actualización y que no cuenten con un identificador único que lo relacione, tratar de evitar duplicar información de unidades económicas e identificar la mayoría de las unidades económicas que se encuentren bajo la misma denominación de Razón Social y/o Nombre del Establecimiento.	Teoría Record Linkage, Estandarización de variables	Herramientas Empate de registros Estandarización de variables Ordenamiento de palabras y combinación de campos	Cargar primer archivo. (Requerimiento no funcional) Cargar segundo archivo Definir blocking. (Requerimiento funcional y no funcional) Definir Filtering, parámetros de q-gram. (Requerimiento funcional y no funcional)

				Definir parámetros de ASM Elimina caracteres erróneos, elimina palabras. (Requerimiento funcional y no funcional) Ordena palabras dentro de una cadena. (Requerimiento no funcional) Combinación de campos. (Requerimiento no funcional)
4	Administración del Sistema	Módulo no revisado en este prototipo		

Los requerimientos se clasifican en Funcionales y No Funcionales. Los requerimientos funcionales son aquellos que representan una serie de acciones en el sistema, mientras que los no funcionales determinan las cualidades del sistema, es decir, como responde el sistema en la interacción con el usuario, por ejemplo requerimientos de seguridad, desempeño, disponibilidad, etc. De seguridad (respaldos, velocidad de respuesta, base de datos, etc.), de capacidad (usuarios simultáneos, terminales, cantidad de información), de disponibilidad, de mantenimiento (modularidad, POO, etc.), de portabilidad (PC o Macintosh, Windows, Linux, etc.).]

Firma Titular	Firma Titular

Área:	Prototipo dirigido la Dirección de Marcos y Muestreo
Fuente:	Protocolo de Tesis para la Maestría de Informática y Tecnologías Computacionales
Analista(s):	Sara J. Palacio Gámez
No. Solicitud:	
Fecha de Captura del Requerimiento:	Enero 2009

1. Factibilidad Técnica

1.1. Hardware y Software Disponibles y utilizados en este momento

Software.

En estos momentos el trabajo de conformación de empresas se realiza mediante Visual Fox 9.0, apoyado en aplicaciones desarrolladas en base al conocimiento empírico del personal.

Hardware

- Equipo
- 1GB en RAM
- Disco Duro de 250 GB
- Tarjeta de Red
- Teclado
- Mouse

1.2. Propuesta de Requerimientos

Software:

Para este proyecto se propone el desarrollo del prototipo, para la creación de la interfaz en Visual Basic y una administrador de bases de datos en SQL Server 2005.

Hardware

Se propone el almacenamiento del sistema en un servidor de producción del instituto para aplicaciones bajo plataforma. NET y un servidor de producción de Administrador de Bases de datos de SQL Server 2005. Con capacidad de almacenamiento de 100 GB

2. Factibilidad Operativa.

Este prototipo propone una sistema que intente disminuir las revisiones manuales o semiautomáticas que en el momento se realizan para la conformación y actualización de empresas, intentando disminuir le tiempo de conformación de ellas, proponiendo la utilización de conceptos bajo la teoría Record Linkage (Empates de registros)

2.1. Beneficios Intangibles:

Intentar disminuir el tiempo de revisión manual de unidades económicas bajo la ,misma denominación de Razón Social y/o Nombre del Establecimiento.

2.2. Beneficios Tangibles

Una mejor administración de los directorios de unidades económicas que conformen directorios de empresas mediante el uso de un administrador de bases de datos, Estandarización automática del proceso de conformación de empresas de más de un establecimiento que se identifique ya sea por Folio Censal y/o por Razón Social y/o Nombre del Establecimiento.

Firma Titular Área de Desarrollo	Firma Titular Área Usuaria

Plantilla 3.-

Versión <1.0 Beta>

3. LISTA DE RIESGOS
Versión <1.0>

Fecha de creación del documento <16/03/10>

CUADRO 2: Lista de Riesgos.

1. 1 Riesgo
01
1.1.1 Descripción
Identificación de las variables candidatas similares entre ficheros para el proceso de empates
1.1.2 Tipo
Directo
1.1.3 Magnitud
Alto
1.1.4 Impacto
De no ser identificadas correctamente las variables similares entre ficheros la incorporación de unidades económicas bajo la denominación de Razón Social o Nombre del Establecimiento no se realizar á por lo que será un fracaso la incorporación de unidades económicas bajo estas denominaciones. No se identificará correctamente las unidades duplicadas y será imposible que se incorporen al directorio.
1.1.5 Estrategia de Mitigación
Deberá recurrir a las aplicaciones existentes en este momento
1.1.6 Plan de Contingencia
La incorporación de unidades económicas deberá hacerse de forma manual o semiautomática. La identificación de unidades económicas entre ficheros deberá hacerse de forma manual o semiautomática dependiendo de las aplicaciones y la plataforma con la que se cuente en el momento.
1. 2 Riesgo
2.-
1.2.1 Descripción

Contenido de caracteres basura dentro de las cadenas que forman las variables similares entre ficheros,
1.2.2 Tipo
Indirecto
1.2.3 Magnitud
Significativo
1.2.4 Impacto
Los errores contenidos en las cadenas de los ficheros como son Razón Social y/o Nombre del Establecimiento pueden provocar que no se identifique correctamente las unidades económicas ya sea para la incorporación de la unidad a la empresa o para detectar si ya esta dentro de alguna provocando duplicidad de información. Identificación de dos empresas en vez de una por considerarla un Nombre o Razón Social diferente
1.2.5 Estrategia de Mitigación
Utilización de aplicaciones existentes hasta este momento,
1.2.6 Plan de Contingencia
La incorporación de unidades económicas deberá hacerse de forma manual o semiautomática. La identificación de unidades económicas entre ficheros deberá haces de forma manual o semiautomática dependiendo de las aplicaciones y la plataforma con la que se cuente en el momento.
1.3. Riesgo
3
1.3.1 Descripción
Identificación correcta de los parámetros de Blocking y Filtering óptimos para el correcto funcionamiento de los algoritmos q-gram y ASM.
1.3.2 Tipo
Directo
1.3.3 Magnitud
Significativo
1.3.4 Impacto
Al no calibrar correctamente los parámetros más óptimos para los algoritmos de q-gram y ASM puede provocar un bajo reconocimiento de unidades económicas o no identificación al 100% de las unidades económicas existentes entre dos ficheros o bien traer mucha basura en las identificaciones trayendo como consecuencia invertir mucho tiempo en revisiones manuales u omitiendo unidades económicas en los procesos que se estén corriendo.

1.3.5 Estrategia de Mitigación
Se realizara el proceso como se hace actualmente
1.3.6 Plan de Contingencia
Realizar el empate de registros y/o la incorporación de unidades económicas de forma semiautomática o manual
1. 4 Riesgo
Carencia u omisión de información considerada básica en los registros contenidos en los ficheros o fichero de directorios de unidades económicas
1.1.1 Descripción
La información básica y necesaria de registros de ficheros para incorporarse dentro del directorio de empresas deberá contener como mínimo, La entidad federativa, municipio, localidad, Razón Social y Nombre del Establecimiento (opcional), ubicación física de la unidad económica (calle, número exterior, número interior, letra, colonia) total de personal ocupado, clase de actividad económica y tipo de establecimiento.
1.1.2 Tipo
Directo
1.1.3 Magnitud
Alto
1.1.4 Impacto
No se incorporara la información de esta unidad económica
1.1.5 Estrategia de Mitigación
Solicitar nuevamente al información de la unidad económica si es posible
1.1.6 Plan de Contingencia
No existe

Firma Titular Área de Desarrollo

Firma Titular Área Usuaria

Plantilla 4.-

4. CASOS DE USO
Versión <1.0 Beta>

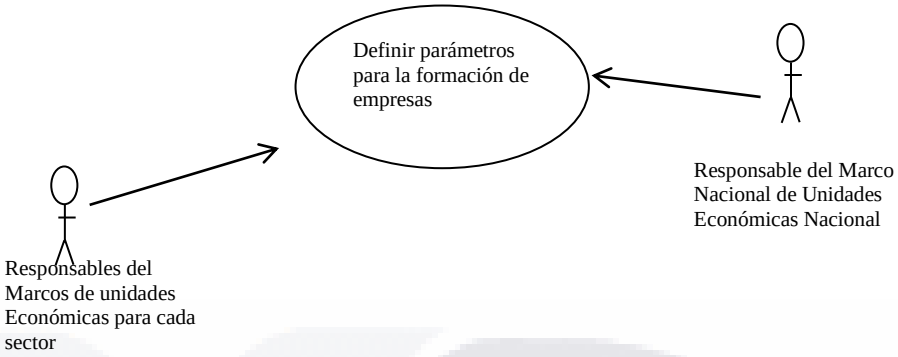
Fecha de última actualización <16/03/10>

CUADRO 3: Casos de uso.

Consecutivo o Identificador	Caso de Uso	Actor Principal
1	Formación de Empresas	Responsable del Marco del Sector Primario de Unidades económicas, Responsable del Sector Secundario de Unidades Económicas, Responsable del Sector Terciario de Unidades económicas o Responsable del Marco Nacional de Unidades Económicas
2	Actualización de empresas	Responsable del Marco del Sector Primario de Unidades económicas, Responsable del Sector Secundario de Unidades Económicas, Responsable del Sector Terciario de Unidades económicas o Responsable del Marco Nacional de Unidades Económicas
3	Herramientas Empates de registros (Q-gram, ASM Ordenamiento de palabras dentro de una cadena Combinación de campos	Responsable del Marco del Sector Primario de Unidades económicas, Responsable del Sector Secundario de Unidades Económicas, Responsable del Sector Terciario de Unidades económicas o Responsable del Marco Nacional de Unidades Económicas
4	Administración del sistema	Administrador del sistema

Firma Titular Área de Desarrollo

Firma Titular Área Usuaria



CUADRO 4: Plantilla de Casos de Uso

ID:	001
Nombre:	Formación de Empresas
Fecha:	Diciembre 2009

Descripción:

Permite definir los parámetros para la conformación del directorio de Empresas

Actor:

Responsable del Marco del Sector Primario de unidades económicas, Responsable del Sector Secundario de unidades Económicas, Responsable del Sector Terciario de unidades económicas o Responsable del Marco Nacional de Unidades Económicas

Precondiciones:

Haber accedido y autentificado dentro del sistema

Flujo Normal:

1. Cargar tabla de Directorio de Unidades Económicas
2. Definir características de la formación de la empresa Nacional

   Entidad Federativa

   Sector de Actividad Económica (Solo para el caso del Responsable del Marco Nacional de Unidades Económicas

3. Definir características para la inicialización de la formación de la empresa Folio Censal

   Razón Social y/ o Nombre del Establecimiento

   Razón Social

4. Definir parámetros para asignar unidad económica matriz para los casos en que la empresa se encuentre en más de un sector de actividad económica, empresa(s) con más de una matriz definido y

empresa(s) sin matriz identificada Sector Manufacturero

Proporción de Sector de Actividad Económica con mayor acumulación de ingresos totales, Unidad Económica con mayor acumulación de ingresos totales

Flujo Alternativo:

De no establecer los parámetros correctos la formación de la empresa deberá realizarse de forma manual y semiautomática como hasta al momento se viene realizando.

