CENTRO DE CIENCIAS BÁSICAS

MAESTRÍA EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES

“TESINA”

“Desarrollo de un Prototipo de Data Mart con Esquema Constelación y Variables Armonizadas para la obtención de Indicadores Principales y Comparables en Proyectos del INEGI”

PRESENTA

I.S.C. Martha Silvia Serrano Rentería

DIRECTOR DE TESIS

M. C. César Eduardo Velázquez Amador

SINODALES

Dr. Juan Muñoz López

M. C. Jorge Eduardo Macías Luévano

Cd. Universitaria, Junio del 2010.

AGRADECIMIENTOS

A Dios.
Por darme la vida, por su amor infinito y por colocarme en el lugar indicado en el momento preciso.

A mi mamá.
Por su cariño, entrega, oraciones y apoyo hasta en los momentos difíciles en los cuáles con solo escucharla todo era remediado.

A mi papá.
Por ser un gran ejemplo de inteligencia y tenacidad, por su dedicación y amor por nuestra familia.

A Rafael.
Por ser parte de mi vida, por brindarme tanta paciencia y comprensión. Y sobre todo por el amor y cariño que nos une.

A mi familia.
A mis hermanos, cuñados, tías, sobrinas. Fer, Montse, Vale y AnaPau gracias por tantos momentos de felicidad que me han regalado.

A mis maestros y compañeros.
A todos aquellos que me han compartido su conocimiento y experiencia, especialmente al Maestro César Velázquez que con su guía fue posible el buen término de este trabajo.

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

I.S.C. MARTHA SILVIA SERRANO RENTERÍA
PASANTE DE LA MAESTRÍA EN INFORMÁTICA
Y TECNOLOGÍAS COMPUTACIONALES
P R E S E N T E .

Estimado (a) Alumno (a) Serrano:

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: **"Desarrollo de un Prototipo de Data Mart con Esquema Constelación y Variables Armonizadas para la obtención de Indicadores Principales y Comparables en Proyectos del INEGI",** 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., 2 de junio de 2010
"LUMEN PROFERRE"
EL DECANO

DR. FRANCISCO JAVIER ÁLVAREZ RODRÍGUEZ

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES
CENTRO DE CIENCIAS BÁSICAS

c.c.p.- Archivo

Por este conducto autorizamos a:

I.S.C. Martha Silvia Serrano Rentería

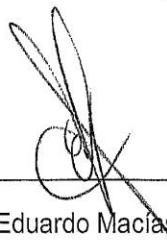
La impresión de su documento final de tesina, ya que cumple con los requisitos de contenido y forma exigidos por la Universidad Autónoma de Aguascalientes

Asesor

M. C. César Eduardo Velázquez Amador

Sinodales

Dr. Juan Muñoz López



M. C. Jorge Eduardo Macías Luévano

Resumen/ Abstract

Actualmente las organizaciones se enfrentan a la generación y explotación de grandes volúmenes de información, sin embargo, no en todos los casos dicha información es útil para ciertos procesos importantes; pues en ocasiones se requiere información menos detallada.

Dentro del Instituto Nacional de Estadística y Geografía se tiene información sobre proyectos importantes como el Censo de Población y Vivienda 2000, Conteo de Población y Vivienda 2005 y Encuesta Nacional de Ocupación y Empleo (ENOE), sin embargo, la Subdirección de Diseño Muestral de Vivienda necesita que dicha información se encuentre armonizada y almacenada de tal manera que se puedan obtener los indicadores principales para el cálculo de tamaños de muestra de encuestas en hogares.

Para la obtención de dichos indicadores se sigue actualmente un método tradicional y se propone seguir una nueva metodología a través de un Data Mart Esquema Constelación, el cuál contiene la información armonizada de los proyectos Censo de Población y Vivienda 2000, Conteo de Población y Vivienda 2005 y Encuesta Nacional de Ocupación y Empleo (ENOE) a partir del año 2005.

Se analizaron las variables de los proyectos mencionados anteriormente y se les asignó un nombre específico basándose en la codificación de las variables del proyecto IPUMS (Integrated Public Use Microdata Series International) de ésta manera se cumplió con la armonización de variables y posteriormente se realizó el modelo del Data Mart basándose en la Metodología HEFESTO.

INDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICAS

CAPÍTULO I. Formulario del Problema

1.1 Contexto y Antecedentes Generales del Problema

El Instituto Nacional de Estadística y Geografía (INEGI) a través de la Subdirección de Diseño muestral de vivienda se encarga de realizar el diseño estadístico de las encuestas en hogares por medio de indicadores principales de encuestas. Por tal motivo, la armonización de variables de diversos proyectos es un tema que compete a dicha área.

El proyecto IPUMS-USA (Integrated Public Use Microdata Series-USA) fue desarrollado en 1997 por Ruggles, Sobek y otros, en el Population Center de la Universidad de Minnesota, dicho proyecto está basado en la armonización de microdatos censales desde el año de 1850 hasta 2000. En 1998 se extendió el proyecto IPUMS y fue llevado como prueba piloto a Colombia, para después extenderse a escala internacional naciendo el proyecto IPUMS-International. El primer grupo de países incorporados incluían Colombia, Francia, Kenya, México, Estados Unidos y Vietnam.

La armonización de microdatos censales internacionales se realizó en dos etapas, la primera etapa consistió en la estandarización de los formatos de los microdatos originales así como la corrección de errores descubiertos, la segunda etapa consistió en la armonización de variables, determinación de la disponibilidad de cada variable comparable, recopilación de la información existente, diseño de los esquemas de codificación de cada variable y su documentación correspondiente.

Sin embargo, no se tiene antecedente de la armonización de variables a nivel de encuestas con el nivel de detalle requerido por la Subdirección de Diseño Muestral de Vivienda.

La misión del Instituto Nacional de Estadística y Geografía (INEGI) con fundamento en Las Reglas Provisionales en Relación con la Gaceta del Senado de la Junta de Coordinación Política de fecha 11 de octubre del año 2006 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" (1), y su estructura organizacional (Figura 1) se conforma de una contraloría interna y siete direcciones generales, entre las que se encuentra la Dirección General de Estadísticas Sociodemográficas, la cual tiene como objetivos:

- Coordinar la generación de información estadística con base en el levantamiento de censos y encuestas, así como en la explotación de registros administrativos de las Unidades del Estado, de manera que contribuyan al conocimiento de la realidad nacional en el ámbito sociodemográfico, de gobierno, seguridad pública e impartición de justicia.
- Dirigir, con el apoyo de las Unidades Administrativas del Instituto, la realización de los censos nacionales de población y vivienda, los conteos nacionales de población, las encuestas en hogares, las encuestas especiales, y la explotación de los registros administrativos de carácter sociodemográfico, de gobierno, seguridad pública e impartición de justicia.
- Coordinar los procesos de diseño, captación, actualización, organización, procesamiento, integración y compilación de la información sociodemográfica, de gobierno, seguridad pública e impartición de justicia, y coadyuvar en la publicación y difusión de dicha información con la Dirección General del Servicio Público de Información y conservarla en los términos que al efecto determine la Dirección General de Coordinación del Sistema Nacional de Información Estadística y Geográfica.
- Fungir como Secretario Técnico de los Comités Ejecutivos del Subsistema Nacional de Información Demográfica y Social, y del Subsistema Nacional de Información de Gobierno, Seguridad Pública e Impartición de Justicia.
- Emitir opinión técnica, respecto de la pertinencia de que Unidades del Estado distintas al Instituto, realicen actividades estadísticas de carácter sociodemográfico, de gobierno, seguridad pública e impartición de justicia, en apego a los programas a que hace referencia el artículo 9 de la Ley, así como hacer del conocimiento de la Dirección General de Coordinación del Sistema Nacional de Información Estadística y Geográfica las recomendaciones para llevarlas a cabo.

- Coordinar la conservación de los metadatos o especificaciones concretas de la aplicación de las metodologías que se hubieren utilizado para la generación de información estadística en el ámbito de su competencia, así como, implementar mecanismos para el control, conservación y resguardo de la información a su cargo, en colaboración con la Dirección General de Coordinación del Sistema Nacional de Información Estadística y Geográfica. (2)

y tiene bajo su mando la Dirección General Adjunta de Encuestas Sociodemográficas y Registros Administrativos que a su vez tiene a su cargo la Dirección de Diseño y Marcos Estadísticos la cual tiene como objetivo el diseño estadístico de las encuestas en hogares, así como de la elaboración, mantenimiento y actualización de los marcos muestrales, a su vez, la Subdirección de Diseño Muestral de Vivienda es la encargada de elaborar, coordinar y supervisar los diseños para la generación de información básica en viviendas, así mismo es la responsable de seleccionar las unidades de muestreo para el levantamiento de las encuestas en hogares y generar los factores de expansión, entre otras actividades.(3) Para cumplir con dicho objetivo la Subdirección de Diseño Muestral de Vivienda se basa en diversos proyectos que se diseñan en el INEGI, entre las cuales se encuentran el Censo de Población y Vivienda 2000, el Conteo de Población y Vivienda 2005 y la Encuesta Nacional de Ocupación y Empleo para realizar el diseño muestral de encuestas especiales.

(1) www.inegi.org.mx

(2) www.inegi.gob.mx/inegi/contenidos/espanol/instituto/presi.asp?c=1610

(3) http://proyectos.inegi.gob.mx/de/ddme/default.aspx

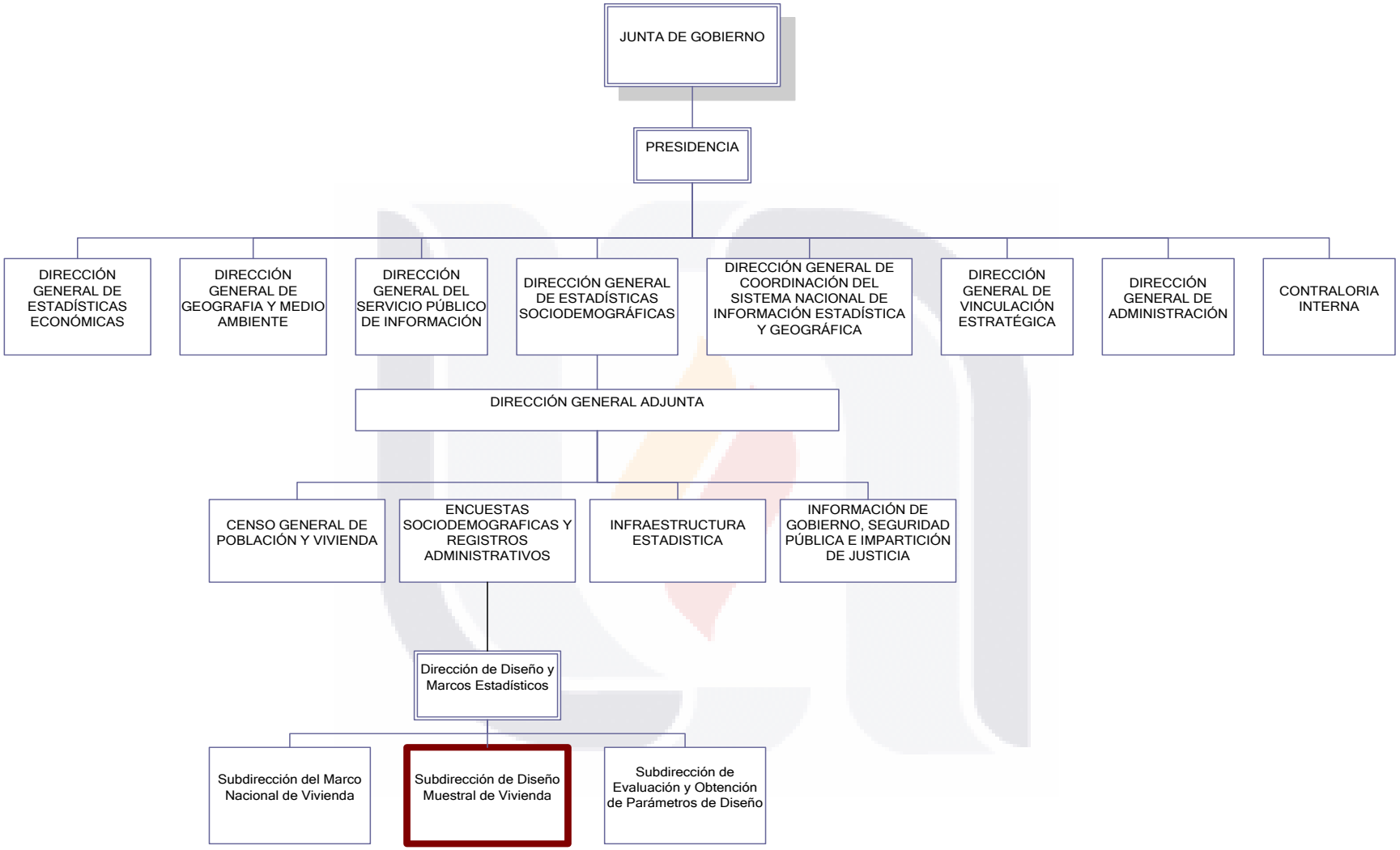


Figura 1; Estructura organizacional del Instituto Nacional de Estadística y Geografía. (DOF, 2009)

Los procesos de negocios se comenzaron a automatizar durante la década de los 80`s, esto ofreció a las organizaciones mejorar desde dentro y hacer realidad los beneficios asociados a dicha mejora, sin embargo con el tiempo el cúmulo de información que se generaba llego a ser poco útil para la organización pues se requería de información menos detallada, menos agregada o más integrada de tal manera que las organizaciones comenzaron a visualizar sus necesidades de información. A partir de ello nace el concepto de data warehouse (DW) proveyendo la facilidad para integrar los datos generados en un mundo de información no integrada. Un DW funcional organiza y almacena los datos necesarios para procesar información de análisis sobre una perspectiva histórica del tiempo. En el mismo camino se encuentra un data mart, siendo un DW a baja escala que apoya a una área particular y no a la empresa completa.

1.2 Situación Problemática

Para realizar el diseño muestral en encuestas especiales, es necesario contar con información de indicadores relevantes al tema de interés a partir de diversas encuestas en hogares que se diseñan en el INEGI, entre las cuales se encuentran el Censo de Población y Vivienda 2000, el Conteo de Población y Vivienda 2005 y la Encuesta Nacional de Ocupación y Empleo, a partir de hacer una comparación entre dichos indicadores se decide por aquella estimación que esté más acorde con las referencias de espacio y tiempo de interés. Para esto es necesario explorar la información disponible a diferentes niveles de dominio, la cual se genera en diferentes áreas y en consecuencia con distinto criterio de almacenamiento, sin embargo esta información no está contenida en un repositorio de datos y mucho menos con un formato armonizado, lo que implica inversión de recursos y tiempo en la explotación de dicha información. Algunos de los indicadores relevantes son publicados en la página web o en intranet del INEGI, sin embargo cada encuesta cuenta con diversos criterios a cubrir y esa flexibilidad no se muestra en las publicaciones.

En los últimos años se han realizado varias encuestas basadas en la ENOE, entre las cuales podemos citar: Modelo de Características de las Viviendas Deshabitadas (MOVIDE), Encuesta sobre Emprendedurismo, Encuesta Nacional sobre Uso de Tecnología de la Información en los hogares (ENDUTIH), etc. Para realizar la Encuesta de Emprendedurismo se requerían indicadores como Población Total, Ingresos, Nivel de Instrucción, Posición en la ocupación (Patrón o Cuenta Propia) y Años escolares cubiertos para personas de 12 a 29 años; dicha información no es posible encontrarla en publicaciones sobre la ENOE y por tal motivo fue indispensable explorar dicha información manualmente para el cálculo posterior del tamaño de muestra requerido; al hacerlo de forma manual se gastó demasiado tiempo en obtenerlo; lo cual nos implica demora en entrega de resultados al usuario final.

1.3 Relevancia del Caso

La información estadística obtenida de encuestas por muestreo es empleada para la toma de decisiones, por tal motivo los usuarios requieren que la información sea confiable, oportuna e integrada debido a que la demanda de encuestas es cada vez mayor, y esto agrava la situación de no contar con información armonizada para una mejor explotación y así proporcionar los resultados indispensables en base a estimación de parámetros de la población, para ello se utilizan diferentes fuentes de información principalmente de las encuestas que se han realizado en el pasado.

Al no contar con una armonización de variables de dichas encuestas, la obtención de información para la definición de los esquemas de muestreo y el cálculo de tamaños de muestra se hace muy compleja y requiere de tiempo para explotar la información, el cual es valioso en cualquier proyecto.

Los proyectos más importantes, en los que se basan la mayoría de las estimaciones para diversas encuestas especiales son Censo de Población y Vivienda 2000, Conteo de Población y Vivienda 2005 y Encuesta Nacional de Ocupación y Empleo (ENOE).

El Censo de Población y Vivienda 2000 tiene como objetivo general generar la información demográfica, socioeconómica y cartográfica necesaria para el país, con la máxima desagregación geográfica posible; enriquecer la serie histórica de datos estadísticos, manteniendo en lo posible la comparabilidad nacional e internacional, y permitir la construcción de marcos de muestreo para realizar encuestas en hogares. Tiene una periodicidad decenal, en años terminados en cero.

El Conteo de Población y Vivienda 2005 tiene como objetivo general producir información sociodemográfica básica, que actualice el conocimiento sobre el tamaño, la composición y la distribución territorial de la población, los hogares y las viviendas existentes en el país. Tiene una periodicidad decenal, en años terminados en cinco.

La ENOE tiene como objetivo principal obtener información estadística sobre las características ocupacionales de la población a nivel nacional, así como otras variables demográficas y económicas que permitan profundizar en el análisis de los aspectos

laborales. Tiene una periodicidad Trimestral, datos por área urbana, entidad federativa y a nivel nacional para cuatro tamaños de localidad, a partir del segundo trimestre de 2000 (con indicadores estratégicos de empleo publicados a través de Internet) y Anual para Datos básicos.

La solución a este problema se alberga en la construcción de un Data Mart que facilite la integración y almacenamiento de la información previamente armonizada, proporcionando a la Subdirección de Diseño Muestral de Vivienda los indicadores principales al tema de interés y de esta se tome la mejor decisión realizando una comparación entre ellos y así adoptar los indicadores más acordes con las referencias de espacio y tiempo de interés de la encuesta a realizar, de tal manera que los tiempos de respuesta en cada proyecto que les soliciten sea menor.

1.4 Objetivos, Preguntas y Proposiciones del Caso.

1.4.1 Objetivo General

Generar el prototipo de un Data Mart con esquema Constelación de proyectos con variables armonizadas para la obtención de sus indicadores principales y comparables.