Postcondiciones:

En el caso de que los parámetros sean los correctos el sistema deberá arrojar un directorio de empresas conformadas por más de una unidad económica y un directorio de empresa de únicos, además en el directorio de unidades económicas diferenciara los establecimientos que pertenecen a una empresa conformada por más de una unidad económica en donde identifique a la matriz y a sus sucursales, por ultimo identifica los únicos

Casos de Uso relacionados:

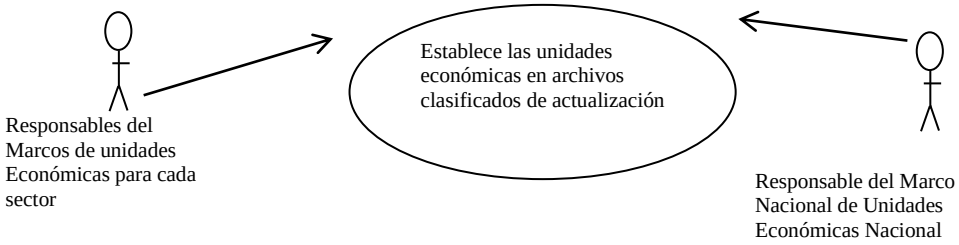
Empate de Registros y Herramientas

Comentarios u observaciones:

La conformación de empresas deberá tener como pre-requisito el de estandarizar previamente los directorios con la finalidad de disminuir el margen de error en el reconocimiento de unidades económicas que e encentren baja la misma denominación de Nombre del Establecimiento y/o Razón Social.

Firma Titular Área de Desarrollo

Firma Titular Área Usuaria



Identificador:	002
Nombre:	Actualización de Empresas
Fecha:	Diciembre 2009

Descripción:

Permite definir los parámetros y tablas de actualización de las empresas cada una de las tablas debe estar clasificada de acuerdo a la problemáticas o actualización que vaya a realizar, y deberá previamente ya haber identificado la empresa y al directorio que pertenece.

Actor:

Responsable del Marco del Sector Primario de unidades económicas, Responsable del Sector Secundario de unidades Económicas, Responsable del Sector Terciario de unidades económicas o Responsable del Marco Nacional de Unidades Económicas

Precondiciones:

Las tablas e actualización y mantenimiento de las empresas deben estar debidamente clasificados y los registros de las unidades económicas que van a actualizar los directorios de empresas deberán estar identificados en relación con la empresa que actualizará

Flujo Normal:

1. El sistema cargara tablas de acuerdo a su clasificación de actualización
   Bajas definitivas

   Altas de unidades económicas

   Cambios de sector, reclasificaciones o cambios de giro

   Cambios de domicilio

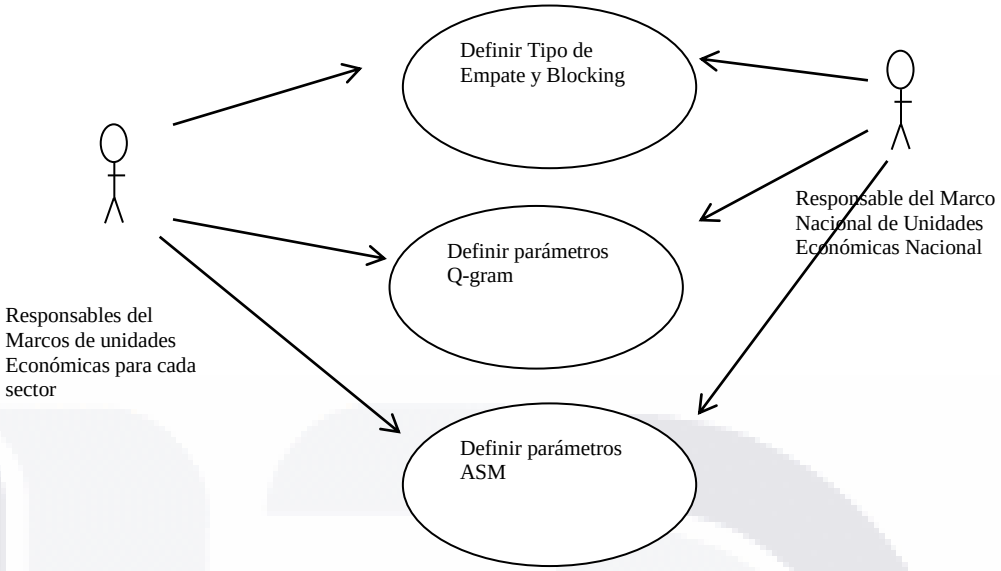
   Fusiones

2. Los registros de las unidades económicas deben haber sido identificados anteriormente a la empresa que pertenecen.
3. Actualizar las empresas con las tablas de alimentación enviadas con el contenido y especificación de la actualización

Flujo Alternativo: De no realizar correctamente las actualizaciones a los directorios de empresas el usuario deberá actualizarlos de forma semiautomática o manual conforme se ha realizado anteriormente este procedimiento
Poscondiciones: El o los directorios de empresas deberán haber sido actualizados.
Casos de Uso Relacionados: Empate de Registros y Herramientas
Comentarios u observaciones: Si los registros de las tablas no están debidamente clasificados y con la información suficiente que especifica el tipo de actualización que se realizará, el sistema incurrirá en una mala o deficiente actualización de los directorios de empresas, de la misma forma deberán previamente identificar cada uno de estos registros a que empresa afectar ya sea que pertenezca a el directorio de las empresas conformadas por más de un establecimiento o bien a el directorio de empresas de únicos, cabe mencionar que derivado de esta actualización es posible que las altas unidas a algún establecimiento del directorio de empresas denominados únicos, puede haber la posibilidad de conformar nuevas empresas de más de un establecimiento, caso contrario el de las bajas definitivas puede traer como consecuencia el que una empresa conformada por más de un establecimiento pase a ser parte del directorio de empresas únicos, debe tenerse cuidado con los cambios de sector, giro o reclasificaciones de tal forma de verificar si la empresa que afecta no cambia de sector.

Firma Titular Área de Desarrollo

Firma Titular Área Usuaria



ID:	003
Nombre:	Empates
Fecha:	Diciembre 2009

Descripción:

Permite la identificación de los registros de un directorio con información de unidades económicas contra otro directorio ya sea de unidades económicas o de empresas

Actor:

Responsable del Marco del Sector Primario de unidades económicas, Responsable del Sector Secundario de unidades Económicas, Responsable del Sector Terciario de unidades económicas o Responsable del Marco Nacional de Unidades Económicas

Precondiciones:

Haber accedido y autentificado dentro del sistema

Flujo Normal:

1. Cargar Primer Directorio de registros de unidades Económicas o Empresas
2. Cargar Segundo Directorio de registros de unidades Económicas o Empresas
3. Definir tipo de empata
   De Empresas

   De Unidades Económicas

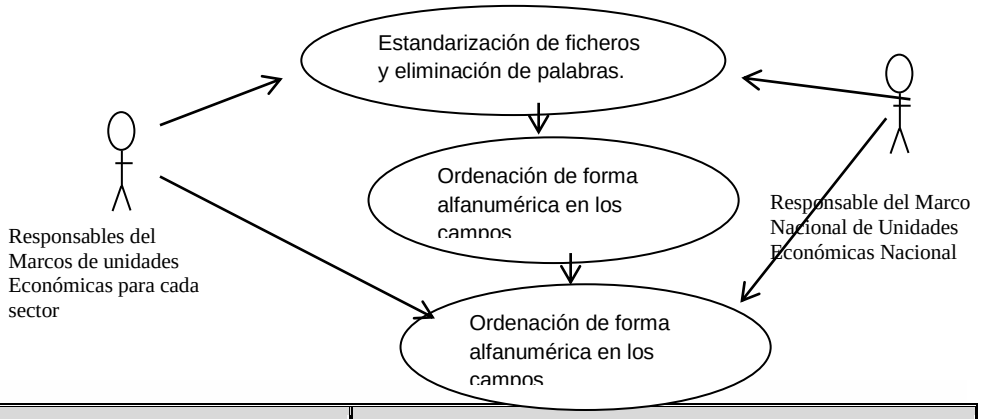
   Empate Exacto sobre cuales campos

   Definir Blocking, por un campo completo, por una o dos palabras de un campo

4. Definir parámetros del algoritmo q-gram
5. Definir parámetros del algoritmo ASM

Flujo Alternativo: Para el proceso de empates de registros entre dos o más tablas de registros de unidades económicas se cuenta con una aplicación desarrollada en el área, que aunque nunca se ha generado sin ningún algoritmo determinístico o probabilístico para el empate de registros si ha demostrado una eficiencia considerable que permite hacer empates de registros y la revisión es manual y tediosa
Postcondiciones: El algoritmo de empate de registros permitirá identificar si ya existe un registro de alguna unidad económica ya existente dentro de una empresa conformada por más de una unidad económica en donde identifique a la matriz y a sus sucursales, por ultimo identifica los únicos
Casos de Uso relacionados: Empate de Registros y Herramientas
Comentarios u observaciones: La conformación de empresas deberá tener como prerrequisito el de estandarizar previamente los directorios con la finalidad de disminuir el margen de error en el reconocimiento de unidades económicas que e encentren baja la misma denominación de Nombre del Establecimiento y/o Razón Social.

Firma Titular Área de Desarrollo	Firma Titular Área Usuaria



Identificador:	004
Nombre:	Herramientas
Fecha:	Diciembre 2009

Descripción:

Lanza un proceso en donde estandariza ciertos caracteres considerados como erróneos o basura como pueden ser caracteres extraños no reconocidos en vez de una ñ, Ñ p caracteres extraños en vez de vocales acentuadas, etc., Lanzar un procedimiento de quitar palabras más convencionales, lanzar un procedimiento que ordene de forma alfabética las cadenas contenidas en ciertas variables y combinar variables si es necesario

Actor:

Módulo de estandarización

Precondiciones:

Cargar archivos y automáticamente al realizar el empate de registros se correrá este procedimiento

Flujo Normal:

4. El sistema cargara tablas para realizar el empate
   Se lanzará el procedimiento de estandarización

   Lanzará el procedimiento de eliminación de palabras

   Lanzará el procedimiento de ordenación de forma alfabética un campo

   Lanzará procedimiento de combinación de variables o campos en caso necesario

Flujo Alternativo:

Correr los procedimientos anteriores independientemente del sistema

Poscondiciones: Al lanzar correctamente estos procedimientos existen mejores condiciones para la optima corrida de los algoritmos q-gram y ASM
Casos de Uso Relacionados: Conformación de Empresas y Actualización de Empresas
Comentarios u observaciones: De no existir estas condiciones cabe la posibilidad de afectar los resultados que arrojen los algoritmos-q-gram y ASM

Firma Titular Área de Desarrollo	Firma Titular Área Usuaria

Plantilla 5.-

5. CONFORMACIÓN DEL EQUIPO DE TRABAJO
Versión <1.0 Beta>

Debido a que este trabajo es un tesina no se consideró un equipo de trabajo para el análisis, diseño e implementación de este prototipo de sistema, sin embargo se deja ejemplificado el formato y la especificación de por que no es requisitado.

CUADRO 5: Conformación del equipo de trabajo.

Integrante	Rol
Al ser un prototipo propuesto en una tesis no existen integrantes de equipo para el análisis y diseño de este prototipo, debido a que una sola persona realizó todos los roles	Se considera que este formato para este caso en especifico no es aplicable
Sara Josefina Palacio Gámez	Analista del Sistema Especificador de Casos de Uso Diseñador de interface Arquitecto Ingeniero de Casos de Uso Ingeniero de Componentes Ingeniero de Pruebas Líder de Proyecto Integrador del Sistema

Tipo de Roles y características que deben tomarse en consideración para el llenado de este formato:

- **Analista de Sistemas**

  Es el encargado de la planeación del nuevo sistema, realizando procesos que sirven para recopilar e interpretar los hechos, diagnosticar problemas y utilizar estos hechos a fin de mejorar el sistema. Es el responsable del conjunto de requerimientos y de delimitar el sistema.

- **Especificador de Caso s de Uso**

  Como parte del análisis, identifica y describe qué es lo que el sistema debe hacer desde el punto de vista del usuario. Es decir, describe un uso del sistema y cómo éste interactúa con el usuario, para poner en marcha el diseño.

- **Diseñador de Interfaces**

  De acuerdo a las necesidades y requerimientos del usuario, se encarga del diseño de las interfaces gráficas que serán la conexión entre el usuario y el sistema.

- **Arquitecto**

  Participa en el flujo de trabajo de los requerimientos para describir la vista de la arquitectura del modelo de casos de uso. Es responsable de la integridad del modelo de análisis, garantizando que éste sea correcto, consistente y legible como un todo.

- **Ingeniero de Casos de Uso**

  Es responsable de garantizar la integridad en la realización de los casos de uso, para que estos cumplan con los requerimientos planteados.

- **Ingeniero de Componentes**

  Encargado de definir y mantener el código fuente de uno o varios componentes, garantizando que cada componente implemente la funcionalidad correcta, asegurándose de cumplir con los requerimientos, de acuerdo a los casos de uso especificados.

- **Ingeniero de Pruebas**

  Encargado de diseñar y crear las pruebas necesarias para la posterior verificación de la funcionalidad de cada módulo del sistema, verificando que se cubran las necesidades para las cuales fue creado, siendo también responsable de las pruebas de integración y del sistema cuando éstas se ejecutan.

- **Líder de Proyecto**

  Encargado de coordinar las distintas actividades de los integrantes del equipo de trabajo, para lograr una buena colaboración entre las diferentes áreas del proyecto, evitando la duplicidad de funciones o la falta de alguna de ellas.

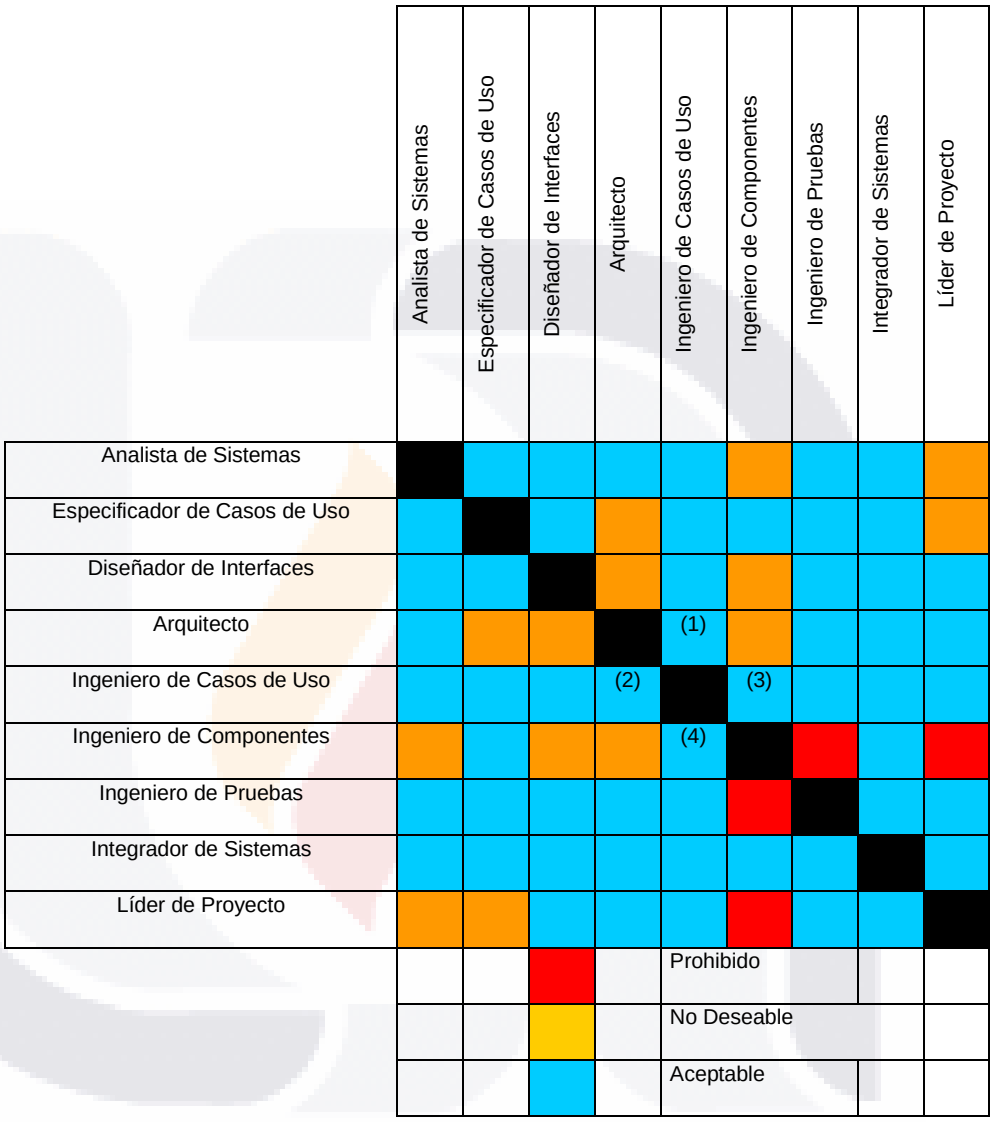
- **Integrador de Sistemas**

  Encargado de llevar a cabo la secuencia de construcciones necesarias en cada iteración y la integración de cada construcción, cuando sus partes han sido implementadas. Describe la funcionalidad que deberá ser implementada y qué partes del modelo de implementación se verán afectadas.

Notas Importantes:

- Un integrante del equipo de trabajo, puede tener más de un rol.
- Para el llenado de este formato, se deberá sustituir el texto que se encuentra entre corchetes con la información que en cada sección se especifica.
- Es recomendable, para efectos de un mejor desempeño de las actividades de cada integrante, tomar en cuenta el siguiente cuadro, en el que se especifican las combinaciones de los roles:

Sugerencia de Combinación de Roles:



(1) Siempre y cuando el ingeniero de casos de uso, no sea el ingeniero de componentes.
(2) Siempre y cuando el ingeniero de casos de uso, no sea el ingeniero de componentes.
(3) Siempre y cuando el ingeniero de casos de uso, no sea el arquitecto.

Siempre y cuando el ingeniero de casos de uso, no sea el arquitecto.

Plantilla 6.-

6. PLAN DE PROYECTO
Versión <1.0 Beta>

Fecha de última actualización <16/03/10>

PLAN DE PROYECTO

Calendarización del Proyecto

CUADRO 6: Plan de proyecto

	Fecha Inicial	Fecha Final
Fase de Gestación [Esta fase desarrolla los requerimientos del producto y establece los casos de uso para el sistema.]		
Iteración Preliminar		
Recopilación de requerimientos generales	01/03/2009	30/05/2009
Recopilación de requerimientos detallados	01/06/2009	30/11/2009
Identificación de Casos de Uso	01/04/2009	30/05/2009
Fase de Elaboración [Al término de esta fase deben estar probados la mayoría de los componentes de la arquitectura]		
Iteración – Desarrollo del Prototipo de Arquitectura		
Detallar casos de uso más importantes.	01/08/2009	30/08/2009
Prueba de las tecnologías a utilizar	01/09/2009	30/11/2009
Desarrollo de los componentes de la arquitectura	010/12/2009	A la fecha
Fase de Construcción [En esta fase debe estar implementada, probada y liberada una versión beta]		
Iteración C1 – Desarrollo Liberación Versión Beta		
Implementación del sistema	01/05/2010	30/05/2010
Pruebas del sistema	01/06/2010	30/06/2010
Liberación de la versión beta	01/07/2010	30/07/2010

	Fecha Inicial	Fecha Final
Fase de Transición [En esta etapa el sistema deberá liberarse para su instalación, proveyendo los requerimientos de soporte, instalación y capacitación del usuario.]		
Iteración – Ajustes y modificaciones de la versión beta		
Ajustes y modificaciones		
Prueba		
Código Completo		
Liberación preliminar		
Iteración – Desarrollo liberación 2 Beta 1		
Ajustes y modificaciones		
Prueba		
Código Completo		
Liberación preliminar		
Iteración – Aceptación y Liberación		
Código Completo		
Liberación del Producto (proyecto liberado)		