1.4.2 Objetivos Específicos.

OE1. Determinar si es posible realizar un Data Mart esquema Constelación con variables armonizadas de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

OE2. Facilitar la obtención de indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

OE3. Comprobar la disminución del tiempo de respuesta para la obtención de indicadores principales de los proyectos ENOE del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 a través de la Metodología Propuesta que hace uso del Data Mart con Esquema Constelación.

1.4.3 Preguntas

Pregunta 1. ¿Es posible realizar un Data Mart con esquema Constelación con variables armonizadas de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2000?

Pregunta 2. ¿Cuántos procesos intervienen en la metodología tradicional y cuántos procesos intervienen en la metodología propuesta usando el Data Mart con Esquema Constelación para obtener indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005?

Pregunta 3. ¿Cuántas personas se necesitan para obtener los indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 con la metodología tradicional y con la metodología propuesta?

Pregunta 4. ¿Cuántas tablas se necesitan acceder para obtener los indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 con la metodología tradicional y con la metodología propuesta?

Pregunta 5. ¿Se obtiene una disminución de tiempo de respuesta para la obtención de indicadores principales de los proyectos ENOE del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 a través de la Metodología Propuesta que hace uso del Data Mart con Esquema Constelación?

1.4.4 Proposiciones

Proposición 1. El proceso de armonización de variables y el modelo del Data Mart con Esquema Constelación permite armonizar y almacenar la información de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

Proposición 2. El modelo del Data Mart con Esquema Constelación facilita la obtención de indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

Proposición 3. Con la metodología propuesta usando el Data Mart con Esquema Constelación se necesita un menor número de procesos para obtener los indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

Proposición 4. Con la metodología propuesta usando el Data Mart con Esquema Constelación se accede a un menor número de tablas para obtener los indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

Proposición 5. Con la metodología propuesta usando el Data Mart con Esquema Constelación se necesitan menos personas para obtener los indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

Proposición 6. A través de la metodología propuesta usando el Data Mart con Esquema Constelación de los proyectos ENOE del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 se disminuye el tiempo de respuesta para la obtención de los indicadores principales.

Capítulo II. Marco Teórico

2.1 Armonización de variables

La Subdirección de Diseño Muestral de Vivienda determina el tamaño de muestra para encuestas en hogares solicitadas por usuarios internos o externos al INEGI, para realizar este proceso se necesita contar con la información conceptual de la encuesta para lo cual realiza una investigación de los conceptos faltantes en encuestas que manejan los mismos indicadores o parecidos dentro de las publicaciones del INEGI, sin embargo cada encuesta cuenta con diversos criterios a cubrir y esa flexibilidad no se muestra en las publicaciones por lo cual se requiere realizar una exploración y explotación de la información de proyectos anteriores que actualmente se tienen almacenadas en diferentes áreas, con distintos criterios de almacenamiento y sin una organización homogénea. De tal manera que se invierte una gran cantidad de recursos humanos y materiales para lograr dicho objetivo pues para cada encuesta se necesita un programador capacitado para explotar la información de cierto proyecto y así obtener los indicadores principales que posteriormente pasarán a los muestristas para su análisis.

La finalidad de contar con información integrada de los proyectos a través del tiempo es tomar la mejor decisión basando el diseño muestral en la estimación más acorde con las referencias de espacio y tiempo de interés de la encuesta a realizar.

El proyecto IPUMS-International (Integrated Public Use Microdata Series, Series de Microdatos Integrados de Uso Público) se extendió en 1998 a los Censos de Colombia con la colaboración del Departamento Nacional de Estadística de Colombia (DANE); posteriormente dio paso a la integración de siete países: China, Colombia, Estados Unidos, Francia, Kenya, México y Vietnam. Se conjuntaron las muestras de microdatos en formato de cómputo del período 1960-2000. (McCaa & Esteve, 2003).

Los objetivos del proyecto IPUMS-International son preservar, integrar y difundir a través del uso de distintas técnicas, métodos y habilidades que a continuación se describen con mayor detalle:

Preservar. Su misión es recabar e inventariar los microdatos y documentos censales en el mundo que hayan sobrevivido hasta nuestros días.

Integrar. La integración es el principal reto al que IPUMS-International debe enfrentarse, pero también su principal fortaleza, el hecho diferencial que singulariza a esta base de datos respecto a las demás. Durante la etapa de integración, los datos son procesados con cuatro finalidades distintas: i) garantizar la confidencialidad de los datos; ii) reformar, limpiar e imputar valores perdidos en la base de datos; iii) armonizar variables; iv) construir variables.

Difundir. Una plataforma de difusión eficiente es esencial para optimizar el uso de los microdatos integrados. IPUMS-International hace un uso extensivo e intensivo de las nuevas tecnologías para satisfacer tan importante objetivo. La difusión se realiza por internet, mediante un sistema que permite al usuario confeccionar su propia base de datos, escogiendo formatos, muestras, variables y casos específicos. (McCaa & Esteve, 2002)

Como conclusión de este proyecto se puede decir que el IPUMS-International es consciente de su potencial, por lo cual sigue trabajando activamente para poner a disposición de la comunidad científica series de microdatos integrados para el máximo número de países posible, en estrecha colaboración con los institutos de estadística nacionales, centros de investigación y profesionales de la Demografía. Hoy, esta ambición es una realidad para el ámbito de América Latina. En cinco años, IPUMS-América Latina prevé difundir datos de más de 70 censos de 17 países. Para ello, se replicará la estrategia de distribución de datos de IPUMS International, de cuyas características se ha informado en este trabajo, otorgando las máximas facilidades en el acceso a los datos a nuestros usuarios. (McCaa & Esteve, 2003).

2.2 Procedimiento para el Cálculo de Tamaños de Muestra de Encuestas Especiales en hogares

Cumpliendo con la misión del Instituto Nacional de Estadística y Geografía (INEGI), la Dirección General de Estadísticas Socio-demográficas solicita a la Dirección General Adjunta de Encuestas Socio-demográficas y Registros Administrativos de Diseño y Marcos Estadísticos, que a su vez delega a la Dirección de Diseño Muestral de Viviendas, a través de una atenta nota, el tamaño de muestra para determinada encuesta en hogares requerida por algún usuario, ya sea interno y/o externo al INEGI. Posteriormente, la Dirección de Diseño y Marcos Estadísticos, solicita el tamaño de muestra a la subdirección de Diseño Muestral de Viviendas, la cual realiza un análisis de los insumos que le fueron entregados por parte de la dependencia solicitante. El objeto es validar aquellos insumos que les fueron entregados, e identificar aquellos conceptos que hacen falta para poder realizar un diseño muestral de la encuesta, estos conceptos son los siguientes: objetivos de la encuesta, población objeto de estudio, cobertura, dominios de interés, indicadores, precisión y confianza.

Una vez que los conceptos del diseño de la encuesta están completos, el muestrista perteneciente a el Área de Diseño Muestral de Encuestas en Hogares y el solicitante de la encuesta en forma conjunta, establecen de manera precisa, las variables a estimar, así como los parámetros poblacionales más importantes a estimar, junto con sus errores de estimación (relativo y/o absoluto). Los estadísticos son generalmente, promedios, totales, tasas y proporciones. En base a los criterios establecidos en el diseño de la encuesta se determina la expresión matemática para determinar el tamaño de muestra para la encuesta.

Para poder determinar el tamaño de muestra que requiere una encuesta, se necesita conocer cierta información estadística como margen de error, estimaciones, efectos de diseño. Esta información varía dependiendo del tipo de encuesta.

La información faltante, se debe de investigar en encuestas que manejan los mismos indicadores o parecidos, dentro de las publicaciones del INEGI.

En caso de no encontrar dicha información directamente en publicaciones, ya sean dentro de la intranet o página web institucional, se solicitan aquellas bases de datos con información estadística que sirva para calcular de manera indirecta el(los) indicador(es) que se necesitan.

El proceso para la obtención de los indicadores principales es el siguiente:

1. Objetivo. Se define el objetivo con una descripción breve y concisa del propósito de la encuesta en términos de la información que se pretende obtener.
2. Revisión. El muestrista y el programador revisan el cuestionario perteneciente a la encuesta para definir las variables que forman parte de ella y así definir las variables principales.
3. Población objeto de estudio. Definición clara y precisa del conjunto de entes o individuos de los que se pretende obtener información.
4. Cobertura. Límites geográficos o sectoriales del universo de estudio.
5. Dominios de interés. Se refiere a los subconjuntos del universo de estudio para los que se pretende obtener las estimaciones.
6. Variables principales. Aquellas variables que son relevantes para la investigación y por tanto deben tomarse en como punto de referencia para fijar la precisión de las estimaciones de la encuesta.
7. Programación. El programador revisa las bases de datos necesarias para la obtención de las variables principales. Revisa la cobertura, el dominio de interés y la población objeto de estudio. Realiza la programación necesaria para obtener las variables principales y pasa la información al muestrista
8. Procesamiento. Es el tiempo que tarda la computadora en realizar los cálculos hechos anteriormente por el programador.
9. Obtener indicadores principales.

2.3 Data Warehouse

El data Warehouse de una organización se mantiene separado de las bases de datos operacionales de la misma, esto se debe a diversas razones, el data warehouse soporta el procesamiento analítico en línea (OLAP) a diferencia del procesamiento de transacciones en línea (OLTP) que es soportado por bases de datos operacionales.

El data warehouse está dirigido a la toma de decisiones, los datos históricos, totalizados y consolidados son más importantes que los registros individuales y detallados. Los datos en el data warehouse son típicamente modelados de forma multidimensional pues contienen diversas dimensiones de interés. *(Chaudhuri y Dayal, 1997)*

Un data warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia. (Inmon, 1996)

2.3.1 Características principales

Orientada al negocio. El diseño y la implementación del data warehouse se realiza para cubrir los aspectos de interés de la organización, es decir, se almacena la información concerniente al proceso de toma de decisiones. Su nivel de detalle excluye datos del día a día que son tan importantes para sistemas operacionales además de la interacción de la información o relaciones entre tablas son bastantes pues el data warehouse requiere que la información sea redundante y dimensionada para evitar recorrer toda la base de datos cuando se realice algún análisis determinado y así disminuir el tiempo de respuesta.

Integrada. El data warehouse está construido a partir de fuentes provenientes de diversos departamentos, secciones, áreas y aplicaciones por lo que se deben consolidar antes de ser agregados al DW. A este proceso se le conoce como Extracción, Transformación y Carga de Datos (ETL – Extraction, Transformation and Load). Como resultado de la integración se obtendrá un modelo globalmente aceptado y usado por un usuario final con la certeza de la solidez de los datos.

Variante en el tiempo. Toda la información del DW posee su propio sello de tiempo. Los datos son almacenados junto con sus respectivos históricos por lo cual se pueden realizar consultas, representando cada una la misma información en diferentes períodos de tiempo. El intervalo de tiempo y periodicidad de los datos debe definirse de acuerdo a la necesidad y requisitos de los usuarios.

No volátil .Los datos cuando entran al data warehouse no cambian, solo existen dos tipos de operaciones: carga y acceso de datos. (Bernabeu, 2007)

2.3.2 Estructura

El Data Warehouse estructura el flujo de los datos en diversos niveles (Figura 2) y se describen de la siguiente manera:

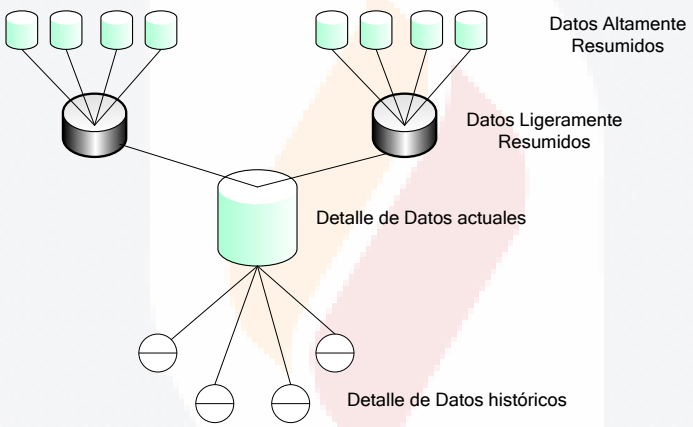


Figura 2; Flujo de datos de un Data Warehouse (Bernabeu, 2007)

- Detalle de datos actuales. Reflejan las ocurrencias más recientes. Posee el más bajo nivel de granularidad, se almacenan los datos a nivel detalle.
- Detalle de datos históricos. Representan aquellos datos antiguos, que no son frecuentemente consultados. También se almacenan a nivel detalle.
- Datos ligeramente resumidos. Provienen de un bajo nivel de detalle y suman los datos bajo algún criterio o condición de análisis.
- Datos altamente resumidos. Son aquellos que compactan aún más los datos ligeramente resumidos.
- Metadatos. Representan la información acerca de los datos. Se sitúa en una dimensión diferente al de los datos del DW. (Bernabeu, 2007)

2.3.3 Arquitectura

La arquitectura del Data Warehouse (Figura 3) está compuesta por:

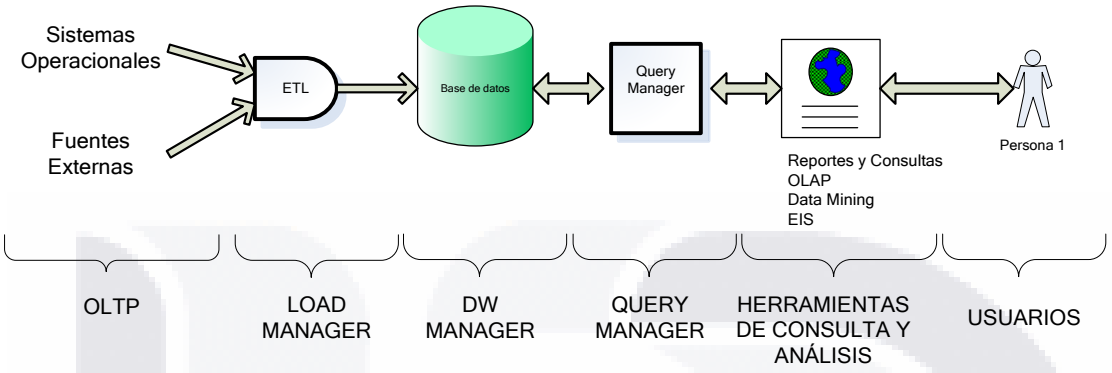


Figura 3; Arquitectura de un Data Warehouse. (Bernabeu, 2007)

1. OLTP (On Line Transaction Processing).

Los datos son extraídos desde aplicaciones, bases de datos, archivos, etc. Esta información generalmente reside en diferentes tipos de sistemas, orígenes y arquitecturas y tienen formatos muy variados.

2. Load manager

Se requiere un sistema que se encargue de la extracción, transformación y carga de los datos (ETL).

- La extracción se encarga de obtener los datos de interés provenientes del OLTP a través de rutinas programadas y consultas en SQL por ejemplo.

- La transformación es la encargada de convertir aquellos datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el DW.

- La carga es el proceso responsable de cargar la estructura de datos del DW con los datos que han sido transformados y que residen en el almacenamiento intermedio y los datos de los OLTP que tienen correspondencia directa con el depósito de datos.

3. Data Warehouse manager

El DW Manager realiza diversas funciones, entre las más importantes están:

1. Transforma e integra los datos fuentes y del almacenamiento intermedio en un modelo adecuado para la toma de decisiones.

2. Gestiona el depósito de datos a través de tablas de hechos y tablas de dimensiones, y lo organiza en torno a una base de datos multidimensional. Esto permite que se puedan crear cubos multidimensionales, Business Models u otras estructuras de datos.

   Bases de datos multidimensionales

   Las bases de datos multidimensionales, proveen una estructura que permite, a través de la creación y consulta a una estructura de datos determinada (Cubo Multidimensional) tener acceso flexible a los datos, para explorar y analizar sus relaciones, y consiguientes resultados.

   Las bases de datos multidimensionales implican tres variantes posibles de modelamiento, que permiten realizar consultas de soporte de decisión:

   - Esquema en estrella (Star Scheme).
   - Esquema copo de nieve (Snowflake Scheme).
   - Esquema constelación o copo de estrellas (Starflake Scheme).

   Los mencionados esquemas pueden ser implementados de diversas maneras, que, independientemente al tipo de arquitectura, requieren que toda la estructura de datos este desnormalizada o semidesnormalizada, para evitar desarrollar uniones complejas para acceder a la información, con el fin de agilizar la ejecución de consultas.

   Los diferentes tipos de implementación son los siguientes:

   - Relacional – ROLAP.

- Multidimensional – MOLAP.
- Híbrido – HOLAP

Tablas de dimensiones.

Las tablas de dimensiones definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio. En un modelo multidimensional bien diseñado, las tablas de dimensiones tienen muchas columnas o atributos. Estos atributos describen las filas de la tabla de dimensiones. Cada dimensión se define por su clave principal. Las tablas de dimensiones son los puntos de entrada en la tabla de hechos.

Cada tabla de dimensión podrá contener los siguientes campos:

- Clave principal o identificador único.
- Clave foráneas.
- Datos de referencia primarios: datos que identifican la dimensión. Por ejemplo: nombre del cliente.
- Datos de referencia secundarios: datos que complementan la descripción de la dimensión. Por ejemplo: e-mail del cliente, fax del cliente, etc.

En un DW, la creación y el mantenimiento de una tabla de dimensión Tiempo es obligatoria, y la definición de granularidad y estructuración de la misma depende de la dinámica del negocio que se esté analizando. Toda la información dentro del depósito, como ya se ha explicado, posee su propio sello de tiempo que determina la ocurrencia de un hecho específico, representando de esta manera diferentes versiones de una misma situación.

Tablas de hechos.

Una tabla de hechos es la tabla principal en un modelo numérico multidimensional. Las tablas de hechos contienen, precisamente, los hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones. Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones. El registro del hecho posee una clave primaria que está compuesta por las claves primarias de las tablas de dimensiones relacionadas a este.

Cubo Multidimensional.

Un cubo multidimensional o hipercubo representa o convierte datos planos que se encuentran en filas y columnas, en una matriz de N dimensiones. Los cubos contienen:

- Indicadores. Sumarizaciones que se efectúan sobre algún hecho, perteneciente a una tabla de hechos.
- Atributos. Campos o criterios de análisis, pertenecientes a tablas de dimensiones.
- Jerarquías. Representa una relación lógica entre dos o más atributos.
- Relación. Una relación representa la forma en que dos atributos interactúan dentro de una jerarquía.
- Granularidad. Representa el nivel de detalle al que se desea almacenar la información sobre el negocio. Los datos que posean granularidad fina (nivel de detalle) podrán ser resumidos hasta obtener una granularidad media o gruesa. No sucede lo mismo en sentido contrario, ya que por ejemplo, los datos almacenados con granularidad media podrán resumirse, pero no tendrán la facultad de ser analizados a nivel de detalle.

Tipos de Modelamiento de un DW

Esquema en Estrella (Star Scheme).

El esquema en estrella, consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves.

Este modelo debe estar totalmente desnormalizado, es decir que no puede presentarse en tercera forma normal (3ra FN). Las ventajas que trae aparejada la desnormalización, son las de obviar uniones (Join) entre las tablas cuando se realizan consultas, procurando así un mejor tiempo de respuesta y una mayor sencillez con respecto a su utilización. Algunas características de este modelo son:

- Posee los mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.
- Existe paralelismo entre su diseño y la forma en que los usuarios visualizan y manipular los datos.
- Simplifica el análisis.

- Facilita la interacción con herramientas de consulta y análisis.

Esquema Constelación. (Starflake Scheme).

Este modelo está compuesto por una serie de esquemas en estrella, está formado por una tabla de hechos principal y por una o más tablas de hechos auxiliares, las cuales pueden ser sumarizaciones de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones.

No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, ya que, las tablas de hechos auxiliares pueden vincularse con solo algunas de las tablas de dimensiones asignadas a la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones.

Su diseño y cualidades son muy similares a las del esquema en estrella, pero posee una serie de diferencias con el mismo, que son precisamente las que lo destacan y caracterizan. Entre ellas se pueden mencionar:

- Permite tener más de una tabla de hechos, por lo cual se podrán analizar más aspectos claves del negocio con un mínimo de esfuerzo adicional de diseño.
- Contribuye a la reutilización de las tablas de dimensiones, ya que una misma tabla de dimensión puede utilizarse para varias tablas de hechos.
- No es soportado por todas las herramientas de consulta y análisis.

4. Query manager

Este componente realiza las operaciones necesarias para soportar los procesos de gestión y ejecución de consultas relacionales, tales como Join y agregaciones, y de consultas propias del análisis de datos, como drill-up y drill-down.

Query Manager recibe las consultas del usuario, las aplica a la estructura de datos correspondiente (cubo multidimensional, Business Models, etc.) y devuelve los resultados obtenidos.

Cabe aclarar que una consulta a un DW, generalmente consiste en la obtención de indicadores a partir de los datos (hechos) de una tabla de hechos, restringidas por las propiedades o condiciones de los atributos que hayan sido creados.

Las operaciones que se pueden realizar sobre modelos multidimensionales y que son las que verdaderamente les permitirán a los usuarios explorar e investigar los datos en busca de respuestas, son:

- Drill-down
- Drill-up
- Drill-across
- Roll-across
- Pivot
- Page.

5. Herramientas de consulta y análisis

Las herramientas de consulta y análisis son sistemas que permiten al usuario realizar la exploración de datos del DW. A través de una amigable interfaz gráfica y una serie de simples pasos, el usuario genera consultas que son enviadas desde la herramienta de consulta y análisis al Query Manager, este a su vez realiza la extracción de información al DW Manager y devuelve los resultados obtenidos a la herramienta que se los solicitó. Luego, estos resultados son expuestos ante el usuario en formatos que le son familiares.

Existen diferentes tipos de herramientas de consulta y análisis, y de acuerdo a la necesidad, tipos de usuarios y requerimientos del negocio, se deberán seleccionar las más propicias al caso. Entre ellas se destacan: Reportes y Consultas, OLAP, Dashboards. Data Mining, EIS

6. Usuarios

Los usuarios que posee el DW son aquellos que se encargan de tomar decisiones y de planificar las actividades del negocio, es por ello que se hace tanto énfasis en la integración, limpieza de datos, etc., para poder conseguir que la información posea toda la calidad posible. Es a través de las herramientas de consulta y análisis, que los usuarios exploran los datos en busca de respuestas para poder tomar decisiones proactivas.

2.3.4 Componentes de un Data Warehouse

Para mostrar el ambiente del Data Warehouse, se describen cuatro elementos básicos (Figura 4).

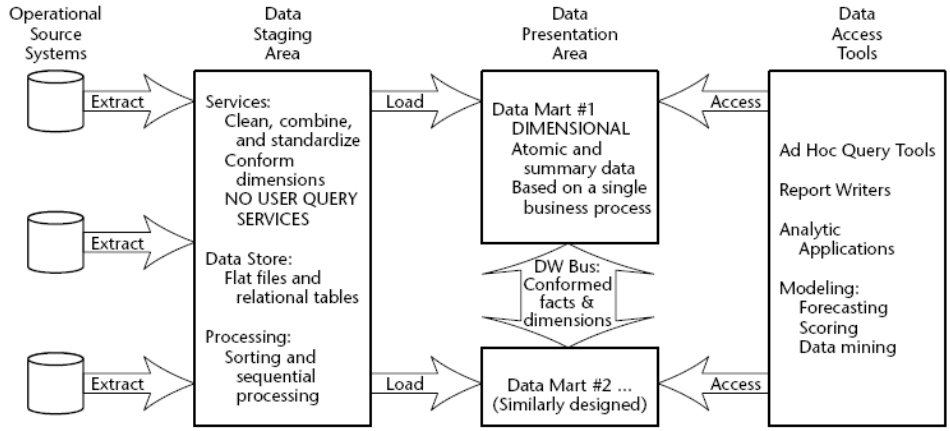


Figura 4; Elementos básicos de un Data Warehouse (Kimball et al, 1998)

Operational Source Systems

Los sistemas operacionales capturan las transacciones del negocio día a día, contienen pocos datos históricos y tienen como función principal el rendimiento del procesamiento y la disponibilidad de los datos.

Data Staging Area

Es a la vez un área de almacén y un conjunto de procesos referentes a extraer, transformar y cargar. La extracción de datos se refiere a leer y entender la información fuente y tomar los datos necesarios para el data warehouse para su posterior manipulación. La transformación es el siguiente paso a seguir, se realiza la depuración de los datos como corrección de faltas ortográficas, trata de elementos faltantes y análisis de formatos estándar, corrección de duplicados, entre otros. Estas transformaciones son todos los precursores de la carga de los datos en el área de presentación del Data Warehouse.

Data Presentation Area

En el área de presentación, los datos se encuentran organizados, almacenados y disponibles para consultas por parte de los usuarios, reportes escritos y otras aplicaciones analíticas. El área de presentación se refiere comúnmente a un serie de data marts integrados, los cuales representan datos de procesos de negocios simples.

Algunas características del área de presentación son:

- Los datos están presentes, almacenados y se acceden en esquemas tridimensionales.

- La industria ha concluido que la modelación tridimensional es la técnica más viable para entregar datos a los usuarios de data warehouse.

- Los data marts deben contener datos atómicos y detallados pese a que también pueden contener resúmenes de datos, agregaciones y sumarizaciones pues el usuario requiere información tanto a ciertos niveles de granularidad como de totales.

- Se necesitan los datos más finamente integrados en el área de presentación para que los usuarios puedan hacer las preguntas más precisas posibles.

- Todos los data marts deberán estar construidos usando dimensiones comunes y tablas de hechos, esto se refiere a conformación. Esto es la base de la arquitectura de bus del data warehouse.

Si el área de presentación se basa en una base de datos relacional, entonces, estas dimensiones del modelo tablas se conocen como esquemas de estrella. Si el área de presentación se basa en la base de datos multidimensional o procesamiento analítico en línea (OLAP), los datos se almacenan en cubos. Si bien la tecnología no era originalmente conocido como OLAP, muchos de los primeros proveedores de sistemas de soporte de decisiones de sus sistemas construidos en torno al concepto del cubo, lo que los vendedores de hoy OLAP natural están alineados con el enfoque tridimensional para almacenamiento de datos. Modelado dimensional es aplicable tanto a las bases de datos relacionales y multidimensionales. Ambos tienen un diseño lógico común con dimensiones reconocibles, sin embargo, la ejecución física es diferente. Si bien las capacidades de la tecnología OLAP están mejorando continuamente, la mayoría de los grandes puestos de datos siguen siendo aplicados sobre bases de datos relacionales. Además, la mayoría de los cubos OLAP se obtienen de perforación o en forma de estrella dimensiones relacionales esquemas usando una variación de la navegación global.

Data Access Tools

Una herramienta de acceso a datos puede ser tan simple como una herramienta de consulta ad hoc o tan complejo como una sofisticada aplicación de minería de datos o modelado.

Metadatos

Otro concepto importante dentro del tema de Data warehouse son los metadatos pues se refiere a toda la información en el entorno del data warehouse que no son los datos en sí. Es como una enciclopedia para el data warehouse que viene en una variedad de formatos que apoyan las diversas necesidades técnicas, administrativas y de grupos de usuarios de negocios. (Kimball et al, 1998)

2.4 Revisión de la Metodología HEFESTO

HEFESTO es una metodología propuesta para la construcción de un Data Warehouse y "está fundamentada en una muy amplia investigación, comparación de metodologías existentes y experiencias propias en procesos de confección de almacenes de datos" (Bernabeu, 2007)

La construcción e implementación de un DW puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, con la salvedad de que para algunas fases en particular, las acciones que se han de realizar serán muy diferentes. (Bernabeu, 2007)

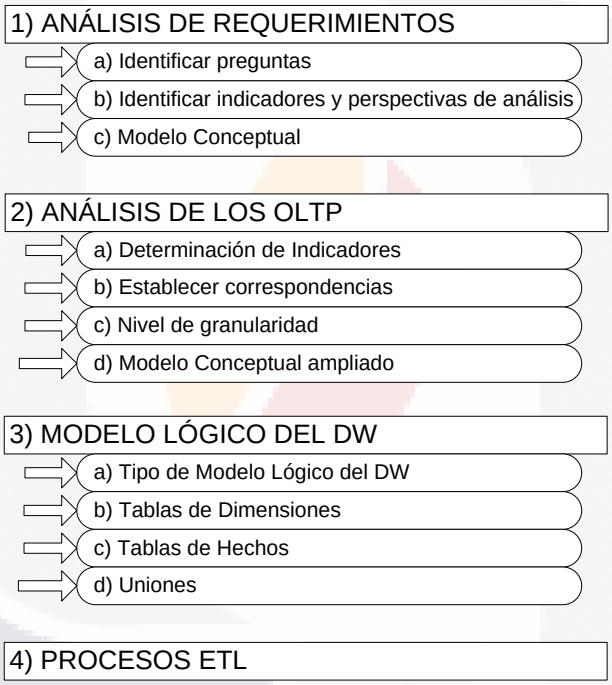


Figura 5; Metodología HEFESTO; (Bernabeu, 2007)

La metodología HEFESTO (Figura 5) cuenta con las siguientes características:

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.

- Se basa en los requerimientos del usuario, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- Reduce la resistencia al cambio, ya que involucra al usuario final en cada etapa para que tome decisiones respecto al comportamiento y funciones del DW.

- Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.

- Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.

- Es independiente de las herramientas que se utilicen para su implementación.

- Es independiente de las estructuras físicas que contengan el DW y de su respectiva distribución.

- Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente.

- Se aplica tanto para Data Warehouse como para Data Mart.

2.4.1 Análisis de Requerimientos

Se identifican los requerimientos del usuario a través de preguntas que expliquen los objetivos de la organización, posteriormente se identifican los indicadores y perspectivas que se tomarán en cuenta para la creación del Data Warehouse.

Un punto importante a tomar en cuenta, es que la información debe estar soportada de alguna manera por un OLTP, ya que de otra manera no se podrá elaborar el DW. Como resultado de esta fase se obtendrá un modelo conceptual.

a) Identificar preguntas

La idea central es que se formulen preguntas complejas sobre el negocio, que incluyen variables de análisis que se consideren relevantes, ya que estas son las que permitirán estudiar la información desde diversas perspectivas. Un punto importante a

tomar en cuenta, es que la información debe estar soportada de alguna manera por un OLTP, ya que de otra manera no se podrá elaborar el DW.

b) Identificar indicadores y perspectivas de análisis.

Las preguntas clave deberán descomponerse para así descubrir los indicadores que se utilizarán y las perspectivas de análisis que intervendrán.

Se debe tomar en cuenta que los indicadores son, en general, valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, etc.

Las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, etc.

c) Modelo conceptual (Figura 6).

Se construye un modelo conceptual a través de los indicadores y perspectivas obtenidas. A través de este modelo, se podrá observar con claridad cuáles son los alcances del proyecto, para luego poder trabajar sobre ellos, además de poseer un alto nivel de definición de los datos, permite que pueda ser presentado ante los usuarios y explicado con facilidad.

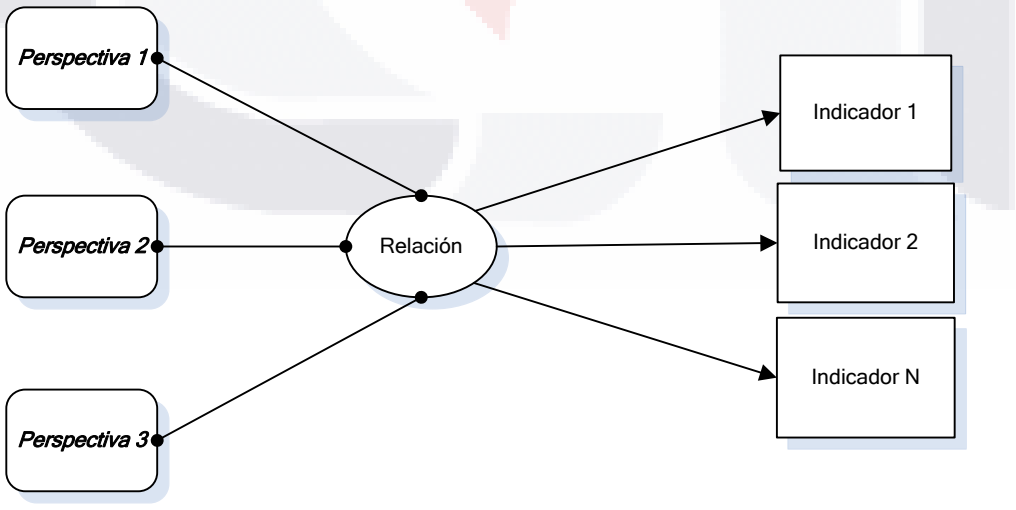


Figura 6; Modelo Conceptual; (Bernabeu, 2007)

2.4.2 Análisis de los OLTP

Seguidamente, se analizarán las fuentes OLTP para determinar cómo serán calculados los indicadores y para establecer las respectivas correspondencias entre el modelo conceptual y las fuentes de datos. Luego, se definirán qué campos se incluirán en cada perspectiva. Finalmente, se ampliará el modelo conceptual con la información obtenida en este paso.

a) Determinación de indicadores.

Se deberán explicitar como se calcularán los indicadores, definiendo los siguientes conceptos para cada uno de ellos:

Hecho/s que lo componen, con su respectiva fórmula de cálculo.

Por ejemplo: Hecho1 + Hecho2.

Función de sumarización que se utilizará para su agregación.

Por ejemplo: SUM, AVG, COUNT, etc.

b) Establecer correspondencias.

El objetivo de este paso, es el de examinar los OLTP disponibles que contengan la información requerida, como así también sus características, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos. La idea es, que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

c) Nivel de granularidad

Una vez que se han establecido las relaciones con los OLTP, se examinarán y seleccionarán los campos que contendrá cada perspectiva, ya que será a través de estos por los que se manipularán y filtrarán los indicadores. Para ello, basándose en las correspondencias establecidas en el paso anterior, se debe presentar al usuario los datos de análisis disponibles para cada perspectiva.

Es muy importante conocer en detalle que significa cada campo y/o valor de los datos encontrados en los OLTP, por lo cual, es conveniente investigar su sentido, ya sea a través de diccionarios de datos, reuniones con los encargados del sistema, análisis de los datos propiamente dichos, etc.

Luego de exponer frente al usuario los datos existentes, explicando su significado, valores posibles y características, este debe decidir cuáles son los que considera relevantes para consultar los indicadores y cuáles no.

Con respecto a la perspectiva “Tiempo”, es muy importante definir el ámbito mediante el cual se agruparán o sumarizarán los datos. Sus campos posibles pueden ser: día de la semana, quincena, mes, trimestres, semestre, año, etc.

Al momento de seleccionar los campos que integrarán cada perspectiva, debe prestarse mucha atención, ya que esta acción determinará la granularidad de la información encontrada en el DW.

d) Modelo Conceptual Ampliado (Figura 7)

Con el fin de graficar los resultados obtenidos en los pasos anteriores, se ampliará el modelo conceptual, colocando bajo cada perspectiva los campos elegidos y bajo cada indicador su respectiva fórmula de cálculo.

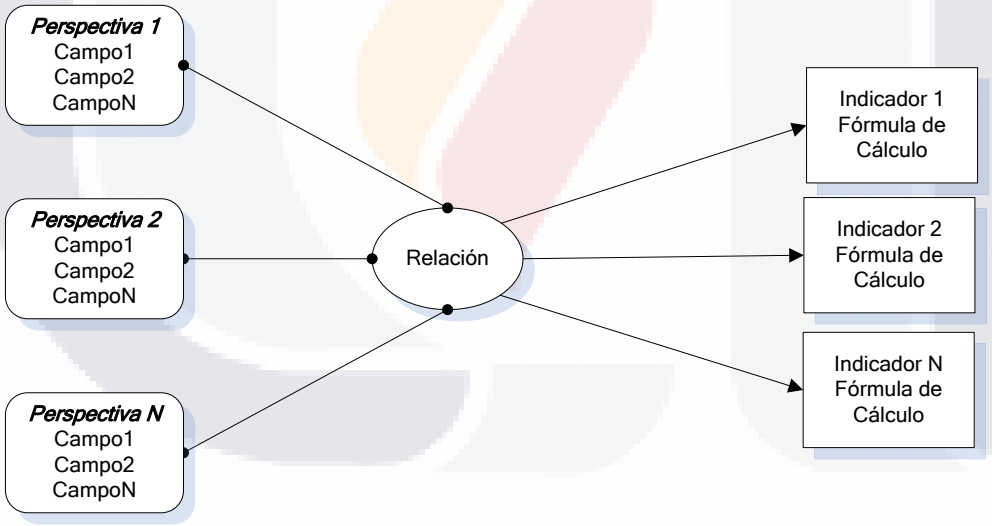


Figura 7; Modelo Conceptual Ampliado; (Bernabeu, 2007)

2.4.3 Modelo Lógico del DW

Basándose en el modelo conceptual creado se hará el modelo lógico del DW. Se debe definir el tipo de modelo que se utilizará y posteriormente se diseñaran las tablas de dimensiones y hechos. Finalmente, se realizarán las uniones pertinentes entre estas tablas.

a) **Tipo de modelo lógico.** Se debe seleccionar cuál será el tipo de esquema que se utilizará para contener la estructura del depósito de datos, que se adapte mejor a los requerimientos y necesidades del usuario. Es muy importante definir objetivamente si se empleará un esquema en estrella, constelación o copo de nieve, ya que esta decisión afectará considerablemente la elaboración del modelo lógico.

b) **Tablas de dimensiones**. Cada perspectiva definida en el modelo conceptual constituirá una tabla de dimensión. Para ello deberá tomarse cada perspectiva con sus campos relacionados y realizarse el siguiente proceso:
   - Se elegirá un nombre que identifique la tabla de dimensión.
   - Se añadirá un campo que represente su clave principal.
   - Se redefinirán los nombres de los campos si es que no son lo suficientemente intuitivos.

c) **Tablas de hechos.** Se definirán las tablas de hechos, que son las que contendrán los hechos a través de los cuales se construirán los indicadores de estudio.