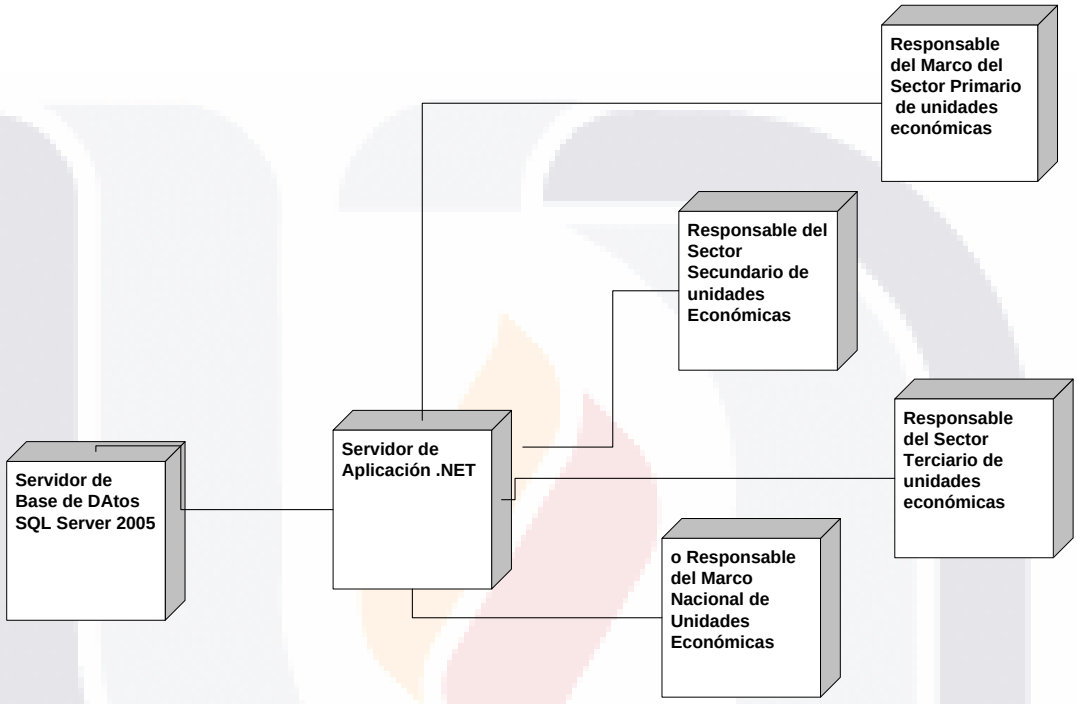
Plan de Fases. En esta sección, se elabora un resumen de las iteraciones que se realizarán en cada fase, así como el tiempo que se llevará trabajar en ellas.

CUADRO 7: Plan de fases

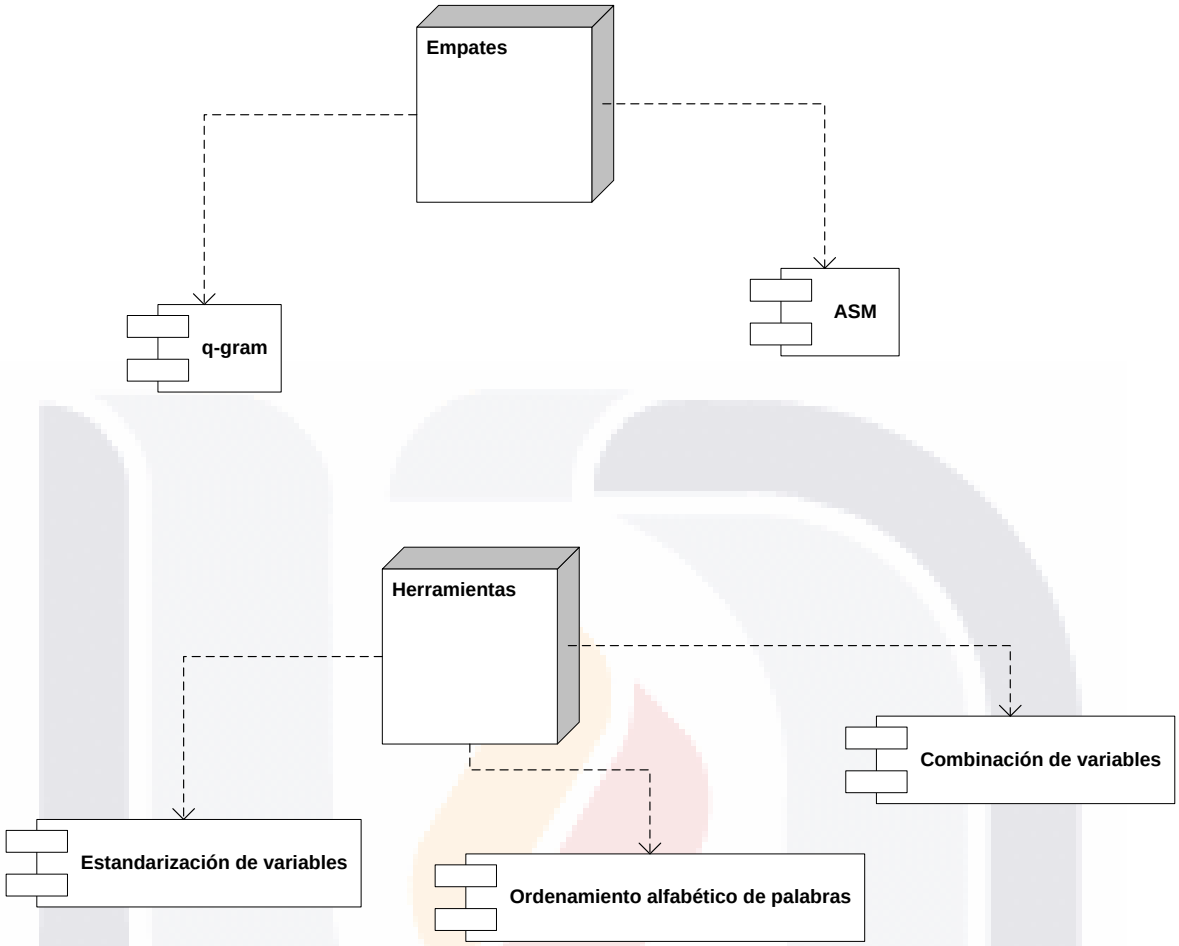
Fase	No. de Iteraciones	Conclusión
Fase de Gestación	1	16 semanas
Fase de Elaboración	1	50 semanas
Fase de Construcción	1	12 semanas
Fase de Transición		

Plantilla 7.-

MODELO DE DESPLIEGUE



La interacción principal del sistema se realiza mediante una serie de pantallas principal de aplicación en NET hacia el servidor de Base de datos en SQL Server 2005, mediante ellos se envían los parámetros hacia los ensamblados y componentes, con cada uno de los clientes en los nodos terminales de los responsables



Plantilla 8.-

8. PLAN DE ITERACIÓN

Versión <1.0>

No utilizado por el momento en esta investigación el prototipo esta en la primera interacción.

Plantilla 9.-

9. IDENTIFICACIÓN DE CLASES
Versión <1.0 Beta>

Fecha de última actualización <16/03/10>

CUADRO 9: Identificación de clases.

Clase: <Formar Empresas>	
Características	**Responsabilidades**
Folcen RazonSocial NombEstab ClaAct TipoEmpre EntFed	FormarF01(). Formar la empresa, por el ámbito geográfico definido AnalizaMatriz(). Identifica la unidad económica que se denominara matriz SectorAct(). Identifica la Clasificación del Sector de Actividad Económica que definirá a la matriz

Clase: <Actualiza Empresa>	
Características	**Responsabilidades**
TipoMov FolEmpre RazonSocial Domicilio TotPO ClaAct	Alta(). Da altas unidades económicas a las empresas Baja(). Da bajas las unidades económicas a un directorio de empresas Actualiza() Realiza actualizaciones en campos de directorio a un directorio de empresas Problemática() Identifica problemáticas en unidades económicas de las empresas

Clase: <Empate>	
Características	**Responsabilidades**
Pa_qgram Pa_ASM estandariza	Filtering(). Genera Filtros gruesos (Entidad, una palabras más larga de una cadena o dos palabras más largas de una cadena ASM() Identifica % de similitud entre dos cadenas comparadas Estandariza(). Homogeniza cadenas, ordena

	alfabéticamente cadenas, concatena campos

Clase: <Filtering>	
Características	**Responsabilidades**
Q_campo Q_op Filt_fino	Blocking() Define parámetros de filtro grueso Q-gram() Define parámetros de filtro fino

Clase: <ASM>	
Características	**Responsabilidades**
Asm_simil Umbral_asm Peso_asm	ASM() Detecta el grado de similitud entre dos cadenas comparadas

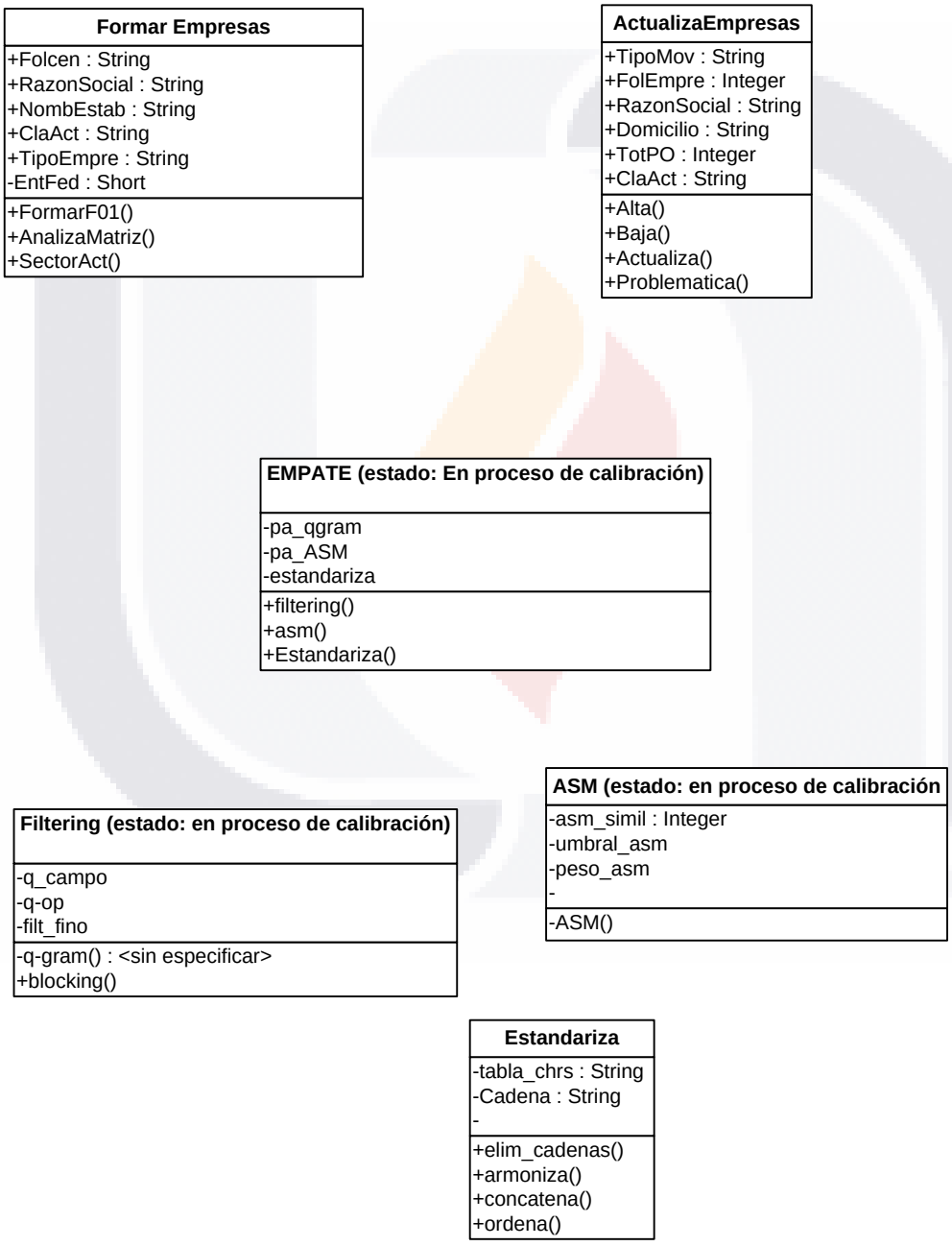
Clase: <Estandariza>	
Características	**Responsabilidades**
Tabla_chrs cadena	Elim_cadenas() Elimina cadenas identificadas como posibles a causar algún problema en la comparación e identificar el grado de similitud Armoniza(). Elimina caracteres basura de las cadenas Concatena() concatena campos Ordena() Ordena alfabéticamente el contenido de un campo