   Para los esquemas en estrella y copo de nieve, se realizará lo siguiente:
   - Se le deberá asignar un nombre a la tabla de hechos que represente la información analizada, área de investigación, negocio enfocado, etc.
   - Se definirá su clave primaria, que se compone de la combinación de las claves primarias de cada tabla de dimensión relacionada.
   - Se crearán tantos campos de hechos como indicadores se hayan definido en el modelo conceptual y se les asignará los mismos nombres que estos. En caso que se prefiera, podrán ser nombrados de cualquier otro modo.

Para los esquemas constelación se realizará lo siguiente:

- Las tablas de hechos se deben confeccionar teniendo en cuenta el análisis de las preguntas realizadas por el usuario en pasos anteriores y sus respectivos indicadores y perspectivas.
- Cada tabla de hechos debe poseer un nombre que la identifique, contener sus hechos correspondientes y su clave debe estar formada por la combinación de las claves de las tablas de dimensiones relacionadas.

d) **Uniones**. Se realizarán las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos.

2.4.4 Procesos ETL

Una vez construido el modelo lógico, se deberá proceder a probarlo con datos, a través de procesos ETL. Para realizar la compleja actividad de extraer datos de diferentes fuentes, para luego integrarlos, filtrarlos y depurarlos, existen varios software que facilitan estas tareas, por lo cual este paso se centrará solo en la generación de las sentencias SQL que contendrán los datos que serán de interés. Antes de realizar la carga de datos, es conveniente efectuar una limpieza de los mismos, para evitar valores faltantes y anómalos.

Al generar los ETL, se debe tener en cuenta cual es la información que se desea almacenar en el depósito de datos, para ello se pueden establecer condiciones adicionales y restricciones. Estas condiciones deben ser analizadas y llevadas a cabo con mucha prudencia para evitar pérdidas de datos importantes.

Cuando se trabaja con un esquema constelación, hay que tener presente que varias tablas de dimensiones serán compartidas con diferentes tablas de hechos, ya que puede darse el caso de que algunas restricciones aplicadas sobre una tabla de dimensión en particular para analizar una tabla de hechos, se puedan contraponer con otras restricciones o condiciones de análisis de otras tablas de hechos. Primero se cargarán los datos de las dimensiones y luego los de las tablas de hechos, teniendo en cuenta siempre, la correcta correspondencia entre cada elemento. En el caso en que se esté utilizando un esquema copo de nieve, cada vez que existan jerarquías de dimensiones, se comenzarán cargando las tablas de dimensiones del nivel más general al más detallado.

Cuando se haya cargado en su totalidad el DW, se deben establecer sus políticas de actualización o refresco de datos. (Bernabeu, 2007)

2.5 Casos Similares

2.5.1 Sistema de Variables INE Portugal (2006)

El INE Portugal se enfrentó al reto de construir un Sistema de Variables para apoyar la producción estadística y difundirla. Esto fue realizado en tres etapas:

La primera etapa se inició en el 2004 con la definición de los requerimientos del sistema, obteniendo como resultado un documento de descripción de los requerimientos y el análisis conceptual.

En la segunda etapa se detalló el análisis y se diseño la base de datos, con lo cual se obtuvo un prototipo de las aplicaciones a ser desarrolladas.

Para terminar, en la tercer etapa se llevó a cabo el desarrollo de la aplicación, terminando así en junio del 2006 dando como resultado un portal estadístico donde se concentra el Sistema de variables armonizado.

Los objetivos principales del Sistema de variables eran los siguientes:

- Crear un repositorio de variables para ayudar a:
  - El diseño de encuestas (operaciones estadísticas)
  - La difusión de datos estadísticos
  - Mejorar la coordinación estadística
  - Ayudar al diseño de cuestionarios
- En el futuro, ayudar a la generación automática de cuestionarios
- Poner a disposición los diferentes usos de una determinada variable en diversas encuestas
- Ayudar en la definición de las variables estandarizadas

Dicho trabajo fue basado en la ISO 11179 (en lo que se refiere a elementos de datos, sus componentes y sus relaciones), además hicieron modificaciones sobre la ISO 11179 para maximizar su utilidad específicamente para su Instituto.

Sin embargo, se enfrentaron a diversos problemas siendo uno de ellos la nomenclatura de variables, la ISO 11179 no contiene reglas específicas para conformar las nomenclaturas por lo cual se crearon reglas concernientes a esto.

Algunos beneficios obtenidos con el Sistema de Variables fueron:

1. Claridad. Los usuarios no solo visualizan datos, también visualizan definiciones, fórmulas, unidades de medida, población estudiada, entre otros.

2. Armonización. Uno de los principales problemas a resolver era la diversidad, pues las variables eran construidas en diferentes partes de la organización y llamadas de diferentes maneras. Se identificaron algunos componentes para el proceso de armonización:

    a. Compilación de la documentación existente
    b. Determinación de las variables disponibles
    c. Diseñar el sistema armonizado de variables y la documentación asociada.
    d. Estandarización (forma específica de la armonización en un esquema definido de acuerdo con que los cambios se realizan)

    Encontraron encuestas, cuestionarios e información en tablas con conjuntos de variables que eran conceptualmente comparables como características básicas de población, actividad económica, educación, etc. (Morgado & Isfan, 2006)

2.5.2. IPUMS-International (Integrated Public Use Microdata Series-International)

El proyecto IPUMS-International comienza con su primera fase en el año 1999 en la Universidad de Minnesota en el cual se incluyen microdatos de ocho países con amplia distribución geográfica.

IPUMS estandariza los formatos, armoniza los códigos, construye variables y elabora la documentación necesaria para facilitar el análisis de la información a los usuarios. A través de la tecnología el proyecto IPUMS publica gratuitamente en internet muestras armonizadas de microdatos censales.

Se llevan a cabo dos etapas para armonizar los microdatos internacionales, la primera es estandarizar la variedad de formatos de los microdatos censales y corregir los posibles errores; la segunda etapa se refiere a armonizar las variables, determinar su disponibilidad y comparabilidad entre variables, recopilar la información existente y diseñar un esquema de códigos de variables y su correspondiente documentación.

Para iniciar con la estandarización, el proyecto IPUMS-International definió el formato a seguir que consiste en transformar cada muestra en un formato jerárquico compuesto por un registro del hogar seguido de registros individuales para cada persona del hogar.

Cualquier información a nivel geográfico o de la vivienda se repitió en cada registro de hogares respectivos.

Una de las metas más importantes del proyecto IPUMS-International era crear variables consistentes en el tiempo y el espacio de tal manera que organizaron el proceso de armonización de variables en tres componentes (Figura 8):

a) Recopilar información existente. Para llevar a cabo esta tarea se buscaron fuentes de información de las variables las cuales podían ser formatos e instrucciones de los encuestadores, manuales de microdatos, etc. Sin embargo no se encontró toda la información necesaria, principalmente en los censos más antiguos. IPUMS-International pidió un ensayo temático a los especialistas de cada país con el fin de

proporcionar una visión sobre los problemas y la perspectiva de las personas con experiencia en datos censales.

b) Determinar la disponibilidad de las variables. No es una tarea fácil determinar la disponibilidad de cientos de variables construidas en su idioma original, sin embargo varios censos en el mundo comparten un conjunto de preguntas conceptualmente comparables, por ejemplo, características de población, educación, economía, etc.

c) Diseñar un esquema de codificación armonizado y documentación asociada. El objetivo fue crear un conjunto comparable de códigos para cada variable en la cual su significado fuera el mismo a pesar del tiempo y el espacio que permitan la comparación entre países y en perspectiva histórica. Para tener un buen resultado se estudió cada variable a través de la información recopilada en cuestionarios, formatos de encuestadores y como recurso final se acudió a la experiencia de las personas expertas en interpretación de códigos de cada país.

La armonización de variables ocupa un lugar central en el proceso de integración. Ante cualquier variable, su codificación final debe satisfacer dos requisitos: garantizar la máxima comparabilidad en el tiempo y en el espacio y, a la vez, retener todo el detalle contenido en las variables originales.

Se obtuvo un resultado satisfactorio del proyecto IPUMS-International, sin embargo en la actualidad se sigue trabajando para incorporar más países a la vasta y útil información que se tiene ya condensada y armonizada en este proyecto.

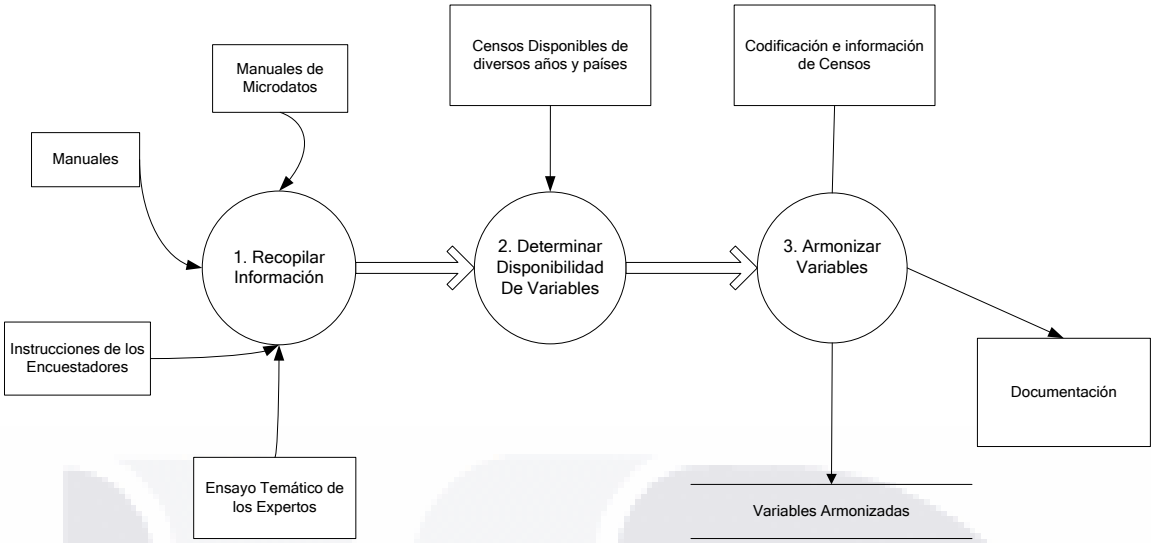


Figura 8; Proceso de Armonización de variables (Proyecto IPUMS)

Capítulo III. Metodología para el Desarrollo del Caso de Estudio

El proceso para facilitar la obtención de indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 fue diseñado en tres etapas (Figura 9); en las cuales se obtuvo como resultado un producto útil para la etapa siguiente, obteniendo como resultado un reporte de los indicadores principales.

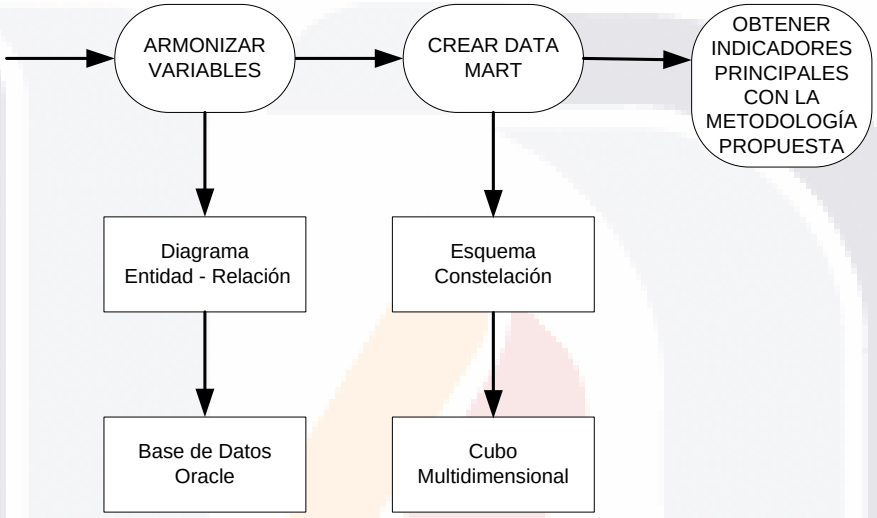


Figura 9; Proceso para la obtención de indicadores principales y comparables

3.1 Armonizar Variables

Dentro del proceso de armonizar variables (Figura 10) intervienen cuatro subprocesos:

- Definir proyectos.
- Definir variables principales.
- Armonizar variables principales.
- Validar armonización de variables principales.

En dichos subprocesos participan tanto informáticos como muestristas para validar la correcta armonización en el tiempo, espacio y dominios de estudio.

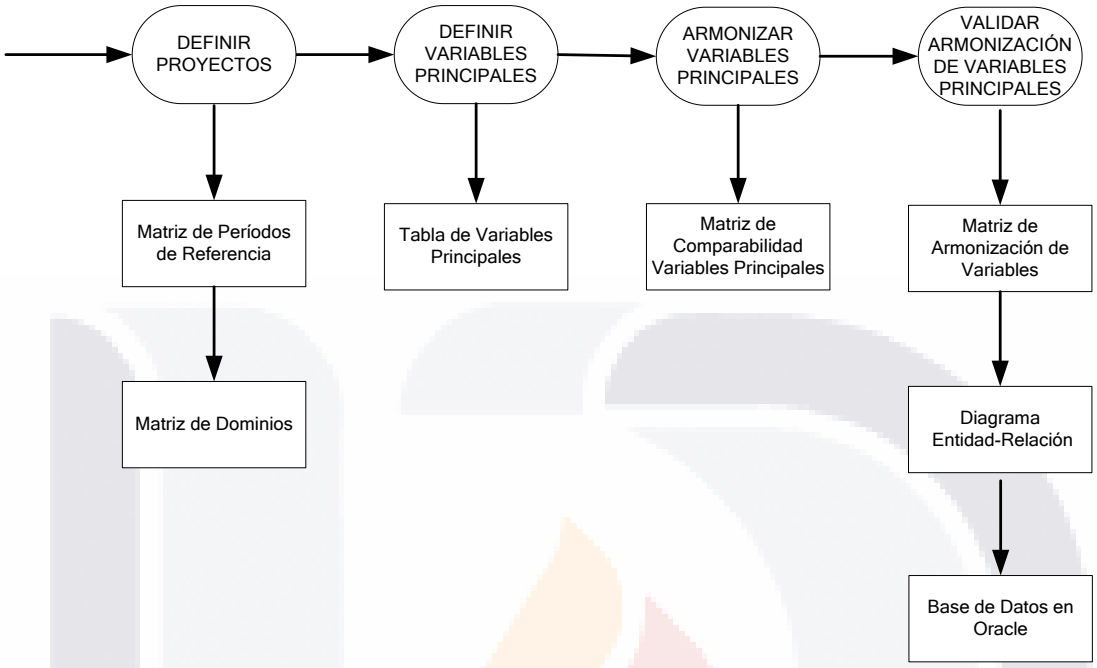


Figura 10; Proceso para armonizar variables

3.1.1 Definir proyectos

Se definieron los proyectos que formarian parte de este estudio basándose en la importancia del proyecto así como la existente comparabilidad tanto de variables como de tiempo, espacio y dominios de estudio.

Se determinaron los siguientes proyectos a trabajar:

- Censo de Población y Vivienda 2000
- Conteo de Población y Vivienda 2005
- Encuesta Nacional de Ocupación y Empleo 2005 a 2009

Matriz de Períodos de Referencia.

Basándose en estos proyectos se realizó la Matriz de Períodos de Referencia (Tabla 1), la cual tiene la finalidad de mostrar los proyectos que pueden ser comparables en el tiempo, se determinó como la mejor opción cortes mensuales del año 2000, 2005 a 2009.

Los valores para dicha matriz son:

‘1’ para el proyecto que se levantó en el mes y año referido.

‘0’ para el proyecto que no se levantó en el mes y año referido.

PER_REF		ENOE	CENSO	CONTEO
MES	AñO			
01	2000	0	0	0
02	2000	0	1	0
03	2000	0	0	0
04	2000	0	0	0
05	2000	0	0	0
06	2000	0	0	0
07	2000	0	0	0
08	2000	0	0	0
09	2000	0	0	0
10	2000	0	0	0
11	2000	0	0	0
12	2000	0	0	0
01	2005	0	0	0
02	2005	0	0	0
03	2005	1	0	0
04	2005	0	0	0
05	2005	0	0	0
06	2005	1	0	0
07	2005	0	0	0
08	2005	0	0	0
09	2005	1	0	1
10	2005	0	0	1
11	2005	0	0	1
12	2005	1	0	1
01	2006	0	0	0
02	2006	0	0	0
03	2006	1	0	0
04	2006	0	0	0
05	2006	0	0	0
06	2006	1	0	0
07	2006	0	0	0
08	2006	0	0	0
09	2006	1	0	0
10	2006	0	0	0
11	2006	0	0	0
12	2006	1	0	0

PER_REF		ENOE	CENSO	CONTEO
MES	AñO			
01	2007	0	0	0
02	2007	0	0	0
03	2007	1	0	0
04	2007	0	0	0
05	2007	0	0	0
06	2007	1	0	0
07	2007	0	0	0
08	2007	0	0	0
09	2007	1	0	0
10	2007	0	0	0
11	2007	0	0	0
12	2007	1	0	0
01	2008	0	0	0
02	2008	0	0	0
03	2008	1	0	0
04	2008	0	0	0
05	2008	0	0	0
06	2008	1	0	0
07	2008	0	0	0
08	2008	0	0	0
09	2008	1	0	0
10	2008	0	0	0
11	2008	0	0	0
12	2008	1	0	0
01	2009	0	0	0
02	2009	0	0	0
03	2009	1	0	0
04	2009	0	0	0
05	2009	0	0	0
06	2009	1	0	0
07	2009	0	0	0
08	2009	0	0	0
09	2009	1	0	0
10	2009	0	0	0
11	2009	0	0	0
12	2009	1	0	0

Tabla 1; Matriz de Períodos de Referencia

Matriz de Dominios.

El Censo de Población y Vivienda 2000, el Conteo de Población y Vivienda 2005 y la Encuesta Nacional de Ocupación y Empleo 2005 a 2009 cuentan con diversos dominios de estudio por lo cual se hizo una selección de ellos y se obtuvo la Matriz de Dominios (Tabla 2).