Plantilla 10.-

10. DIAGRAMA DE CLASES
Versión <1.0 Beta>

Fecha de última actualización <16/03/10>

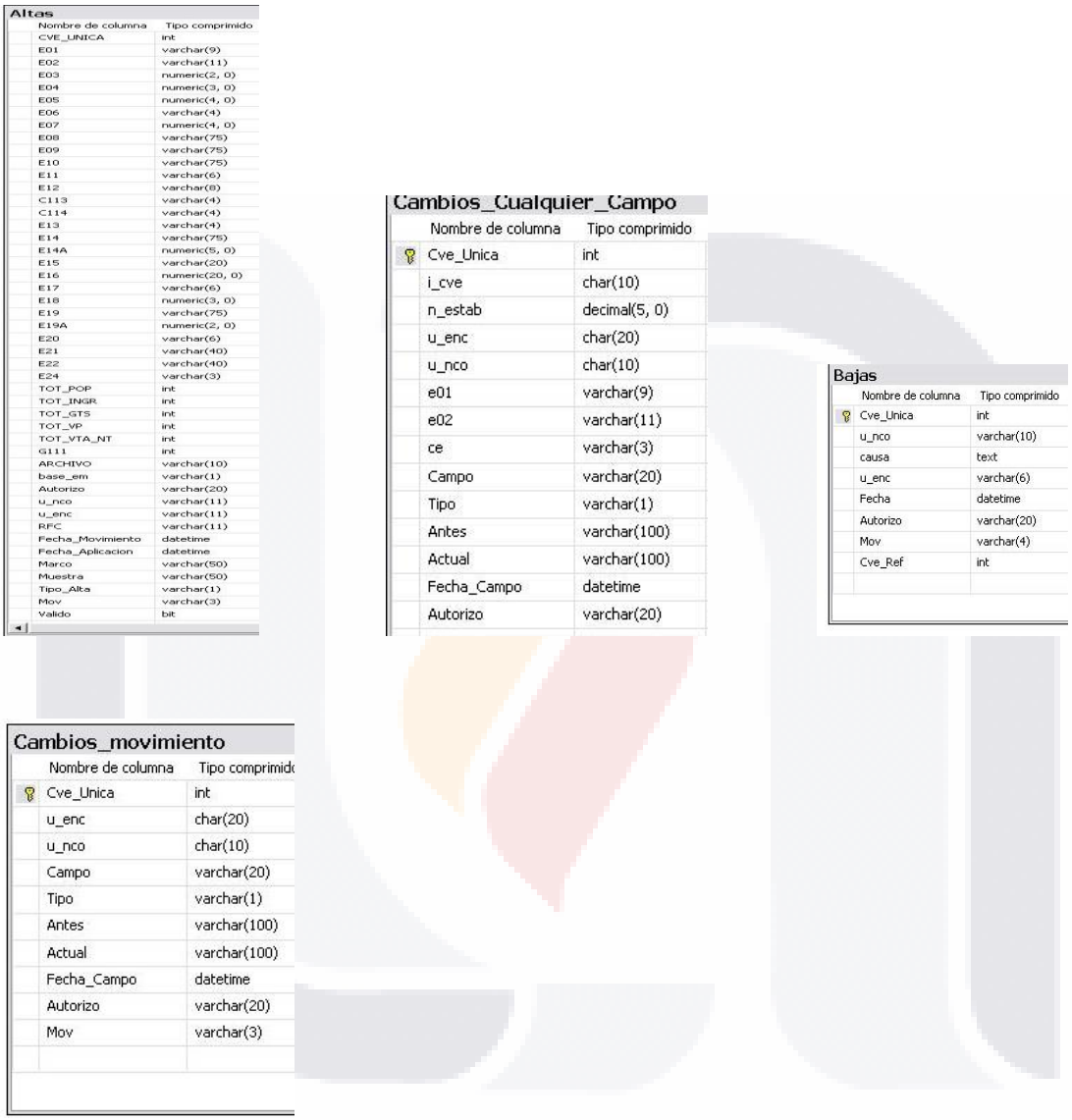
DIAGRAMAS DE CLASES



Plantilla 11.-

11. MODELO DE DATOS

Versión <1.0 Beta>



Altas

Nombre de columna	Tipo comprimido
CVE_UNICA	int
E01	varchar(9)
E02	varchar(11)
E03	numeric(2, 0)
E04	numeric(3, 0)
E05	numeric(4, 0)
E06	varchar(4)
E07	numeric(4, 0)
E08	varchar(75)
E09	varchar(75)
E10	varchar(75)
E11	varchar(6)
E12	varchar(8)
C113	varchar(4)
C114	varchar(4)
E13	varchar(4)
E14	varchar(75)
E14A	numeric(5, 0)
E15	varchar(20)
E16	numeric(20, 0)
E17	varchar(6)
E18	numeric(3, 0)
E19	varchar(75)
E19A	numeric(2, 0)
E20	varchar(6)
E21	varchar(40)
E22	varchar(40)
E24	varchar(3)
TOT_POP	int
TOT_INGR	int
TOT_GTS	int
TOT_VP	int
TOT_VTA_NT	int
G111	int
ARCHIVO	varchar(10)
base_em	varchar(1)
Autorizo	varchar(20)
u_nco	varchar(11)
u_enc	varchar(11)
RFC	varchar(11)
Fecha_Movimiento	datetime
Fecha_Aplicacion	datetime
Marco	varchar(50)
Muestra	varchar(50)
Tipo_Alta	varchar(1)
Mov	varchar(3)
Valido	bit

Cambios_Cualquier_Campo

Nombre de columna	Tipo comprimido
Cve_Unica	int
i_cve	char(10)
n_estab	decimal(5, 0)
u_enc	char(20)
u_nco	char(10)
e01	varchar(9)
e02	varchar(11)
ce	varchar(3)
Campo	varchar(20)
Tipo	varchar(1)
Antes	varchar(100)
Actual	varchar(100)
Fecha_Campo	datetime
Autorizo	varchar(20)

Bajas

Nombre de columna	Tipo comprimido
Cve_Unica	int
u_nco	varchar(10)
causa	text
u_enc	varchar(6)
Fecha	datetime
Autorizo	varchar(20)
Mov	varchar(4)
Cve_Ref	int

Cambios_movimiento

Nombre de columna	Tipo comprimid
Cve_Unica	int
u_enc	char(20)
u_nco	char(10)
Campo	varchar(20)
Tipo	varchar(1)
Antes	varchar(100)
Actual	varchar(100)
Fecha_Campo	datetime
Autorizo	varchar(20)
Mov	varchar(3)

Plantilla 12.-

12.- ARQUITECTURA DE SOFTWAREPRUEBAS

Compuesta por:

- Casos de Uso
- Diagrama de Despliegue
- Diagrama de Clases
- Modelo de Datos.

Plantilla 13.-

13. PRUEBAS
Versión <1.0 BETA>

Fecha de la prueba <27/02/10>

PRUEBAS

CUADRO 10: Pruebas

Prueba No.	**1**				
Tipo de Prueba:	**Unitaria**	**Componentes**	**Integral**	**Desempeño**	**Aceptación**
	1				
Responsable:					
Objetivo de la Prueba: Verificar la eficiencia de el algoritmo al realizar los empates, calibrar los parámetros para la identificación de unidades económicas bajo la denominación de Razón Social y/o Nombre del Establecimiento					
Casos de Uso Relacionado: Filtering, ASM, Empate, Estandariza					
Entradas:					
Restricciones: El Servidor SQL Server 2005 esta montado en un equipo con ¡GB de RAM					
Salidas Obtenidas			**Satisfactorio**		
			SI	**NO**	
Excepciones: [Casos excepcionales. Situaciones que debemos tener en cuenta que pueden pasar. Se indica también qué se hace cuando ocurre la excepción. Por ejemplo: División por cero, etc.]					
Observaciones:					

[Nombre y rol]	[Nombre y rol]	[Nombre y rol]
Realiza	Verifica	Autoriza

Plantilla 14.-

14. SEGUIMIENTO DE PRUEBAS
Versión <1.0 Beta>

Fecha de creación del documento <dd/mm/aa>

SEGUIMIENTO DE PRUEBAS

CUADRO 11: Seguimiento de pruebas.

No. Prueba	Observaciones	Estatus	Revisó	Atendió	Fecha de Liberación
1	Parámetros de algoritmos para identificación de unidades económicas bajo la denominación de Nombre del Establecimiento y/o Razón Social	Corriendo prueba	Sara	David	

Status:

NP No Procede

P Pendiente

C Corregido

[Revisó.- Corresponde a la persona que realiza la prueba.]

[Atendió.- Se refiere al programador que atiende la observación.]

Plantilla 15.-

15. SOLICITUD DE SERVICIO PARA CENTRO DE PRUEBAS
Versión <1.0>

Fecha de creación del documento <dd/mm/aa>

Plantilla no utilizada en este prototipo.

Plantilla 16.-

16. LIBERACIÓN Y ACEPTACIÓN DE SISTEMA
Versión <1.0>

Fecha de creación del documento <dd/mm/aa>
Fecha de última actualización <dd/mm/aa>

Nota Aclaratoria: Las Plantillas 15 y 16 no se llenan debido a que como es una propuesta de prototipo hasta estar aceptada la propuesta se realizará el trámite correspondiente.

Plantilla 17.-

17. CONTROL DE CAMBIOS
Versión <1.0 Beta>

Fecha de Solicitud del Cambio <dd/mm/aa>

CONTROL DE CAMBIOS

CUADRO 12: Control de cambios

Módulo:	Empates
Caso de uso:	Q-gram
Justificación:	Calibrar los parámetros que identificarán las unidades económicas que se encuentren baja la misma denominación de Nombre del Establecimiento y/o Razón Social
Descripción:	Debe realizarse un análisis cuales son los mejores parámetros para reducir los universos de empates de tal forma que se optimicé la búsqueda con el menor número de registros a comprar, sin arriesgar que la creación de los subuniverso de búsqueda deje fuera registros posibles a encontrarse bajo la misma denominación de Nombre de Establecimiento o Razón Social
Responsable:	Analista del prototipo
Observaciones:	Las primeras pruebas son utilizando una longitud de caracteres de 3 para el q-gram, utilizando antes de este el subduniverso de tomar los registros que contengan las dos palabras más largas de la cadena

Módulo:	Empates
Caso de uso:	ASM
Justificación:	Calibrar los parámetros que identificarán las unidades económicas que se encuentren baja la misma denominación de Nombre del Establecimiento y/o Razón Social
Descripción:	Debe realizarse un análisis de cuales son los mejores parámetros para reducir los universos de empates de tal forma que se optimicé la búsqueda con el menor número de registros a comprar, sin arriesgar que la creación de los subuniverso de búsqueda deje fuera registros posibles a encontrarse bajo la misma denominación de Nombre de Establecimiento o Razón Social
Responsable:	Analista del prototipo
Observaciones:	Las primeras pruebas son utilizando una longitud de caracteres de 3 para el q-gram, utilizando antes de este el subduniverso de tomar los registros que contengan las dos palabras más largas de la cadena

Si el cambio es en pantalla:

Especificaciones de Cambios:
Motivos del Cambio:
Pantalla afectada por el cambio

Si el cambio es en validaciones

Especificaciones de Cambios
Motivos del cambio

Solicitante del cambio	Recibe solicitud

Nota: Para el llenado de este formato se deberá sustituir el texto que se encuentra entre corchetes con la información que en cada sección se especifica

ANEXO III

Lista de programas SAS, SHAMSA

Listing of SAS programs and macros for the IDB linking process

---

#asm.sas Calculates 'distance' between strings using aproximate string matching algorithm.

#bigram.sas Calculates 'distance' between strings using bigram string comparison.

#compare.sas Combine and data and compare identifying variables.

#dobcomp.sas Calculates the relative 'distance' between two birth dates.

#iddata.sas Variable attributes for IDMASTER data.

#initwt.sas Determine initial weights for all variables.

#iterwts.sas Developes the weights for all variables.

#join.sas Combines two data sets using specified criteria and creates a data view of joined data.

#joindata.sas Combines client obs to determine weights and links.

#mtchcls.sas Classifys joined client pairs as matched or not matched using scores and the upper threshold.

#ncomp.sas Compares two (arrays) of names and returns a numerical score indicating the degree of agreement.

#nysiis.sas Creates phonetic code for name besed upon the New York State Identification and Intelligence System

#review.sas Creates a report for manual review of linked client record pairs.

#scale.sas Creates a data set listing the variables of the supplied data set.

#scalewt.sas Adjusts weights for first and last names, date of birth, and ZIP codes according to the appropriate

#thrshld.sas Calculates matched/nonmatched/uncertain threshold from scored data and creates upper and lower bound

#topn.sas Generates frequency distributions listing only the top n most numerous values for standard variables

#trans.sas Translates multiple, associated links to a common link.

#twinchk.sas SAS macro to identify high scores resulting from the pairing of twins rather than multiple IDs for the

#weights.sas summarizes data by matched/nonmatched and calculates new weights.

#winkler.sas Calculates 'distance' between strings using algorithm devised by William Winkler.

#zipdist.sas Calculates distance between ZIP codes based upon the longitude and latitude of each ZIP code's centroid.

zips.csv ZIP Code centroid information used with the mcaro "zipdist"

Step350.sas First step in client linking program. Combines and scores the identification data.

Step352.sas Second step of client linking program, Completes the linking of ID files.

ANEXO IV

Estandarización de Caracteres

Tabla 3.- Estandarización de caracteres.

CHR	Carácter	Obs	Nombre del Establecimiento	Razón Social	Calle	Num. Ext	Colonia	Teléfono
			9	3	0	1	5	0
				Eliminar				
	No detectados		1	11	0	0	0	0
	No detectados		4	0	0	0	0	0
!			52	4	3	1	0	0
				Ø Eliminar= Ø				
"			19086	602	580	123	472	0
				Ø				
#			2172	356	1783	1862	231	0
				Ø Sustituir por Ñ cuando esta en medio de vocales a) Si antes y después tiene una vocal: YA#EZ, PE#UELAS. b) Aunque en algunas ocasiones es basura se recomienda dejarla, ya que algunos casos es parte de la Razón Social. BACHILLERATO TECNICO #18, EST #17. c) Existen casos en que viene al principio o al final, eliminarlo en estos casos.	Sustituir por Ñ 1766 Registros a) Eliminarla cuando esta al principio de la variable y al final. Nota: Se detectan algunos casos especiales INTERIOR DE LA SECUNDARIA 16, AV. FRANCISCO I. MADERO ESQUINA CON SUR 101 FRENTE AL # 120, CALZADA IGNACIO ZARAGOZA FRENTE #32, AV. FRANCISCO I. MADERO #9, CALLE 11, #137	Ø + casos Eliminar del principio y del final de la variable#19, 809#, Eliminar #, SIN#, S/#, #S-N, S-#, S#, S/# S/N#	Eliminar si viene al principio y al final de la variable.	
$			26	14	1	1	4	0
				Ø casos =Ñ siempre que tanga una vocal anterior y una posterior vocal$vocal and PERE$, Excepciones: CHICHIMECA$, MACILLA$.	3	Eliminar	Cambiar por Ñ	
%			129	45	80	0	8	0
				Por Ñ Cuando vocal%vocal	en algunos casos se utiliza como separador. Podemos sustituirla y eliminarla: SENDERO NACIONAL % AV. PROGREOS Y PRIV. MANUEL CASTAÑEDA , QUINTANA ROO % RAMIREZ Y LOPEZ PORTILLO (ES UN SEPARADOR ENTRE CALLES) Sustituir por Ñ cuando vocal%vocal NI%OS HEROES		Por Ñ, vocal%vocal, sólo una excepción COLONIA%VENUSTIANO	
&			84926	149416	43228	25	26315	0

				Por Ñ Cuando vocal&vocal	Por Ñ Cuando vocal&vocal	Eliminar	Por Ñ Cuando vocal&vocal, Revisar Barrio &IKAHUA, BARRIO &UCAHUA, BARRIO &UKAHUA, &UNDASU, &USACA, &UZACA, BARRIO I&I&UU, BARRIO SHINI&UU. Opción 1: 36 casos, sería cambiar todo a Ñ. Opción 2: 159 casos, dejar estos casos tal cual como están. Rev_e14_&.dbf	
'			8773	402	106	7	35	0
				397 reg. SPORT'S Es la razón social y si es posible que este correcta: 1º Se puede quedar tal cual 2º Se elimina - Eliminar (quitar) CHARLY'O, CHARLY'S	104 Casos POR ERROR D' en vez de 15D' SEPTIEMBRE, CALLLE 'B' CORRECTO: D' ANUNCIO, D' GYVES, SAM'S 1ª. Se propone sustituir un Ø al final quedaría D en el caso de error.	6 reg. Eliminar	34 reg. Error: D' SAN JUAN, Correcta: SAM'S, M'QUIRASCO, K'OLOK'IN	
(			5172	1309	1956	9	1223	67
				Eliminar (PARROCO), (SUPERVISOR), (VIUDA DE GUTIERREZ)	Eliminar (ESQUINA), (DEPORTIVO PLATEROS), (INSURGENTES NORTE)	Eliminar	Eliminar (colonia), (STA. ROSA)	
)			5000	1245	1909	4	1212	67
				Eliminar (SAGARPA), (SECUD), (ARZOBISPADO)	Eliminar	Eliminar	Eliminar (Portales)	
*			63	19	17	8	2	20
				Por Ñ cuando vocal*vocal - Eliminar al principio, al final ó que no se encuentre en medio de dos vocales.	Sustituir por un Ø	8 reg. Sustituir por un Ø	Sustituir por un Ø	
+			187	59	107	266	53	2
				59 reg. Revisar, puede ser Ó, É o puede ser OK	Eliminar al principio y al final de la cadena Excepciones: OBREG+N, PROLONGACI+N, M+XICO, H+ROES, PERIF+RICO, AM+RICA, P+RICON, S+PTIMA	263 reg. Sustituir por un Ø	Eliminar al principio y al final de la cadena. ANDR+S, CUAUTEP+C, Excepciones: G+MES, CONGREGACI+N.	
,			11873	9714	2321	172	1899	1
				9442 reg. Eliminar por un Ø	2008 reg. Eliminar por un Ø	151 reg. Sustituir por un Ø	1689 reg. Sustituir por un Ø	
-			23194	3529	23032	100296	2900	31
				Eliminar por un Ø				
.			153073	163763	100179	19313	47351	4
				Eliminar (quitar)	Eliminar (quitar)	18612 reg. Dejar OK	43891 reg. Eliminar (quitar)	
/			2188	126	1256	31125	1082	2