Dominios	
CDA	Área Metropolitana
ENT	Entidad
TLN	Tamaño de Loc.Nnal.
NNAL	Total Nacional
EST	Estrato

Tabla 2; Matriz de Dominios

3.1.2 Definir variables principales

Tabla de Variables Principales.

Se recabó la información de cada proyecto concerniente a este estudio, se estudian las variables disponibles y se determinan cuáles de ellas pasarán a la siguiente etapa como variables principales y se obtiene la Tabla de Variables Principales (Tabla 3)

No.	NOMBRE DE VARIABLE
1	Sexo
2	Edad
3	Estado conyugal o civil.
4	Condición del Derechohabiente al IMSS
5	Condición del Derechohabiente al ISSSTE
6	Condición del Derechohabiente a PEMEX
7	Condición del Derechohabiente tiene a Otra institución
8	No tiene Derechohabiencia
9	Discapacidad
10	Condición de discapacidad brazos
11	Condición de discapacidad al moverse
12	Condición de discapacidad auditiva
13	Condición de discapacidad lenguaje
14	Condición de discapacidad visual
15	Condición de discapacidad mental

No.	NOMBRE DE VARIABLE
17	Alfabetismo
18	Asistencia Escolar
19	Material de las paredes o muros
20	Material de los techos
21	Material de los pisos
22	Tiene cocina
23	Dispone de Electricidad
24	Dispone de Televisión
25	Dispone de Refrigerador
26	Dispone de Lavadora
27	Dispone de Computadora
28	Dispone de Sanitario
29	Dispone de Drenaje
30	Dispone de Agua Entubada
31	Dispone de Teléfono

16	Habla lengua indígena	32	Dispone de Automóvil

Tabla 3; Variables Principales

3.1.3 Armonizar variables principales

Se realizó un estudio exhaustivo de los metadatos existentes de los proyectos Censo de Población y Vivienda 2000, Conteo de Población y Vivienda 2005 y Encuesta Nacional de Ocupación y Empleo en el Data Warehouse institucional.

De cada proyecto mencionado anteriormente se obtuvo su caracterización general y se realizó una búsqueda para cada variable principal a estudiar. Después se estudió la codificación de cada variable y su significado; de tal manera que se fueron guardando dichos datos para su uso posterior.

Por otra parte, se extrajo la descripción de cada variable principal determinada en la etapa anterior a través de la página web de IPUMS-International en donde se encontró oportunamente la descripción de las variables similares a este estudio.

Matriz de Comparabilidad de Variables Principales.

Con los datos recabados del Data Warehouse institucional se procedió al llenado de la parte derecha de la Matriz de Comparabilidad de Variables Principales (Tabla 4) en la cual se observa el nombre dado a la variable principal y su descripción detallada.

De la misma manera, con los datos recabados a través de la página web del proyecto IPUMS-International se realizó el llenado de la parte izquierda de la Matriz de Comparabilidad de Variables Principales (Tabla 4), en donde se puede observar el código y descripción de la variable principal a estudiar. Cabe mencionar que dicho código usado en IPUMS-International será el nombre asignado a nuestras variables principales dentro de este proyecto de armonización.

IPUMS		Data Warehouse	
CÓDIGO	**DESCRIPCIÓN**	**NOMBRE**	**DESCRIPCIÓN**
SEX	Sex of respondent.	Sexo	Condición biológica que distingue a las personas en hombres y mujeres
AGE	AGE gives age in years as of the person's last birthday prior to or on the day of enumeration.	Edad	Número de años cumplidos por la persona, desde la fecha de su Nacimiento hasta el momento del hecho.
MARST1	Single/never married	Estado conyugal o civil.	Soltero/Nunca casado
MARST2	Married/in union		Casado/En unión libre
MARST3	Separated/divorced/spouse absent		Separado/Divorciado
MARST4	Widowed		Viudo
HLTHCOV10	IMSS only	Condición del Derechohabiente al IMSS	Identifica si la persona es derechohabiente al IMSS
HLTHCOV20	ISSSTE only	Condición del Derechohabiente al ISSSTE	Identifica si la persona es derechohabiente al ISSSTE
HLTHCOV30	Pemex, military, or naval coverage only	Condición del Derechohabiente a PEMEX	Identifica si la persona es derechohabiente a PEMEX, Defensa o Marina
HLTHCOV40	Other coverage only	Condición del Derechohabiente tiene a Otra institución	Identifica si la persona tiene derechohabiencia a otra institución diferente de IMSS, ISSSTE, PEMEX, Defensa o Marina, Seguro Popular o seguro por Institución Privada
HLTHCOV60	No coverage	No tiene derechohabiencia	
DISABLE	Indicates whether the person reported a disability of any kind.	Discapacidad	Identifica si la persona presenta o no alguna limitación física o mental
DISUPPR	Indicates whether the person reported a disability of any kind.	Condición de discapacidad brazos	Identifica si la persona presenta o no discapacidad para mover los brazos
DISMOBL	Indicates whether the respondent had any physical or mental health condition, lasting 6 months or more, that made it difficult or impossible to go outside the home alone. This did not include temporary health problems.	Condición de discapacidad al moverse	Disfunción en el aparato psicomotor que limita al individuo para realizar actividades físicas

Tabla 4 Parte 1; Matriz de Comparabilidad de Variables Principales

IPUMS		Data Warehouse	
CÓDIGO	**DESCRIPCIÓN**	**NOMBRE**	**DESCRIPCIÓN**
DISDEAF	Indicates whether the person was deaf or had limited hearing.	Condición de discapacidad auditiva	Identifica si la persona presenta o no discapacidad al oír
DISMUTE	Indicates if the person could not speak or had a significant speech impediment.	Condición de discapacidad lenguaje	Identifica si la persona presenta o no discapacidad para hablar
DISBLND	Indicates whether the person was blind or had limited vision.	Condición de discapacidad visual	Identifica si la persona presenta o no discapacidad al ver
DISMNTL	Indicates whether the person suffered a mental disability in the form of diminished capacity.	Condición de discapacidad mental	Identifica si la persona presenta o no discapacidad mental
SPKIND	Indicates whether a person speaks an indigenous language, and for many Latin American samples whether they also speak Spanish.	Habla Lengua Indígena	Especifica si el poblador habla o no alguna lengua indígena
LIT	Indicates whether or not the respondent could read and write in any language. A person is typically considered literate if he or she can both read and write. All other persons are illiterate; including those who can either read or write but cannot do both.	Alfabetismo	Situación que distingue a la población, según declare saber leer y escribir un recado.
SCHOOL	Indicates whether or not the person attended school at the time of the census or within some specified period of time prior to the census.	Asistencia Escolar	Situación que distingue a la población de 5 años y más, según su asistencia pasada o actual a cualquier establecimiento de enseñanza del Sistema Educativo Nacional como preescolar, primaria, secundaria, preparatoria, profesional o postgrado independientemente de su modalidad, ya sea pública o privada, escolarizada, abierta, de estudios técnicos o comerciales, educación especial o de educación para adultos

Tabla 4 Parte 2; Matriz de Comparabilidad de Variables Principales

IPUMS		Data Warehouse	
CÓDIGO	**DESCRIPCIÓN**	**NOMBRE**	**DESCRIPCIÓN**
WALL201	Waste, scrap, or discarded material	Material de las paredes o muros	Material de desecho
WALL204	Cardboard sheet		Lámina de cartón
WALL546	Metal or asbestos sheet		Lámina de asbesto o metálica
WALL405	Reed, bamboo, or palm		Carrizo, bambú o palma
WALL532	Clay or clay-covered sticks		Embarro o bajareque
WALL300	Wood		Madera
WALL523	Adobe		Adobe
WALL501	Brick, block, stone, or cement		Tabique, ladrillo, block, piedra, cantera, cemento o concreto
ROOF64	Discarded or scrap material	Material de los techos	Material de desecho
ROOF65	Cardboard		Lámina de cartón
ROOF34	Metal or asbestos		Lámina de asbesto o metálica
ROOF40	Wood and other plant materials		Palma, tejamanil o madera
ROOF12	Tile, unspecified		Teja
ROOF10	Masonry, concrete, clay tile, or tiles of unspecified type		Losa de concreto, tabique, ladrillo o terrado con viguería
FLOOR100	None (earth)	Material de los pisos	Tierra
FLOOR202	Cement		Cemento o firme
FLOOR231	Other finished, n.e.c.		Madera, mosaico u otros recubrimientos
KITCHEN20	Yes, have a kitchen	Tiene cocina	Disponen de cocina
KITCHEN10	No kitchen		No disponen de cocina
ELECTRC20	Indicates whether the household had access to electricity.	Dispone Electricidad	Energía eléctrica para alumbrar la vivienda, sin considerar la fuente de donde provenga, la cual puede ser un acumulador, el servicio público de energía, una planta particular, una planta de energía solar, entre otras
ELECTRC10		No dispone Electricidad	

Tabla 4 Parte 3; Matriz de Comparabilidad de Variables Principales

IPUMS		Data Warehouse	
CÓDIGO	**DESCRIPCIÓN**	**NOMBRE**	**DESCRIPCIÓN**
TV20	TV indicates whether the household had a television.	Dispone Televisión	Se refiere a la disponibilidad de televisión con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
TV10		No dispone Televisión	
REFRIG20	REFRIG indicates whether the household had a refrigerator.	Dispone Refrigerador	Se refiere a la disponibilidad de refrigerador con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
REFRIG10		No dispone Refrigerador	
WASHER20	WASHER indicates whether the household had a clothes washing machine.	Dispone Lavadora	Se refiere a la disponibilidad de lavadora con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
WASHER10		No dispone lavadora	
COMPUTR20	COMPUTR indicates whether the household had a personal computer.	Dispone Computadora	Se refiere a la disponibilidad de computadora con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
COMPUTR10		No dispone Computadora	
TOILET20	TOILET indicates whether the household had access to a toilet and, in most cases, whether it was a flush toilet or other type of installation.	Dispone Sanitario	Especifica si la vivienda cuenta o no con servicio sanitario.
TOILET10		No Dispone Sanitario	
SEWAGE20	SEWAGE indicates whether the household has access to a sewage system or septic tank.	Dispone Drenaje	
SEWAGE10		No Dispone Drenaje	
WATSUP11	WATSUP describes the physical means by which the housing unit receives its water. The primary distinction is whether or not the household had piped (running) water.	Dispone Agua Entubada	Instalación de tuberías que se planea y construye para abastecer de agua a las viviendas, edificios y escuelas, entre otros. Puede ser administrada por la entidad, el municipio, la comunidad o una empresa particular. No necesariamente es una instalación subterránea construida con tubos, puede ser superficial sin importar el tipo de material
WATSUP20		No dispone Agua Entubada	
PHONE20	PHONE indicates the availability of a fixed-line telephone in the dwelling.	Dispone Teléfono	Se refiere a la disponibilidad de teléfono con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
PHONE10		No Dispone Teléfono	
AUTOS20	AUTOS records whether a member of the household owned or had use of a vehicle, and in many samples the variable indicates the number of such vehicles.	Dispone Automóvil	Se refiere a la disponibilidad de automóvil o camioneta propios con que cuenta la vivienda respecto al estrato o nivel de bienestar de sus habitantes.
AUTOS10		No Dispone Automóvil	

Tabla 4 Parte 4; Matriz de Comparabilidad de Variables Principales

3.1.4 Validar armonización de variables principales

Matriz de Armonización de Variables.

La matriz de armonización de variables (Tabla 5) muestra las variables principales definidas anteriormente con su respectivo código que está siendo utilizado en el proyecto IPUMS.

Por otra parte se muestran los proyectos con los que trabajaremos por dominio de estudio, siendo: ENOE, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005. Dicha matriz es llenada con los siguientes valores:

‘1’ para la variable que está en el proyecto y dominio especificado.
‘0’ para la variable que no se encuentra dentro del proyecto y dominio especificado.

Una vez que se llenó la matriz, los muestristas deben validar que dicha información sea correcta, a través de sus conocimientos teóricos y prácticos de cada proyecto, esto es, cada proyecto cuenta con diferentes dominios de estudio y cada proyecto es realizado para distintos fines, de tal manera que no todos los proyectos son válidos para los dominios y variables principales que estamos estudiando. Por tal motivo, es de suma importancia que los muestristas revisen la validez de esta información.

No.	COD IPUMS	Descripción de Variable Armonizada	ENOE					CENSO 2000					CONTEO 2005				
			CDA	ENT	TLN	NNAL	EST	CDA	ENT	TLN	NNAL	EST	CDA	ENT	TLN	NNAL	EST
1	SEX1	Sexo	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	SEX2	Sexo	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	AGE	Edad	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	MARST1	Soltero/Nunca casado	1	1	0	1	0	1	1	1	1	1	0	0	0	0	0
5	MARST2	Casado/En unión libre	1	1	0	1	0	1	1	1	1	1	0	0	0	0	0
6	MARST3	Separado/Divorciado	1	1	0	1	0	1	1	1	1	1	0	0	0	0	0
7	MARST4	Viudo	1	1	0	1	0	1	1	1	1	1	0	0	0	0	0
8	HLTHCOV10	Derechohabiencia IMSS	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
9	HLTHCOV20	Derechohabiencia ISSSTE	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
10	HLTHCOV30	Derechohabiencia PEMEX	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
11	HLTHCOV40	Derechohabiencia OtraInst	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
12	HLTHCOV60	No tiene derechohabiencia	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	DISUPPR	Discapacidad Brazos	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
14	DISMOBL	Discapacidad Motriz	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
15	DISDEAF	Discapacidad Auditiva	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
16	DISMUTE	Discapacidad Lenguaje	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
17	DISBLND	Discapacidad Visual	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
18	DISMNTL	Discapacidad Mental	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
19	SPKIND	Habla Lengua Indígena	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
20	LIT	Alfabetismo	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1
21	SCHOOL	Asistencia Escolar	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1
22	WALL201	Material de desecho	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
23	WALL204	Lámina de cartón	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
24	WALL546	Lámina de asbesto o metálica	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
25	WALL405	Carrizo, bambú o palma	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
26	WALL532	Embarro o bajareque	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
27	WALL300	Madera	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
28	WALL523	Adobe	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
29	WALL501	Tabique, ladrillo, block, piedra, cantera, cemento o concreto	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
30	ROOF64	Material de desecho	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
31	ROOF65	Lámina de cartón	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0

Tabla 5, Parte 1; Matriz de armonización de variables

No.	COD IPUMS	Descripción de Variable Armonizada	ENOE					CENSO 2000					CONTEO 2005				
			CDA	ENT	TLN	NNAL	EST	CDA	ENT	TLN	NNAL	EST	CDA	ENT	TLN	NNAL	EST
32	ROOF34	Lámina de asbesto o metálica	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
33	ROOF40	Palma, tejamanil o madera	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
34	ROOF12	Teja	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
35	ROOF10	Losa de concreto, tabique, ladrillo o terrado con viguería	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
36	FLOOR100	Tierra	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
37	FLOOR202	Cemento o firme	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
38	FLOOR231	Madera, mosaico u otros recubrimientos	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
39	KITCHEN20	Disponen de cocina	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
40	KITCHEN10	No disponen de cocina	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
41	ELECTRC20	Dispone Electricidad	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
42	ELECTRC10	No dispone Electricidad	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
43	TV20	Dispone Televisión	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
44	TV10	No dispone Televisión	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
45	REFRIG20	Dispone Refrigerador	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
46	REFRIG10	No dispone Refrigerador	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
47	WASHER20	Dispone Lavadora	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
48	WASHER10	No dispone lavadora	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
49	COMPUTR20	Dispone Computadora	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
50	COMPUTR10	No dispone Computadora	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
51	TOILET20	Dispone Sanitario	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
52	TOILET10	No Dispone Sanitario	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
53	SEWAGE20	Dispone Drenaje	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
54	SEWAGE10	No Dispone Drenaje	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
55	PHONE20	Dispone Teléfono	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
56	PHONE10	No Dispone Teléfono	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
57	WATSUP11	Dispone Agua Entubada	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
58	WATSUP20	No dispone Agua Entubada	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
59	AUTOS20	Dispone Automóvil	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
60	AUTOS10	No Dispone Automóvil	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0

Tabla 5, Parte 2; Matriz de armonización de variables

Diagrama Entidad – Relación.

Una vez que se realizó la matriz de armonización de variables basándose en las tablas anteriores, se obtuvo como resultado final una base de datos que es representada por el diagrama Entidad-Relación (Figura 11).

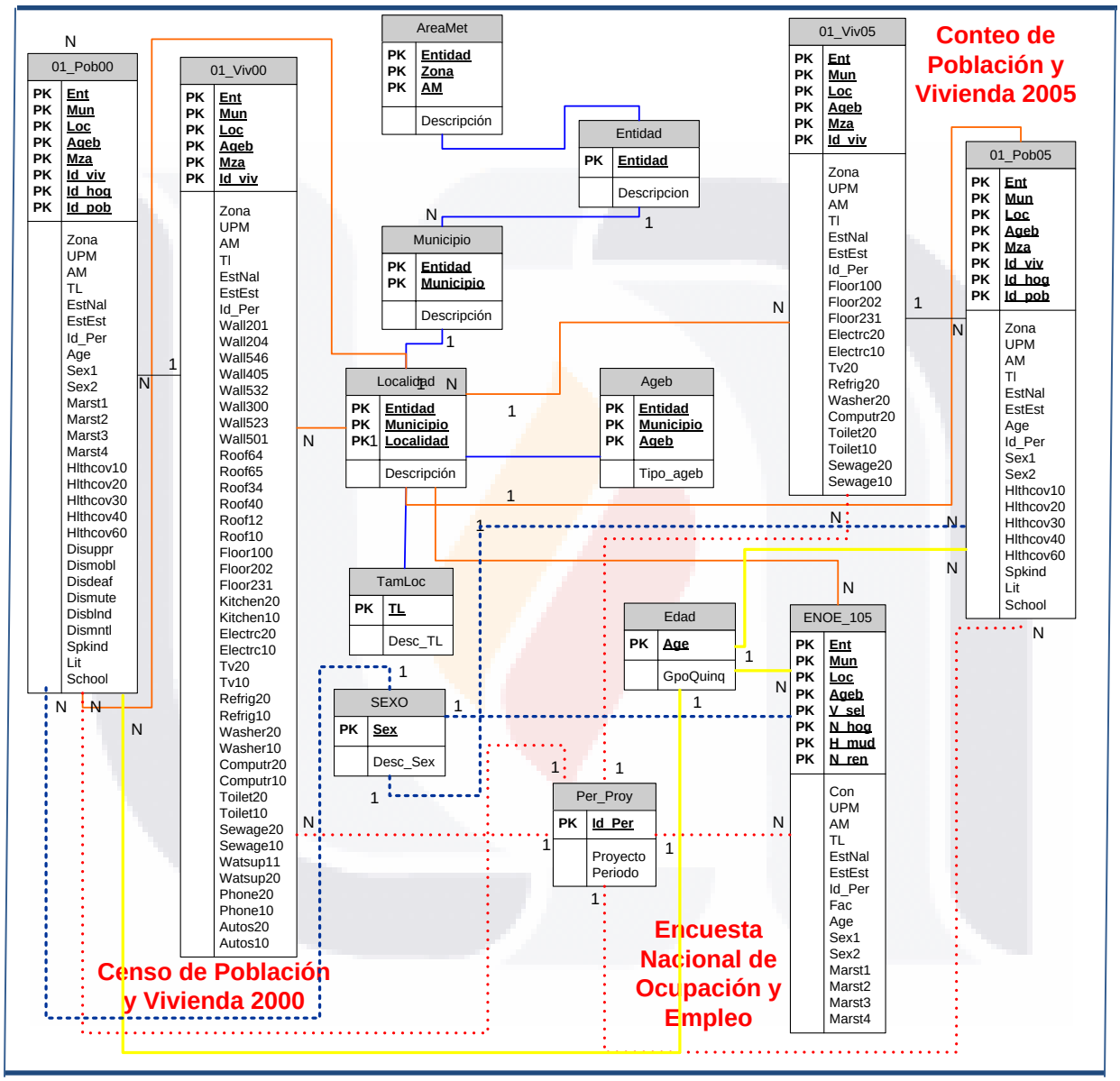


Figura 11; Diagrama entidad-relación de variables armonizadas

Base de Datos en Oracle.