				Eliminar	Eliminar (quitar)	31 150 reg. Eliminar por un Ø	Eliminar (quitar) (Tip eliminar S/N)	
0		Válido	26932	8775	100081	812979	22659	3967
1		Válido	24539	6802	321646	1293776	36405	586560
2		Válido	25908	6316	230222	861536	40002	58282
3		Válido	16185	3721	134393	659267	15181	80589
4		Válido	11343	3336	106671	546176	7306	37012
5		Válido	10349	3184	157709	559122	14372	136904
6		Válido	7870	2522	128082	427048	8974	106096
7		Válido	8518	2801	80841	393573	7338	52831
8		Válido	7013	2425	74850	367524	7582	95725
9		Válido	6403	2584	64986	355013	7249	40556
:			118	55	453	5	90	0
				53 reg. Eliminar por un Ø	Eliminar (quitar), tener cuidado con: AVENIDA:BENTO	5 reg. Eliminar por un Ø	Eliminar (quitar) por un Ø	
;			58	39	64	78	46	0
				39 reg. Eliminar (quitar) en caso de estar entre dos dígitos, sustituir por Ø	64 REG. Eliminar (quitar)	Eliminar (quitar)	42 reg. Eliminar (quitar)	
<			9	19	17	0	5	0
				19 reg. Eliminar (quitar)	Eliminar (quitar)	No tiene	Eliminar (quitar)	
=			13	5	5	0	1	0
				Eliminar (quitar) 1 caso 1+1=3 asociados	Eliminar (quitar)	No tiene	1 reg. Eliminar por un Ø	
>			1	1	0	0	0	0
				Eliminar (quitar)		No tiene	No hay	
?			20	9	7	8	4	0
				4 reg. Eliminar (quitar)	Eliminar por un Ø Excepción: TREVI?O=TREVIÑO	3 reg. Eliminar palabras que contengan ? Eliminar ? (quitar)	4 reg. Corregir a mano algunas ocasiones es Ñ en otras Ó etc. É	
@			1227	34	3	0	3	0
				Ñ Cuando este entre dos vocales vocal@vocal	Sustituir por Ñ	No hay	3 reg. Revisar a mano son pocos registros y no hay relación.	
A		Válido	9747835	11265873	9946261	50587	8059887	0
B		Válido	1794996	863330	699527	28530	706406	0
C		Válido	3569001	3162716	4160069	9187	5230494	1
D		Válido	3244697	2886622	2575440	3485	1395191	0
E		Válido	9131429	8238197	7719294	4830	4232001	10
F		Válido	557321	554018	358092	24280	361652	0
G		Válido	717467	1823977	880972	540	523825	0
H		Válido	472877	752562	410447	447	291988	0
I		Válido	6469403	5425398	3947139	5082	5211357	1
J		Válido	372914	838892	415975	427	331254	0
K		Válido	82756	38193	20082	30450	9941	0
L		Válido	3954015	4363112	6588532	5397	4769668	0
M		Válido	2420522	2451683	1496130	32733	1082054	0
N		Válido	5122157	4488419	3978584	998588	5977070	4

O		Válido	5934594	5987890	4404843	3908	9153380	1
P		Válido	1594823	1080154	946872	1440	795141	0
Q		Válido	190801	234244	157547	346	76053	0
R		Válido	6588542	6974995	3869546	843	3685297	0
S		Válido	4937285	3861428	1953676	1019976	2039294	2
T		Válido	3891844	2352997	1868806	4767	2461226	5
U		Válido	1898337	2490983	1562532	3439	1123257	0
V		Válido	873559	1271982	1443591	52	435577	0
W		Válido	22806	15386	4606	102	1415	0
X		Válido	170939	95468	144536	232	105833	5
Y		Válido	599834	336046	313658	690	127356	0
Z		Válido	480579	2511004	943651	660	551119	0
[			5	2	1	0	1	0
				Eliminar (quitar)	Eliminar (quitar)	No hay	Eliminar (quitar)	
****			2	1	0	15	1	0
				1 reg. Eliminar	No hay	15 reg. Eliminar	1 reg. Eliminar	
]			5	4	1	0	1	0
				4 reg.	Eliminar (quitar)	No hay	1 reg. Eliminar (quitar)	
_			194	52	283	797	39	0
				Eliminar, sustituir por Ø	Eliminar, sustituir por Ø	798 reg. Eliminar, sustituir por Ø	39 reg. Eliminar por Ø	
`			107	16	5	0	7	0
				Eliminar (quitar)	Eliminar (quitar)	No hay	Eliminar (quitar)	
a		Convertir a Mayúsculas	167	36	298	2	201	0
b		Convertir a Mayúsculas	32	0	23	0	12	0
c		Convertir a Mayúsculas	80	26	68	0	34	0
d		Convertir a Mayúsculas	80	7	59	0	64	0
e		Convertir a Mayúsculas	177	57	200	1	92	0
f		Convertir a Mayúsculas	3	9	7	0	6	0
g		Convertir a Mayúsculas	1	2	19	0	11	0
h		Convertir a Mayúsculas	7	0	12	0	6	0
i		Convertir a Mayúsculas	169	38	110	0	106	0
j		Convertir a Mayúsculas	10	0	12	0	9	0
k		Convertir a Mayúsculas	0	0	2	12	4	0

l		Convertir a Mayúsculas	80	2	120	0	92	0
m		Convertir a Mayúsculas	82	25	41	29	18	0
n		Convertir a Mayúsculas	115	21	95	80	81	0
o		Convertir a Mayúsculas	111	15	171	4	82	1
p		Convertir a Mayúsculas	16	0	20	0	17	0
q		Convertir a Mayúsculas	1	8	9	0	26	0
r		Convertir a Mayúsculas	91	13	181	0	114	0
s		Convertir a Mayúsculas	73	17	75	80	66	0
t		Convertir a Mayúsculas	58	10	75	0	64	1
u		Convertir a Mayúsculas	57	10	42	0	76	0
v		Convertir a Mayúsculas	3	0	36	0	9	0
w		Convertir a Mayúsculas	0	0	1	0	0	0
x		Convertir a Mayúsculas	4	18	5	1	1	0
y		Convertir a Mayúsculas	20	8	11	0	9	0
z		Convertir a Mayúsculas	7	0	20	0	7	0
{			2	2	1	0	1	0
				Eliminar (quitar)	Eliminar (quitar)	No hay	Eliminar (quitar)	
|			5	2	26	12	18	0
				Eliminar (quitar)	Eliminar (quitar)	Eliminar	Eliminar (quitar)	
}			67	65	71	74	70	0
				Eliminar (quitar) por un Ø	Eliminar (quitar) por un Ø	Eliminar (quitar)	Eliminar (quitar)	
~			1	0	0	0	0	0
						Eliminar	Eliminar (quitar)	
é	negrilla		0	0	1	0	1	0
				negrilla				
í	=		27	10	4	0	3	0
				=				
ó	Ë		26	4	4	0	1	0
ñ	¤		1	1	1	0	0	0
				Sustituir				

Ñ	¥	50	52	15	0	12	0
			Sustituir				
ª	¦	20	8	30	0	3	0
			Eliminar (quitar) por un Ø	Eliminar quitar	No hay	Eliminar quitar	
¿	¨	36	2	0	0	0	0
			Eliminar quitar	Eliminar quitar			
¬	¼	10	3	32	0	10	0
			Eliminar quitar	Eliminar quitar			
½	«	1	0	0	0	0	0
			Eliminar quitar	Eliminar quitar			
¼	╝	1	0	25	0	5	0
░	°	894	190	1561	127	493	0
▒	±	0	0	3	0	0	0
┤		3057	154	71	0	40	0
À	└	0	0	2	0	0	0
╢		69	22	188	14	116	0
╝	¼	6	0	0	0	0	0
¢	½	0	0	0	3	0	0
			Sustituir por O	Sustituir por O	No hay	Sustituir por O	
¥	Ñ	2	2	2	0	0	0
			Sustituir por Ñ	Sustituir por Ñ	No hay	Sustituir por N	
┐	ć	32	25	31	2	18	0
└	À	3	2	13	0	2	0
┴	À	82	105	101	0	76	0
├	Ã	0	0	0	0	2	0
├		0	0	1	0	0	0
			Eliminar	Eliminar			
├		0	1	0	0	0	0
			Eliminar				
─		1	0	0	0	0	0
			Ä				
Ã	Ç	7	6	6	1	2	0
╚	È	2	5	5	0	1	0
╒	É	57	100	128	0	31	0
╩	Ê	4	3	0	0	3	0
╤	Ë	8	3	32	0	14	0

╠	Ì		0	0	4	0	0	0
═	Í		93	133	95	0	49	0
╬	Î		0	1	0	0	0	0
¤	ñ		0	0	0	0	31	0
				Sustituir por Ñ	Sustituir por Ñ			
¤	ñ		0	0	2	0	0	0
				Sustituir por Ñ	Sustituir por Ñ			
ð	Ð		3596	4537	490	0	426	0
Ð	Ñ		4301	8262	5193	4	2999	0
Ê	Ò		4	1	9	0	5	0
Ë	Ó		122	151	176	0	94	0
È	Õ		0	0	1	0	0	0
ı	Ö		0	1	0	0	0	0
Í	═		9	8	7	0	1	0
┘	Ù		0	0	0	0	1	0
┘			0	0	3	0	0	0
┘			0	2	0	0	0	0
┌	Ú		39	39	16	0	5	0
▉	Û		8	2	0	0	0	0
▄	Ü		1023	260	63	0	13	0
¦			6	1	0	0	0	0
				Eliminar quitar	Eliminar quitar	Quitar	Eliminar quitar	
■	ß		0	0	0	0	1	0
Ó	à		16	11	1	0	9	0
ß	á		47	103	38	0	4	0
Ú	é		31	67	6	1	6	0
ý	ì		1	0	0	0	0	0
Ý	í		83	110	25	0	2	0
	ð		593	847	3	0	2	0
±	ñ		1279	2269	1287	0	436	0
¾	ó		128	86	54	0	25	0
¶	ô		2	0	0	0	0	0
÷	ö		1	1	0	0	0	0
¸	÷		0	0	8	0	0	0
·	ú		12	18	0	0	28	0
¹	û		1	0	0	0	0	0
³	ü		48	18	5	0	0	0

20.- Referencias Bibliográficas.

Hal Berghel & David Roach. (1996). An extension of Ukkonen's enhanced dynamic programming ASM algorithm. ACM Transactions on Information Systems (TOIS), Volume 14, Issue 1 (January 1996) (Issue 1), 94 - 106

Su Yan, Dongwon Lee, Min-Yen Kan, & C. Lee Giles. (s.d.). Adaptive sorted neighborhood methods for efficient record linkage (págs. 185 - 194). Presented at the International Conference on Digital Libraries archive Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, Vancouver, BC, Canada. Recuperado a partir de http://portal.acm.org/citation.cfm?id=1255175.1255213.

Carlos Avendaño, Claudia Feregrino, & Gonzalo Navarro. Mejorando un Algoritmo para Búsqueda Aproximada (pág. 14). Presented at the XIII International Congress on Computing 2004, México, D.F. Recuperado a partir de http://www.dcc.uchile.cl/~gnavarro/ps/cic04.pdf.

CEPAL. (2003, Mayo 6). DIRECTORIOS ESTADÍSTICOS DE EMPRESAS ELABORADOS A PARTIR DE REGISTROS ADMINISTRATIVOS. LC/L.1892(CEA.2003/7) 6 de mayo de 2003. Recuperado a partir de http://www.eclac.cl/ceacepal/documentos/lcl1892p.pdf.

CEPAL. (2001, Enero 24). ESTABLECIMIENTO DE LA CONFERENCIA ESTADÍSTICA DE LAS AMÉRICAS DE LA COMISIÓN ECONÓMICA PARA AMÉRICA LATINA Y EL CARIBE (Resolución 2000/7 del Consejo Económico y Social de las Naciones Unidas). LC/L.1475(CEA.2001/5) 24 de enero de 2001. Recuperado a partir de http://www.un.org/spanish/ecosoc/docs/index.shtml.

Comunidad Andina. (279, Septiembre 30). LEGISLACIÓN ESTADÍSTICA COMUNITARIA Comunidad Andina. Comunidad Andina.