Se realizó la programación pertinente a través de consultas de SQL para la construcción de las variables principales definidas para cada proyecto (Anexo), posteriormente se realizó la carga de los resultados a Oracle, lo cual fue un proceso demasiado tardado debido al número de registros que cada tabla contenía por lo cual se solicitaron 50 Gb de espacio en el servidor de Oracle.

Para los proyectos de Censo de PyV 2000 y Conteo de PyV 2005 se creó una tabla a nivel población y otra a nivel vivienda para cada entidad, teniendo como total 64 tablas para cada proyecto respectivamente. Al finalizar la carga de dichas tablas se unieron las 32 entidades en una sola a nivel población y otra a nivel vivienda dando como resultado 2 tablas por proyecto. (Tabla 6)

Para el proyecto ENOE se obtuvo una tabla por año de interés para nuestro estudio, siendo éstos del 2005 al 2009. Al realizar la carga al servidor de Oracle, se conjuntaron y se obtuvo una sola tabla para dicho proyecto. (Tabla 6)

Proyecto	Nombre de laTabla	No. de Registros
Censo de Población y Vivienda 2000	POB_00	97,483,412
	VIV_00	21,954,733
Conteo de Población y Vivienda 2005	POB_05	103,263,388
	VIV_05	24,719,029
Encuesta Nacional de Ocupación y Empleo	ENOE_05	1,698,242

Tabla 6; Registros totales en la Base de Datos de Oracle

3.2 Crear Data Mart

Para la construcción del Data Mart Constelación se propone seguir la Metodología HEFESTO anteriormente descrita.

1. Análisis de Requerimientos

a) Identificar preguntas. La subdirección de diseño muestral de Viviendas está interesado en contar con información de indicadores relevantes a diversos temas de interés a partir de diversas encuestas en hogares que se diseñan en el INEGI, entre las cuales se encuentran el Censo de Población y Vivienda 2000, el Conteo de Población y Vivienda 2005 y la Encuesta Nacional de Ocupación y Empleo. Existe una gran diversidad de indicadores, pues los temas que se tratan en los proyectos anteriormente mencionados son vastos y provienen de las variables de cada uno de estos proyectos, sin embargo se desea obtener los indicadores por Entidad y Tamaño de Localidad (Tabla 7).

No.	Indicador	No.	Indicador
1	Población total	**18**	Niños de 0 a 4 años
2	Población con derechohabiencia	**19**	Población de 0 a 14 años
3	Población que es derechohabiente al IMSS	**20**	Población de 15 a 64 años
4	Población que es derechohabiente al ISSSTE	**21**	Población de 65 años y más
5	Población que es derechohabiente al seguro de pemex	**22**	Mujeres de 12 a 49 años
6	Población que es derechohabiente a otra institución	**23**	Total de viviendas
7	Población sin derechohabiencia	**24**	Viviendas con drenaje
8	Población con derechohabiencia al IMSS ó ISSSTE	**25**	Viviendas con electricidad
9	Población de 6 a 17 que asiste a la escuela	**26**	Viviendas con piso diferente de tierra
10	Población de 6 a 17 años	**27**	Viviendas con televisión
11	Población de 5 años y más que habla lengua indígena	**28**	Viviendas con refrigerador
12	Población mayor a 5 años	**29**	Viviendas con lavadora
13	Población de 6 a 14 asiste a la escuela	**30**	Viviendas con computadora
14	Población de 6 a 14 años	**31**	Viviendas con piso de tierra
15	Población de 15 años y más que es alfabeta	**32**	Viviendas con sanitario
16	Población de 15 años y más	**33**	Viviendas con televisión y lavadora
17	Población que asiste a la escuela y habla lengua indígena	**34**	Viviendas con drenaje y electricidad

Tabla 7; Indicadores a obtener.

b) Identificar indicadores y perspectivas de análisis.

A continuación se analizan las respuestas del paso anterior para obtener los indicadores y perspectivas.

"Población total por entidad a nivel TL en cierto período de tiempo"

INDICADOR — PERSPECTIVAS

En síntesis:

- Indicadores:
  - Población total.
- Perspectivas
  - Entidad.
  - TL.
  - Tiempo.

"Población con derechohabiencia por entidad a nivel TL en cierto período de tiempo"

INDICADOR — PERSPECTIVAS

En síntesis:

- Indicadores:
  - Población con derechohabiencia.
- Perspectivas
  - Entidad.
  - TL.
  - Tiempo.

Se analizaron los treinta indicadores que se desean obtener (Tabla 7) y se observó en trece de ellos la dimensión de la "Edad".

Por otra parte, los muestristas nos señalaron la importancia de hacer la distinción de los indicadores de población por género masculino o femenino; por lo tanto, basándonos en dichas observaciones y en los ejemplos anteriores podemos deducir que para los indicadores de población que deseamos obtener las perspectivas de análisis son:

- Entidad.
- Tamaño de Localidad. (TL)
- Tiempo.
- Edad.
- Sexo.

Sin embargo, para los indicadores de vivienda que deseamos obtener las perspectivas de análisis son:

- Entidad.
- Tamaño de Localidad. (TL)
- Tiempo.

Debido al comportamiento de las dimensiones generadas se observaron tres situaciones en los indicadores a obtener:

1. El indicador se mantiene, es decir, se realizará el proceso pertinente para su cálculo apropiado.

2. Cambió de nombre, es decir, el indicador se compone de una de las dimensiones que se van a generar, por lo tanto, en este proceso sólo se calculará el indicador independiente a dichas dimensiones y el nombre del indicador original cambió.

3. Eliminación del indicador, debido a que el indicador se compone en su totalidad de una de las dimensiones que se van a generar, por lo tanto, dicho indicador se elimina en este proceso y al momento de obtener los resultados finales, se retomará dicho indicador.

Los indicadores que cambian o se eliminan son:

No.	Nombre	Situación
1	Población de 6 a 17 que asiste a la escuela	Cambio (Población que asiste a la escuela)
2	Población de 6 a 17 años	Eliminación
3	Población de 5 años y más que habla lengua indígena	Cambio (Población que habla lengua indígena)
4	Población mayor a 5 años	Eliminación
5	Población de 6 a 14 asiste a la escuela	Cambio(Población que asiste a la escuela)
6	Población de 6 a 14 años	Eliminación
7	Población de 15 años y más que es alfabeta	Cambio (Población que es alfabeta)
8	Población de 15 años y más	Eliminación
9	Niños de 0 a 4 años	Eliminación
10	Población de 0 a 14 años	Eliminación
11	Población de 15 a 64 años	Eliminación
12	Población de 65 años y más	Eliminación
13	Mujeres de 12 a 49 años	Eliminación

Tabla 8; Situación de indicadores a obtener.

c) Modelo conceptual.

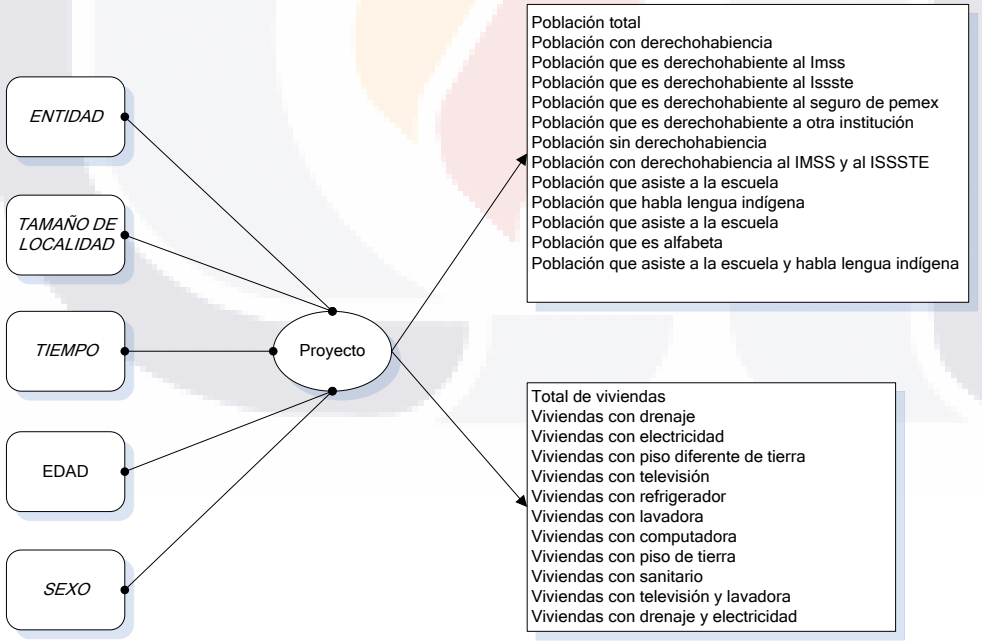


Figura 12; Modelo conceptual

2. Análisis de los OLTP

a) Determinación de indicadores.

Los indicadores se calcularán de la siguiente manera:

- “Población Total”
  - Hechos: Población Total
  - Función: COUNT ().

- “Población con Derechohabiencia”
  - Hechos:HLTHCOV10 + HLTHCOV20 + HLTHCOV30 + HLTHCOV40
  - Función: SUM

- “Población que es derechohabiente al IMSS”
  - Hechos: HLTHCOV10
  - Función: SUM

- “Población que es derechohabiente al ISSSTE”
  - Hechos: HLTHCOV20
  - Función: SUM

- “Población que es derechohabiente a PEMEX”
  - Hechos: HLTHCOV30
  - Función: SUM

- “Población sin derechohabiencia”
  - Hechos: HLTHCOV60
  - Función: SUM

- “Población que asiste a la escuela”
  - Hechos: SCHOOL
  - Función: SUM

- “Población que asiste a la escuela”
  - Hechos: SCHOOL
  - Función: SUM

- “Población que es alfabeta”
  - Hechos: LIT
  - Función: SUM

- “Población que asiste a la escuela y habla lengua indígena”
  - Hechos: SCHOOL AND SPKIND
  - Función: SUM

- “Total de viviendas”
  - Hechos: TOTVIVS
  - Función: COUNT()

- “Viviendas con drenaje”
  - Hechos: SEWAGE20
  - Función: SUM

- “Viviendas con electricidad”
  - Hechos: ELECTRC20
  - Función: SUM

- “Viviendas con piso diferente a tierra”
  - Hechos: FLOOR202 OR FLOOR231
  - Función: SUM

- “Viviendas con televisión”
  - Hechos: TV20
  - Función: SUM

- “Viviendas con refrigerador”
  - Hechos: REFRIG20
  - Función: SUM

- “Viviendas con lavadora”
  - Hechos: WASHER20
  - Función: SUM

- “Viviendas con computadora”
  - Hechos: COMPUTR20
  - Función: SUM

- “Viviendas con piso de tierra”
  - Hechos: FLOOR100
  - Función: SUM

- “Viviendas con sanitario”
  - Hechos: TOILET20
  - Función: SUM

- “Viviendas con televisión y lavadora”
  - Hechos: TV20 AND WASHER20
  - Función: SUM

- “Viviendas con drenaje y electricidad”
  - Hechos: SEWAGE20 AND ELECTRC20
  - Función: SUM

b) Establecer correspondencias.

La base de datos generada del proceso de armonización de variables dió como resultado el diagrama entidad – relación (Figura 11) que será nuestra base para establecer las correspondencias pertinentes (Figura 13) con el modelo conceptual.

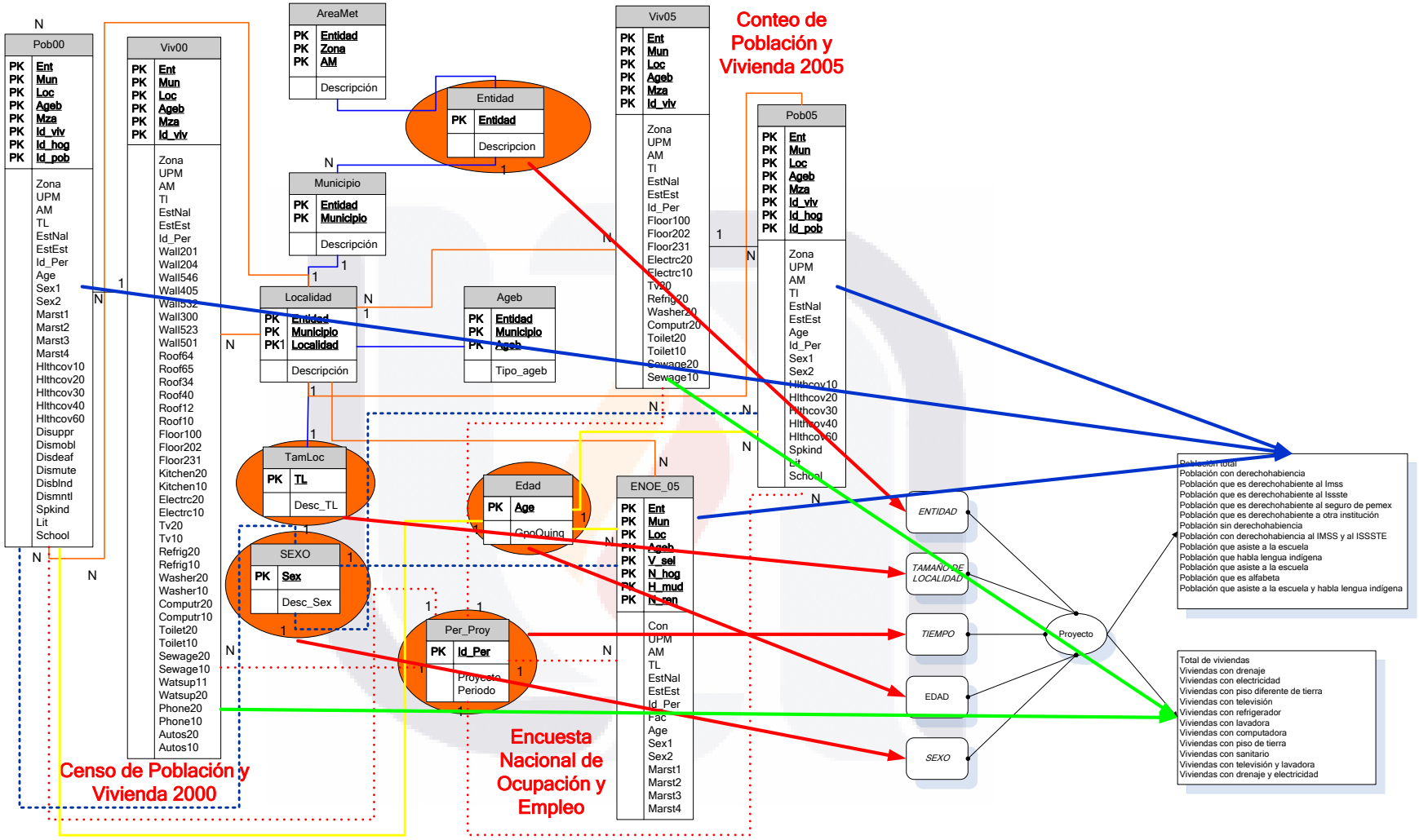


Figura 13; Correspondencias

Las relaciones identificadas fueron las siguientes:

- La Tabla Entidad se relaciona con la perspectiva Entidad.
- La Tabla Tam_Loc se relaciona con la perspectiva Tamaño de Localidad (TL).
- La Tabla Per_Proy se relaciona con la perspectiva Tiempo.
- La Tabla Edad se relaciona con la perspectiva Edad.
- La Tabla Sexo se relaciona con la perspectiva Tiempo.
- La Tabla Pob00, Pob05 y Enoe_05 se relacionan con los indicadores referentes a Población
- La Tabla Viv00, Viv05 se relacionan con los indicadores referentes a Vivienda.

c) Nivel de granularidad.

Con respecto a la perspectiva Entidad se tiene la tabla Entidad, la cual se relaciona con las tablas Municipio y Localidad de las cuales podemos obtener los siguientes datos:

- Entidad: Es un número representativo de una entidad en particular.
- Descripcion: Nombre de la Entidad.
- Municipio: Es un número representativo de un municipio particular.
- Descripcion: Nombre del Municipio.
- Localidad: Es un número representativo de una localidad en particular.
- Descripcion: Nombre de la Localidad.

Con respecto a la perspectiva TL se puede obtener de las tablas:

- TL: Es un número representativo de un tamaño de localidad en particular.
- Desc_TL: Descripción del Tamaño de Localidad

Con respecto a la perspectiva Tiempo se tienen los siguientes datos:

- Id_Per: Es la clave primaria de la tabla Per_proy y representa unívocamente a un proyecto en un período específico.
- Proyecto: Nombre del proyecto.
- Período: Período correspondiente al proyecto, formado por mes y año.

Con respecto a la perspectiv Edad se tienen los siguientes datos:

- Age: Es la clave primaria, además de representar el número de años cumplidos por la persona, desde la fecha de su Nacimiento hasta el momento del hecho.
- GpoQuinq: Se refiere a la agrupación de edades por quinquenio (Tabla 9)

Grupo Quinquenal	Rango de Edad
1	0 - 4 años
2	5 - 9 años
3	10 - 14 años
4	15 - 19 años
5	20 - 24 años
6	25 - 29 años
7	30 - 34 años
8	35 - 39 años
9	40 - 44 años
10	45 - 49 años
11	50 - 54 años
12	55 - 59 años
13	60 - 64 años
14	65 y más años

Tabla 9; Grupos quinquenales

Con respecto a la perspectiva Sexo se tienen los siguientes datos:

- IdSex: Condición biológica que distingue a las personas en hombres y mujeres Hombre = 1 y mujer = 2.
- DescSex: Nombre del género al que se pertenece; HOMBRE, MUJER.

d) Modelo Conceptual ampliado.

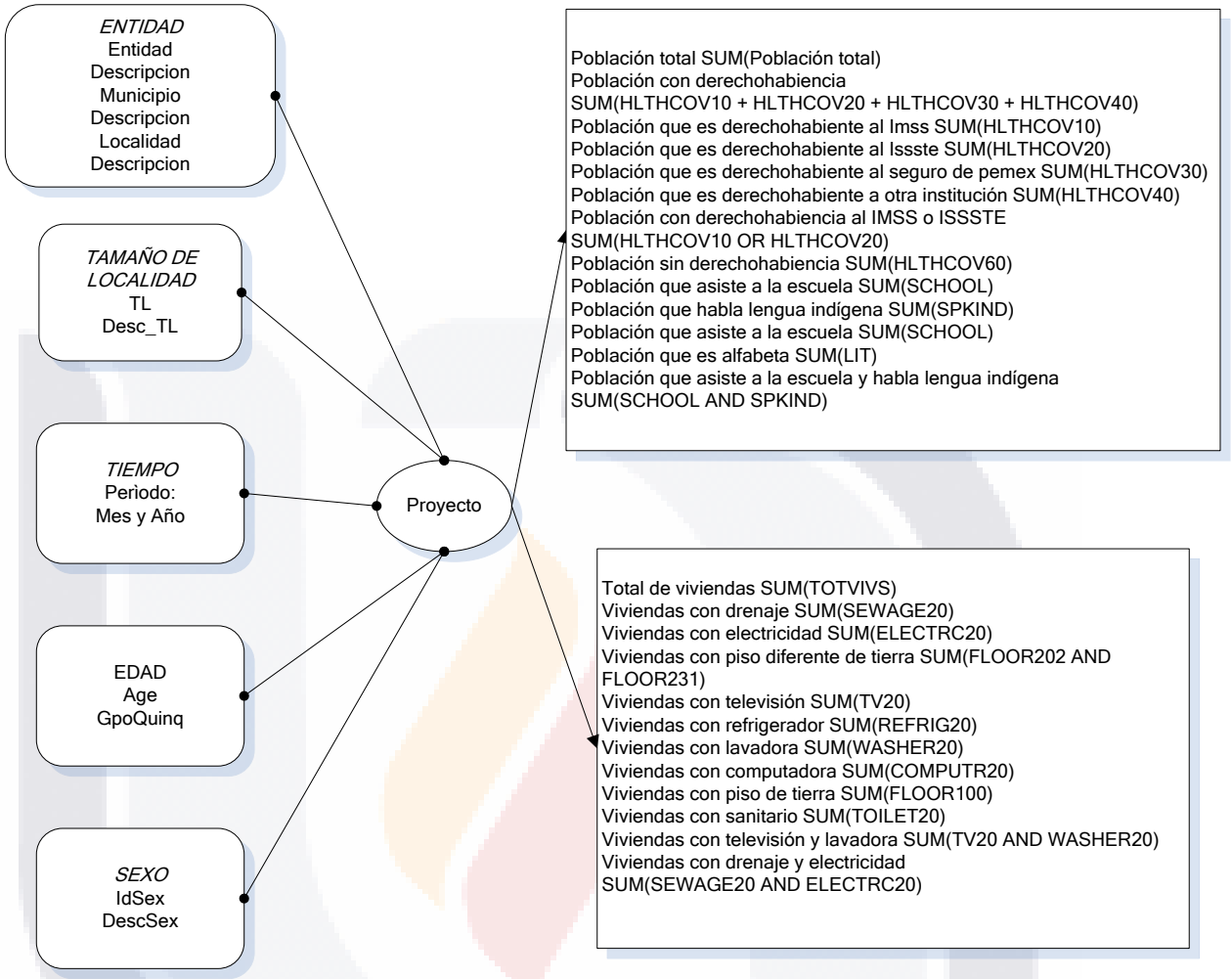


Figura 14; Modelo conceptual ampliado

3) Modelo Lógico del DW

a) Tipo de Modelo Lógico del DW.

El esquema será en Constelación debido a sus características.

b) Tablas de dimensiones.

- Perspectiva “Entidad”:

La Tabla de dimensión tendrá el nombre de “ENTIDAD”.

Se le asigna una clave principal con el nombre de “idEntidad”.

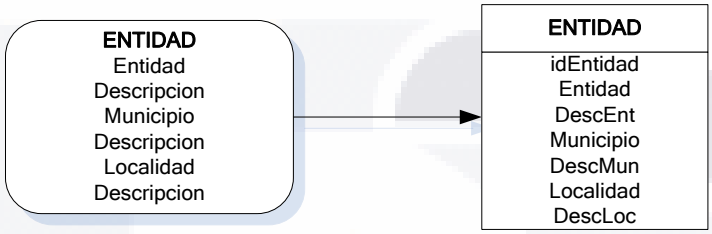


Figura 15; Tabla de Dimensión “ENTIDAD”

- Perspectiva “TL”:

La Tabla de dimensión tendrá el nombre de “TAM_LOC”.

Se le asigna una clave principal con el nombre de “idTL”.

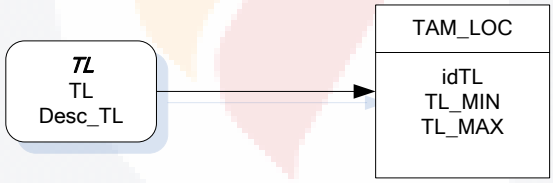


Figura 16; Tabla de Dimensión “TAM_LOC”

- Perspectiva “Tiempo”:

La Tabla de dimensión tendrá el nombre de “TIEMPO”.

Se le asigna una clave principal con el nombre de “idPer”.

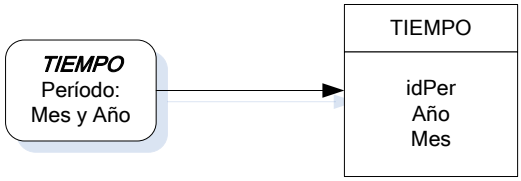


Figura 17; Tabla de Dimensión “TIEMPO”

- Perspectiva “Edad”:

La Tabla de dimensión tendrá el nombre de “EDAD”.

Se le asigna una clave principal con el nombre de “Age”.

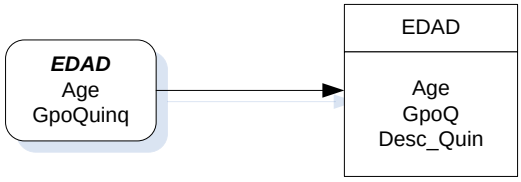


Figura 18; Tabla de Dimensión “EDAD”

- Perspectiva “Sexo”:

La Tabla de dimensión tendrá el nombre de “SEXO”.

Se le asigna una clave principal con el nombre de “SEX1” o “SEX2”, según corresponda.

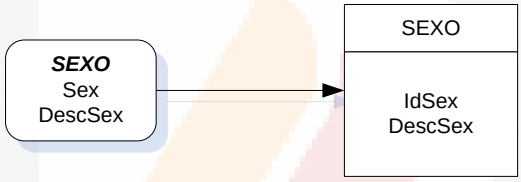


Figura 19; Tabla de Dimensión “SEXO”

c) Tablas de hechos.

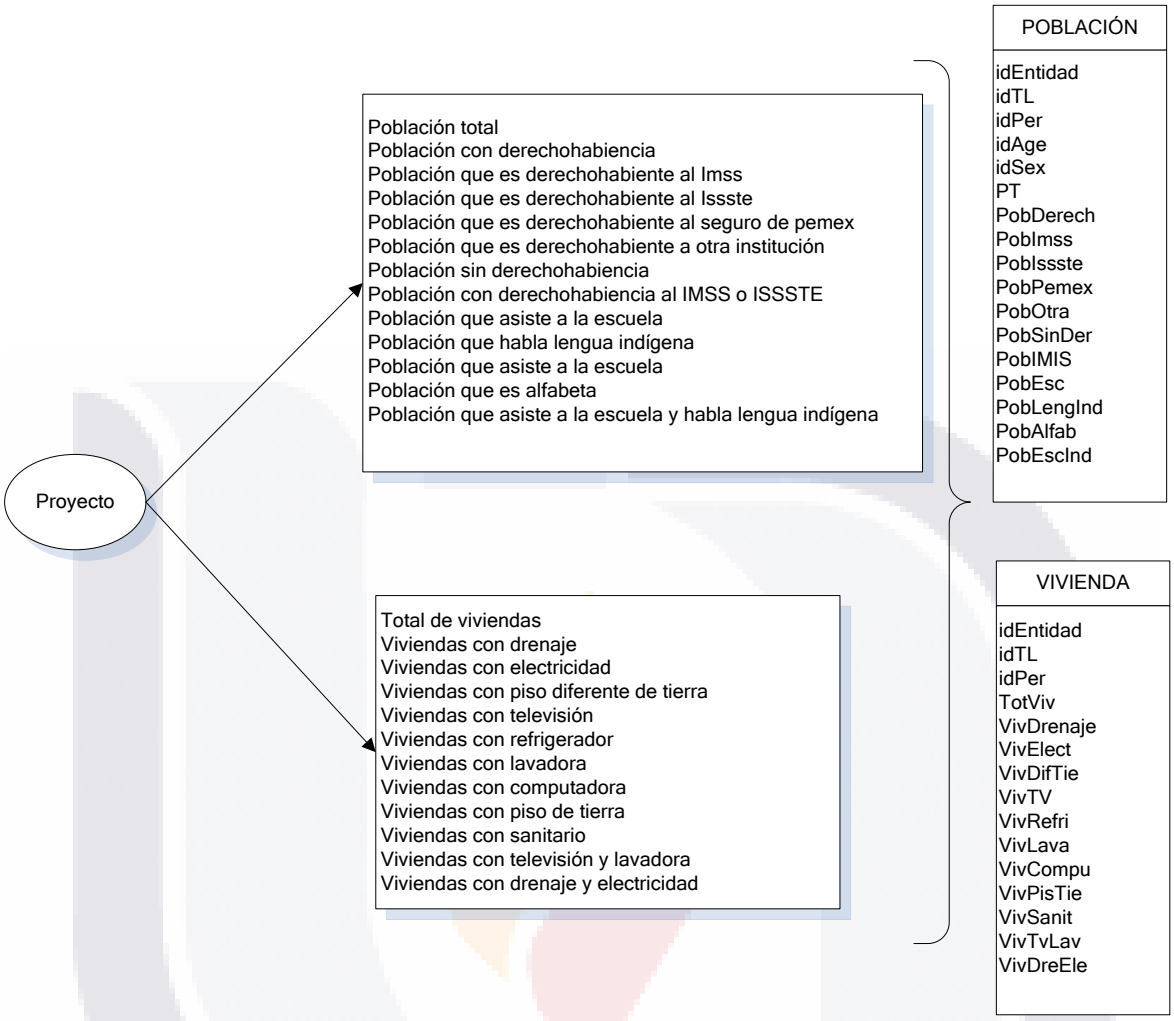


Figura 20; Diseño de la Tabla de Hechos

d) Uniones.

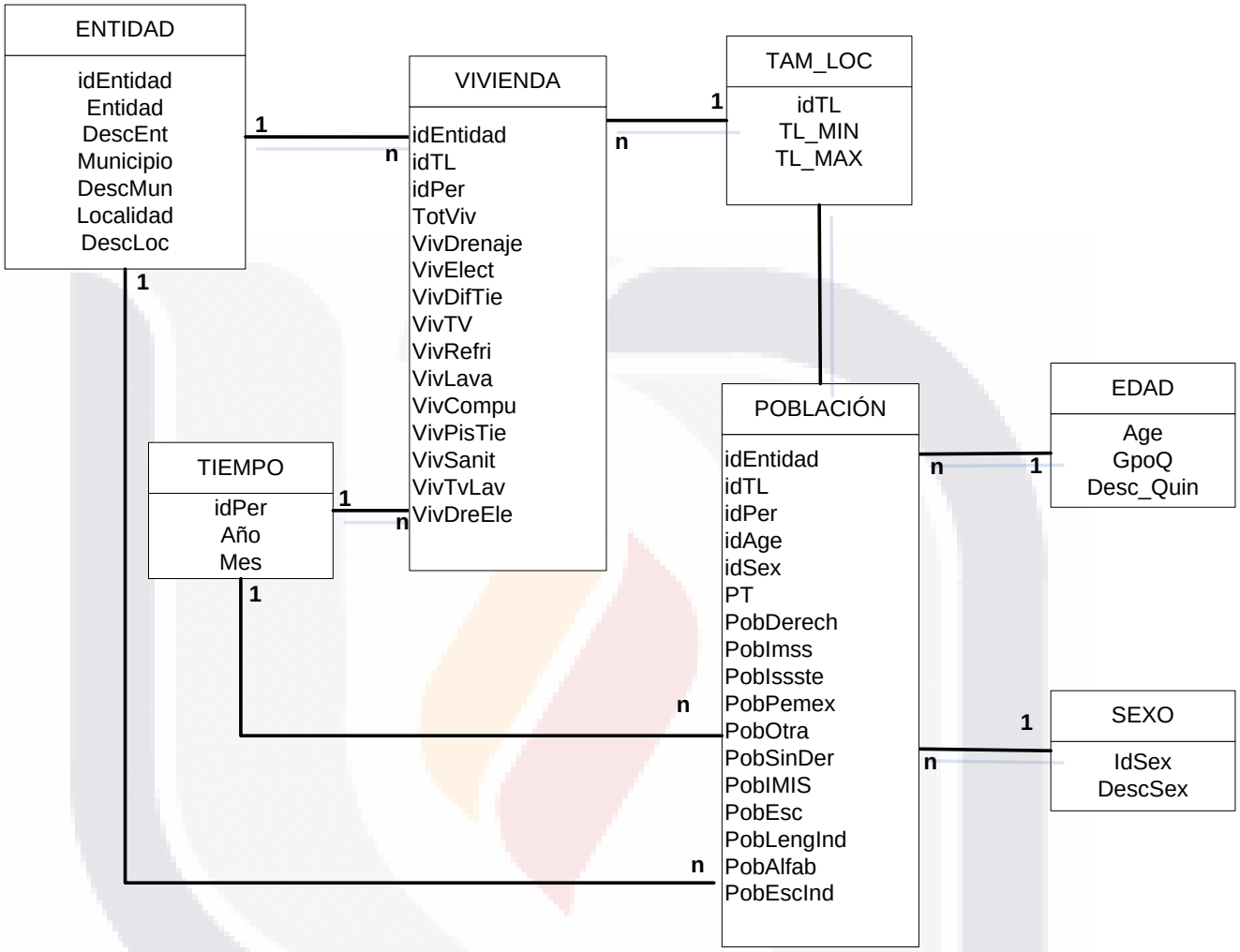


Figura 21; Esquema Constelación

4) Procesos ETL

Se generaron las sentencias SQL para cargar las tablas de dimensiones y la tabla de hechos. Las tablas fuente se tomarán del Diagrama Entidad – Relación (Figura 11) y de Correspondencias (Figura 13) descritos anteriormente.

TABLAS DE DIMENSIÓN

Tabla de Dimensión “ENTIDAD”.

Se toman como fuentes las tablas Entidad, Municipio, Localidad.

```
Select a.cve_ent+b.Municipio+c.Localidad as IdEntidad,;
a.cve_Ent as Entidad,a.Desc_ent as DescEnt,;
 b.Municipio as Municipio, b.Descripcio as Desc_mun,;
c.Localidad as Localidad,c.Descloc as DescLoc;
from "dbf fuentes\Entidades" a;
inner join "dbf fuentes\tc_Municipios" b;
on a.cve_ent=b.Entidad;
inner JOIN tc_localidades c;
on b.Entidad=c.Entidad AND b.Municipio =c.Municipio;
INTO TABLE Entidad
```

Tabla de Dimensión “TAM_LOC”

Se toma como fuente la tabla Tam_Loc.

```
Select  tl as idTL, tl as TL, desc_tl as .DescTL;
from "dbf fuentes\Tam_Loc";
INTO TABLE Tam_Loc
```

Tabla de Dimensión “TIEMPO”

Se toma como fuente la tabla Per_proy.

```
Select id_per as Id_per,;
 SUBSTR(Periodo,1,2) as Mes, SUBSTR(Periodo,3,4) as anio;
from "Per_ref";
INTO TABLE Tiempo
```

Tabla de Dimensión “EDAD”

Se toma como fuente la tabla Edad.

```
SELECT Age as IdAge, Gpo_quin as GpoQ;
 FROM “Edad”;
 INTO TABLE EDAD
```

Tabla de Dimensión “SEXO”

Se toma como fuente la tabla Sexo.

```
SELECT Sex as IdSex, DescSex as DescSex;
 FROM “Sexo”;
 INTO TABLE SEXO
```

TABLAS DE HECHOS.