Comunidad Andina. (s.d.). RESOLUCIÓN 1274 Guía para la construcción de los Directorios de Empresas con fines estadísticos en la Comunidad Andina.

Comunidad Andina. Recuperado a partir de http://www.comunidadandina.org/andestad/areas_tematicas.asp?id=13&m=2.

Comunidad Andina. (s.d.). RESOLUCION 1218 Cobertura de los Directorios de Empresas Comunidad Andina. Recuperado a partir de http://www.comunidadandina.org/normativa/res/R1218sg.htm.

Comunidad Andina. (s.d.). RESOLUCIÓN 1273 Manual de Recomendaciones sobre los Directorios de Empresas con fines estadísticos en la Comunidad Andina. Comunidad Andina. Recuperado a partir de http://www.comunidadandina.org/normativa/res/R1273sg.htm.

CONSEJO DE LAS COMUNIDADES EUROPEAS. (1993, Marzo 15). REGLAMENTO (CEE) No 696/93 DEL CONSEJO de 15 de marzo de 1993 relativo a las unidades estadísticas de observación y de análisis del sistema de producción en la Comunidad EL CONSEJO DE LAS COMUNIDADES EUROPEAS. EUROSTAT. Recuperado a partir de http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993R0696:ES:HTML.

Daniel Whalen, Anthony Pepitone, Linda Graver, & Jon D. Busch. (2000, Julio). Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration. Recuperado a partir de http://www.csat.samhsa.gov/IDBSE/idb/modules/linking/recordlink.aspx.

e u r o s t a t. (2003, Marzo). MANUAL DE RECOMENDACIONES SOBRE LOS REGISTROS DE EMPRESAS DE LA Unión Europea (e u r o s t a t). Recuperado a partir de http://secgen.comunidadandina.org/andestad/adm/upload/file/manual.pdf.

EL CONSEJO DE LAS COMUNIDADES EU. (1993, Julio 22). Reglamento (CEE) nº 2186/93 del Consejo, de 22 de julio de 1993, relativo a la coordinación comunitaria del desarrollo de los registros de empresas utilizados con fines

estadísticos. EUROSTAT. Recuperado a partir de http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993R2186:ES:HTML.

EUROSTAT. (2005, Mayo 24). Código de buenas prácticas de las estadísticas europeas. Recuperado a partir de http://secgen.comunidadandina.org/andestad/adm/upload/file/codigo.pdf.

, I. (2004). METODOLOGIA DE LOS CENSOS ECONOMICOS 2004. Recuperado a partir de http://www.inegi.org.mx/est/contenidos/espanol/metodologias/censos/metodo_ce2004.pdf.

Iván Amón, & Claudia Jiménez. (2009). Hacia una Metodología para la Selección de Técnicas de Depuración de Datos (pág. 6). Presented at the IV CONGRESO COLOMBIANO DE COMPUTACIÓN - 4CCC 2009, BUCARAMANGA, COLOMBIA. Recuperado a partir de http://serverlab.unab.edu.co:8080/wikimedia/memorias/fullpapers/15.pdf.

Ivan P. Fellegi, & Alan B. Sunter. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, Vol. 64, No. 328 (Dec., 1969),(328), pp. 1183- 1210.

José Antonio González Madrid, & Alberto Lezcano Lastra. (2008). Normalización de direcciones postales (pág. 17). SANTANDER, ICANE, Instituto Cantabro de Estadística. Recuperado a partir de http://www.jecas.org/descargas.html.

LA COMISIÓN DE LA COMUNIDAD ANDINA, (s.d.). DECISION 698 Creación y actualización de Directorios de Empresas. Comunidad Andina. Recuperado a partir de http://www.comunidadandina.org/normativa/dec/D698.htm.

Lifang Gu, & Rohan Baxter. (2004). Adaptive Filtering for Efficient Record Linkage. En e.g. 001 (pág. 5). Presented at the Proceedings of the 2004 SIAM International Conference on Data Mining, Neshville, NT. Recuperado a partir de http://www.siam.org/proceedings/datamining/2004/dm04.php.

Mexicanos, F. D. J. C. H, P. D. L. E. U., & Honorable Congreso de la Unión. (2008, Abril 16). LEY DEL SISTEMA NACIONAL DE INFORMACIÓN ESTADÍSTICA Y GEOGRÁFICA.

Rohan Baxter, Peter Christen, & Tim Churches. (2003). A Comparison of Fast Blocking Methods for Record Linkage (pág. 3). Presented at the The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC. Recuperado a partir de http://datamining.anu.edu.au/publications/.

Rohan Baxter, Peter Christen, & Tim Churches . (2003). A Comparison of Fast Blocking Methods for Record Linkage (pág. 6). Presented at the The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC. Recuperado a partir de http://datamining.anu.edu.au/publications/.

Vicenta Mardones, & Mauricio Ponce. (2008). Directorios de Empresas y Establecimientos: Revisión de la experiencia internacional y situación en algunos países de América Latina 1 Vicenta Mardones y Mauricio Ponce. En CEPAL, Banco Interamericano de Desarrollo (pág. 71). Presented at the Taller “Directorios de empresas y establecimientos: Desarrollos recientes y desafíos actuales y futuros en América Latina” Santiago de Chile, 22 al 23 de Septiembre de 2008, Santiago de Chile. Recuperado a partir de http://www.eclac.org/scaeclac/documentos/2008_09_CEA_tallerDirectorios_MARDONESPAPER.pdf.

Vitoria-Gasteiz, Josu Iradi Arrieta, Leire Legarreta, & Laura Otero. (2007, MRZO). MÉTODOS AUTOMÁTICOS DE FUSIÓN DE REGISTROS Y SU UTILIZACIÓN EN EUSTAT. EUSKAL ESTATISTIKA ERAKUNDEA INSTITUTO VASCO DE ESTADISTICA.

Páginas WEB

[2] new_ley383 LEY DEL SISTEMA NACIONAL DE INFORMACIÓN ESTADÍSTICA Y GEOGRÁFICA
http://www.inegi.org.mx/est/contenidos/espanol/metodologias/censos/metodo_ce2004.pdf
Instituciones y otros Órganos de la Unión Europea.
http://europa.eu/institutions/index_es.htm)
[9] DOUE. Diario Oficial de la Unión Europea, http://vlex.com/source/doue-23/issue/1993/8/5
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993R2186:ES:HTML
Reglamento (CEE) nº 2186/93 del Consejo, de 22 de julio de 1993, relativo a la coordinación comunitaria del desarrollo de los registros de empresas utilizados con fines estadísticos.
DOUE http://vlex.com/source/doue-23/issue/1993/8/5

[10] Reglamento (CEE) nº 696/93 del Consejo, de 15 de marzo de 1993, relativo a las unidades estadísticas de observación y de análisis del sistema de producción en la Comunidad, Diario Oficial n° L 076 de 30/03/1993 p. 0001 – 0011, Edición especial en finés ...: Capítulo 13 Tomo 24 p. 0007, Edición especial sueca...: Capítulo 13 Tomo 24 p. 0007

[11] Manual de recomendación sobre directorios de empresas, manual.pdf, http://secgen.comunidadandina.org/andestad/adm/upload/file/manual.pdf
http://www.comunidadandina.org/andestad/normativa_europea.asp?m=3.

[12] Código de Buenas Practicas de las Estadísticas Europeas
http://secgen.comunidadandina.org/andestad/adm/upload/file/codigo.pdf
http://www.comunidadandina.org/andestad/areas_tematicas.asp?id=13

(código de buenas practicas Legislación Estadística, Normativa Europea

http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC&StrLanguageCode=EN&CFID=17616854&CFTOKEN=16f6680-00017fcc-f9el-1513-976c-8:

[14] http://www.comunidadandina.org/andestad/areas_tematicas.asp?id=13&m=2

[15] NORMATIVIDAD ANDINA, DESICIONES, http://www.comunidadandina.org/normativa/dec/decnum.htm, http://www.comunidadandina.org/normativa/dec/D698.htm

[16] NORMATIVIDAD ANDINA, RESOLUCIONES, http://www.comunidadandina.org/normativa/res/resoluciones.htm, http://www.comunidadandina.org/normativa/res/R1218sg.htm

[17] NORMATIVIDAD ANDINA, RESOLUCIONES, http://www.comunidadandina.org/normativa/res/resoluciones.htm, http://www.comunidadandina.org/normativa/res/R1273sg.htm, Publicado en la Gaceta Oficial 1748

[18] NORMATIVIDAD ANDINA, RESOLUCIONES, http://www.comunidadandina.org/normativa/res/resoluciones.htm, http://www.comunidadandina.org/normativa/res/R1274sg.htm, Publicado en la Gaceta Oficial 1749

[19] http://www.comunidadandina.org/andestad/areas_tematicas.asp?id=13&m=2

[20] http://www.cinu.org.mx/onu/estructura/ecosoc.htm, Consejo Económico y Social (ECOSOC)

[21] http://www.eclac.org/cgi-bin/getprod.asp?xml=/noticias/paginas/4/21324/P21324.xml&xsl=/tpl/p18f-st.xsl&base=/tpl/top-bottom_acerca.xsl ACERCA de CEPAL

[22] http://www.eclac.org/cgi-bin/getprod.asp?xml=/noticias/paginas/9/21469/P21469.xml&xsl=/tpl/p18f-st.xsl&base=/tpl/top-bottom_acerca.xsl MANDATO Y MISION CEPAL

[23] Resolución 2000/7 http://www.un.org/spanish/ecosoc/docs/index.shtml

[24] “Directorios estadísticos de empresas elaborados a partir de registros administrativos” (LC/L.1892(CEA.2003/7 http://www.eclac.cl/ceacepal/documentos/lcl1892p.pdf

[25] http://www.eclac.cl/publicaciones/xml/8/30028/LCL2795e.pdf

[26] http://www.eclac.org/scaeclac/taller_directorios_empresas_2008.htm

[30] Normalización de Direcciones Postales, http://www.jecas.org/descargas.html

[31] http://serverlab.unab.edu.co:8080/wikimedia/memorias/full_papers.html

[32] http://www.gartner.com/it/page.jsp?id=501733

ASM

[33] http://csat.samhsa.gov/programs.aspx

[34] http://www.csat.samhsa.gov/IDBSE/idb/modules/linking/recordlink.aspx

[35] http://www.csat.samhsa.gov/IDBSE/idb/modules/linking/recordlink.aspx
módulos linkage SAMHSA
http://www2.inegi.gob.mx/sneig/contenidos/espanol/superior/lonuevo.aspx

[36] http://www.berghel.net/publications/asm/asm.php

[37] http://www.cs.helsinki.fi/u/ukkonen/InfCont85.PDF

[38] http://portal.acm.org/citation.cfm?id=214183