1. Construir tabla de Hechos POBLACIÓN con los indicadores de los proyectos Censo de PyV 2000, Conteo de PyV2005 y ENOE 2005 a 2009.

```
SELECT ENT || MUN || LOC IDENTIDAD,TL IDTL,ID_PER IDPER, AGE IDAGE,
DECODE(SEX1,1,'1',0,'2' ) IDSEX,COUNT(*) PT,
SUM(HLTHCOV10 + HLTHCOV20 + HLTHCOV30 + HLTHCOV40) POBDERECH,
SUM(HLTHCOV10) POBIMSS, SUM(HLTHCOV20) POBISSSTE,
SUM(HLTHCOV30) POBPEMEX, SUM(HLTHCOV40) POBOTRA,
SUM(HLTHCOV60) POBSINDER,
SUM(CASE WHEN HLTHCOV10=1 OR HLTHCOV20=1 THEN 1 ELSE 0 END)  POBIMIS,
SUM(SCHOOL)  POBESC, SUM(SPKIND)  POBLENGIND, SUM(LIT)  POBALFAB,
SUM(CASE WHEN SCHOOL=1 AND SPKIND=1 THEN 1 ELSE 0 END)  POBESCIND
FROM POB_00 GROUP BY   ENT, MUN, LOC, TL, ID_PER, AGE, SEX1
UNION ALL
SELECT ENT || MUN || LOC IDENTIDAD,TL IDTL,ID_PER IDPER, AGE IDAGE,
DECODE(SEX1,1,'1',0,'2' ) IDSEX,COUNT(*) PT,
SUM(HLTHCOV10 + HLTHCOV20 + HLTHCOV30 + HLTHCOV40) POBDERECH,
SUM(HLTHCOV10) POBIMSS,SUM(HLTHCOV20) POBISSSTE,
SUM(HLTHCOV30) POBPEMEX,SUM(HLTHCOV40) POBOTRA,
SUM(HLTHCOV60) POBSINDER,
SUM(CASE WHEN HLTHCOV10=1 OR HLTHCOV20=1 THEN 1 ELSE 0 END)  POBIMIS,
SUM(SCHOOL)  POBESC, SUM(SPKIND)  POBLENGIND, SUM(LIT)  POBALFAB,
SUM(CASE WHEN SCHOOL=1 AND SPKIND=1 THEN 1 ELSE 0 END)  POBESCIND
FROM POB_05 GROUP BY   ENT, MUN, LOC, TL, ID_PER, AGE, SEX1
UNION ALL
SELECT ENT || MUN || LOC IDENTIDAD,TL IDTL,ID_PER IDPER, AGE IDAGE,
DECODE(SEX1,1,'1',0,'2' ) IDSEX,COUNT(*) PT,
SUM(HLTHCOV10 + HLTHCOV20 + HLTHCOV30 + HLTHCOV40) POBDERECH,
SUM(HLTHCOV10) POBIMSS,SUM(HLTHCOV20) POBISSSTE,
SUM(HLTHCOV30) POBPEMEX,SUM(HLTHCOV40) POBOTRA,
SUM(HLTHCOV60) POBSINDER,
SUM(CASE WHEN HLTHCOV10=1 OR HLTHCOV20=1 THEN 1 ELSE 0 END)  POBIMIS,
SUM(SCHOOL)  POBESC, SUM(SPKIND)  POBLENGIND, SUM(LIT)  POBALFAB,
SUM(CASE WHEN SCHOOL=1 AND SPKIND=1 THEN 1 ELSE 0 END)  POBESCIND
FROM ENOE GROUP BY   ENT, MUN, LOC, TL, ID_PER, AGE, SEX1
```

2. Construir tabla de Hechos VIVIENDA con indicadores de los proyectos: Censo de PyV 2000 y Conteo de PyV 2005.

```
SELECT ENT || MUN || LOC IDENTIDAD,TL IDTL,ID_PER IDPER,
COUNT(*) VIVS,
SUM(SEWAGE20)  VIVDRENAJE,
SUM(ELECTRC20)  VIVELECT,
SUM(CASE WHEN FLOOR202=1 OR FLOOR231=1 THEN 1 ELSE 0 END)  VIVDIFTIE,
SUM(TV20)  VIVTV,
SUM(REFRIG20)  VIVREFRI,
SUM(WASHER20)  VIVLAVA,
SUM(COMPUTR20)  VIVCOMPU,
SUM(FLOOR100)  VIVPISTIE,
SUM(TOILET20)  VIVSANIT,
SUM(CASE WHEN TV20=1 AND WASHER20=1 THEN 1 ELSE 0 END)  VIVTVLAV,
SUM(CASE WHEN SEWAGE20=1 AND ELECTRC20=1 THEN 1 ELSE 0 END)  VIVDREELE
FROM VIV_00  GROUP BY   ENT, MUN, LOC, TL, ID_PER
UNION ALL
SELECT ENT || MUN || LOC IDENTIDAD,TL IDTL,ID_PER IDPER,
COUNT(*) VIVS,
SUM(SEWAGE20)  VIVDRENAJE,
SUM(ELECTRC20)  VIVELECT,
SUM(CASE WHEN FLOOR202=1 OR FLOOR231=1 THEN 1 ELSE 0 END)  VIVDIFTIE,
SUM(TV20)  VIVTV,
SUM(REFRIG20)  VIVREFRI,
SUM(WASHER20)  VIVLAVA,
SUM(COMPUTR20)  VIVCOMPU,
SUM(FLOOR100)  VIVPISTIE,
SUM(TOILET20)  VIVSANIT,
SUM(CASE WHEN TV20=1 AND WASHER20=1 THEN 1 ELSE 0 END)  VIVTVLAV,
SUM(CASE WHEN SEWAGE20=1 AND ELECTRC20=1 THEN 1 ELSE 0 END)  VIVDREELE
FROM VIV_05  GROUP BY   ENT, MUN, LOC, TL, ID_PER
```

Cubo Multidimensional

Se crearán tres cubos multidimensionales, tomando los datos de las tablas de hechos POBLACIÓN y VIVIENDA además de las tablas de dimensión ENTIDAD, TAMAÑO DE LOCALIDAD, TIEMPO, EDAD y SEXO; según corresponda.

1. " Población y Viviendas en México "

Creación de Indicadores:

- De la tabla POBLACIÓN se sumarizará PT para crear el indicador:
  "Población mexicana" = SUM(POBLACION.PT)
- De la tabla VIVIENDAS se sumarizará VIVS para crear el indicador:
  "Viviendas en México" = SUM(VIVIENDA.VIVS)

Creación de atributos:

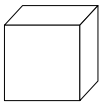
- De la tabla " ENTIDAD " se toma " DESC_ENT " para el atributo "Entidad"
- De la tabla " ENTIDAD " se toma " DESC_MUN " para el atributo " Municipio "
- De la tabla "TAM_LOC" se toma "DESC_TL" para el atributo "Tam de Localidad"
- De la tabla " TIEMPO " se toma " AÑO " para el atributo "Año"
- De la tabla " TIEMPO " se toma " MES " para el atributo "Mes"

Creación de jerarquías.

- Se definió la Jerarquía Geográfica que se aplicará a los atributos Entidad y Municipio, en donde:
  Una entidad tiene uno o más municipios.
  Un Municipio pertenece a una sola entidad

- Se definió la Jerarquía Tiempo que se aplicará a los atributos Año y Mes.
  Un año tiene uno o más meses.

El Cubo Multidimensional de Población y Viviendas en México queda conformado por cinco atributos, dos indicadores y dos jerarquías (Figura 22).



Población y Viviendas en México

- Entidad [ENTIDAD.DescEnt]
- Municipio [ENTIDAD.DescMun]
- Tamaño de Localidad [TamLoc.DescTL]
- Año [TIEMPO.Año]
- Mes [TIEMPO.Mes]
- Población Mexicana [Sum(POBLACIÓN.Pt)]
- Viviendas en México [Sum(VIVIENDA.Vivs)]

Jerarquía Geográfica
- Entidad
  - Municipio

Jerarquía de Tiempo
- Año
  - Mes

Figura 22; Cubo Multidimensional de Población y Viviendas en México

Para la obtención del Cubo de Población y Vivienda en México (Figura 20), se realizó la siguiente consulta:

```
SELECT  C.DESCENT, C.DESC_MUN, A.IDTL, A.IDPER, D.MES, D.ANIO,;
A.VIVS AS VT, SUM(B.PT) AS PT;
FROM VIVIENDA A INNER JOIN POBLACION B;
ON A.IDENTIDAD=B.IDENTIDAD AND A.IDTL = B.IDTL;
INNER JOIN ENTIDAD C ON A.IDENTIDAD=C.IDENTIDAD;
INNER JOIN TIEMPO D ON A.IDPER=D.ID_PER;
GROUP BY DESCENT,DESC_MUN,A.IDTL,A.IDPER,D.MES,D.ANIO,A.VIVS;
INTO TABLE “CUBO_POBYVIV”
```

2. “ Servicios de Salud en México “

Creación de Indicadores:

- De la tabla POBLACIÓN se sumarizará PT para crear el indicador:
  “Población mexicana” = SUM(POBLACION.PT)
- De la tabla POBLACIÓN se sumarizará PobDerech para crear el indicador:
  “Población con Derechohabiencia” = SUM(POBLACION.PobDerech)
- De la tabla POBLACIÓN se sumarizará PobImss para crear el indicador:
  “ Población con Derechohabiencia al IMSS” = SUM(POBLACION.PobImss)
- De la tabla POBLACIÓN se sumarizará PobIssste para crear el indicador:
  “Población con Derechohabiencia al ISSTE” = SUM(POBLACION.PobIssste)
- De la tabla POBLACIÓN se sumarizará PobPemex para crear el indicador:
  “Población con Derechohabiencia a PEMEX”= SUM(POBLACION.PobPemex)
- De la tabla POBLACIÓN se sumarizará PobOtra para crear el indicador:
  “Población con Derechohabiencia a otra Institución de Salud” =
  SUM(POBLACION.PobOtra)
- De la tabla POBLACIÓN se sumarizará PobSinDer para crear el indicador:
  “Población sin Derechohabiencia” = SUM(POBLACION.PobSinDer)

Creación de atributos:

- De la tabla “ ENTIDAD ” se toma “ DESC_ENT ” para el atributo “Entidad”
- De la tabla “ ENTIDAD ” se toma “ DESC_MUN ” para el atributo “ Municipio ”
- De la tabla “TAM_LOC” se toma “DESC_TL” para el atributo “Tam de Localidad”
- De la tabla “ TIEMPO ” se toma “ AÑO ” para el atributo “Año”
- De la tabla “ TIEMPO ” se toma “ MES ” para el atributo “Mes”
- De la tabla “ EDAD ” se toma “ AGE ” y “GPO_Q ” para los atributos “ Edad ” y “Grupo Quinquenal ”, respectivamente.
- De la tabla de dimensión “SEXO” se toma “DESCSEX” para el atributo “Sexo”.

Creación de jerarquías.

- Se definió la Jerarquía Geográfica que se aplicará a los atributos Entidad y Municipio, en donde:

  Una entidad tiene uno o más municipios.

  Un Municipio pertenece a una sola entidad

- Se definió la Jerarquía Tiempo que se aplicará a los atributos Año y Mes.

  Un año tiene uno o más meses.

El Cubo Multidimensional queda conformado por nueve atributos, siete indicadores y dos jerarquías (Figura 23).

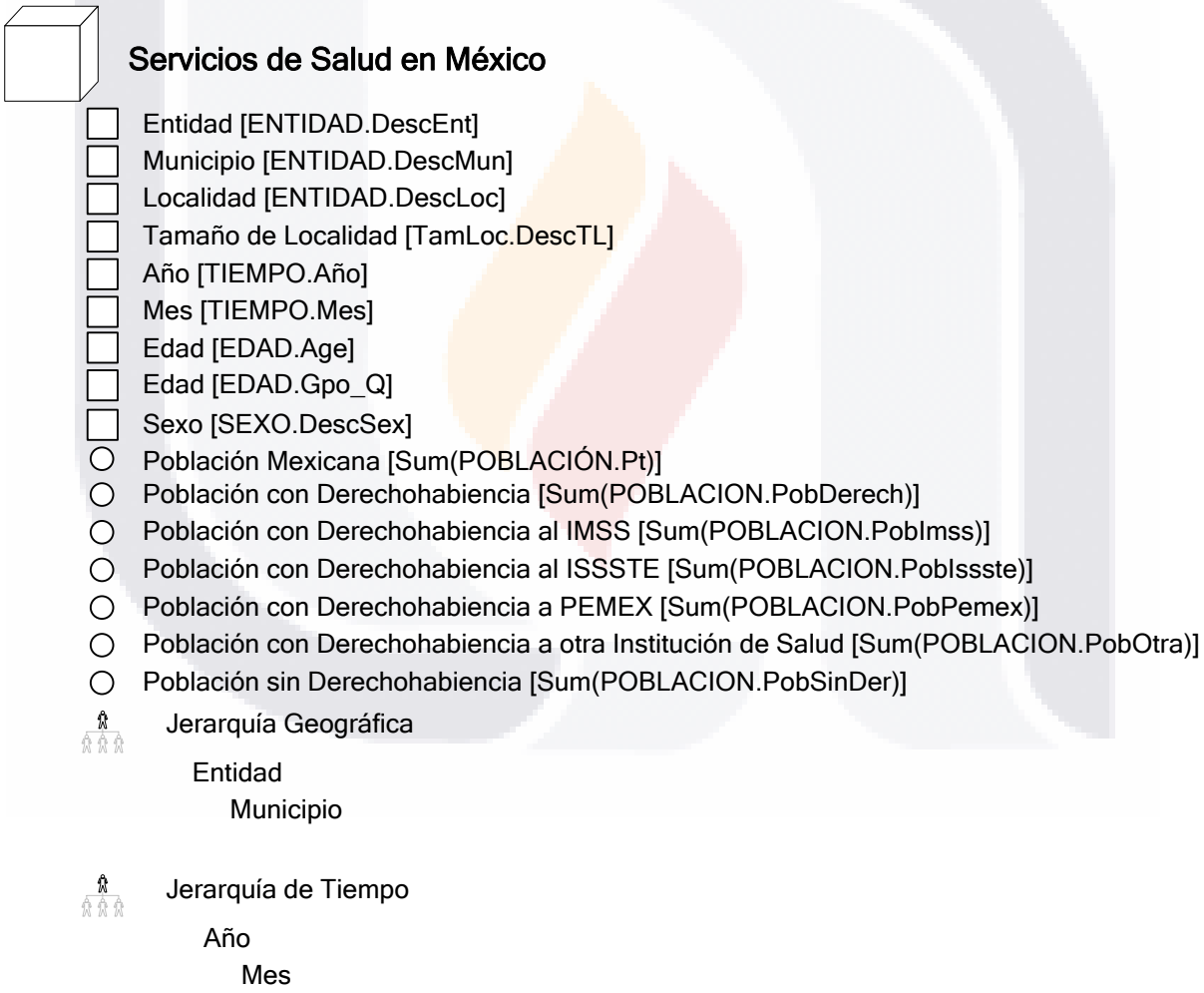


Figura 23; Cubo Multidimensional de Servicios de Salud en México

Para la obtención del Cubo Servicios de Salud en México (Figura 23), se realizó la siguiente consulta:

```
SELECT  C.DESCENT, C.DESC_MUN, A.IDTL, A.IDPER, D.MES,
D.ANIO,E.AGE,E.GPO_QUIN,F.DESCSEX,;
SUM(A.PT) AS PT,;
SUM(A.POBDERECH) AS POBDER,;
SUM(A.POBIMSS) AS POBIMSS,;
SUM(A.POBISSSTE) AS POBISSSTE,;
SUM(A.POBPEMEX) AS POBPEMEX,;
SUM(A.POBOTRA) AS POBOTRA,;
SUM(A.POBSINDER) AS POBSINDER;
FROM POBLACION A;
INNER JOIN ENTIDAD C ON A.IDENTIDAD=C.IDENTIDAD;
INNER JOIN TIEMPO D ON A.IDPER=D.ID_PER;
INNER JOIN EDAD E ON A.IDAGE=E.AGE;
INNER JOIN SEXO F ON A.IDSEX=F.IDSEX;
GROUP BY DESCENT, DESC_MUN ,A.IDTL, A.IDPER, D.MES, D.ANIO,;
 E.AGE, E.GPO_QUIN, F.DESCSEX;
INTO TABLE "CUBO_SERV_SALUD"
```

3. “ Bienes de las Viviendas en México “

Creación de Indicadores:

- De la tabla VIVIENDA se sumarizará TOTVIV para crear el indicador:
  “Viviendas mexicanas” = SUM(VIVIENDA.TotViv)
- De la tabla VIVIENDA se sumarizará VIVDRENAJE para crear el indicador:
  “Viviendas con drenaje” = SUM(VIVIENDA.VivDrenaje)
- De la tabla VIVIENDA se sumarizará VIVELECT para crear el indicador:
  “Viviendas con electricidad” = SUM(VIVIENDA.VivElect)
- De la tabla VIVIENDA se sumarizará VIVTV para crear el indicador:
  “Viviendas con televisión” = SUM(VIVIENDA.VivTV)
- De la tabla VIVIENDA se sumarizará VIVREFRI para crear el indicador:
  “Viviendas con refrigerador” = SUM(VIVIENDA.VivRefri)
- De la tabla VIVIENDA se sumarizará VIVLAVA para crear el indicador:
  “Viviendas con lavadora” = SUM(VIVIENDA.VivLava)
- De la tabla VIVIENDA se sumarizará VIVCOMPU para crear el indicador:
  “Viviendas con computadora” = SUM(VIVIENDA.VivCompu)

Creación de atributos:

- De la tabla “ ENTIDAD ” se toma “ DESC_ENT ” para el atributo “Entidad”
- De la tabla “ ENTIDAD ” se toma “ DESC_MUN ” para el atributo “ Municipio ”
- De la tabla “TAM_LOC” se toma “DESC_TL” para el atributo “Tam de Localidad”
- De la tabla “ TIEMPO ” se toma “ AÑO ” para el atributo “Año”

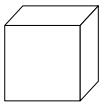
Creación de jerarquías.

- Se definió la Jerarquía Geográfica que se aplicará a los atributos Entidad y Municipio, en donde:

Una entidad tiene uno o más municipios.

Un Municipio pertenece a una sola entidad

El Cubo Multidimensional queda conformado por nueve atributos, siete indicadores y dos jerarquías (Figura 24).



Bienes de las Viviendas en México

- Entidad [ENTIDAD.DescEnt]
- Municipio [ENTIDAD.DescMun]
- Localidad [ENTIDAD.DescLoc]
- Tamaño de Localidad [TamLoc.DescTL]
- Año [TIEMPO.Año]
- Viviendas Mexicanas [Sum(VIVIENDA.TotViv)]
- Viviendas con Drenaje [Sum(VIVIENDA.VivDrenaje)]
- Viviendas con Electricidad [Sum(VIVIENDA.VivElect)]
- Viviendas con Televisión [Sum(VIVIENDA.VivTV)]
- Viviendas con Refrigerador [Sum(VIVIENDA.VivRefri)]
- Viviendas con Lavadora [Sum(VIVIENDA.VivLava)]
- Viviendas con Computadora [Sum(VIVIENDA.VivCompu)]

Jerarquía Geográfica

Entidad

Municipio

Figura 24; Cubo Multidimensional de Bienes de las Viviendas en México

Para la obtención del Cubo Servicios de Salud en México (Figura 24), se realizó la siguiente consulta:

```
SELECT  C.DESCENT, C.DESC_MUN, A.IDTL, A.IDPER, D.MES, D.ANIO,;
A.VIVS AS VT, A.VIVDRENAJE, A.VIVELECT,;
A.VIVTV, A.VIVREFRI, A.VIVLAVA, A.VIVCOMPU;
FROM VIVIENDA A;
INNER JOIN ENTIDAD C ON A.IDENTIDAD=C.IDENTIDAD;
INNER JOIN TIEMPO D ON A.IDPER=D.ID_PER;
INTO TABLE "CUBO_BIENES_VIVS"
```

3.3 Obtener Indicadores principales con la Metodología Propuesta

El proceso para la obtención de los indicadores principales con el Método Tradicional se compone de nueve pasos explicados con detalle en el Capítulo 2 (Procedimiento para el cálculo de tamaños de muestra de Encuestas Especiales en hogares). Sin embargo, algunos pasos de dicho Método Tradicional, pueden ser omitidos o cambiados (Tabla 10) debido a la Armonización de Variables realizada y al Data Mart Esquema Constelación que se construyó; por tal motivo se requiere de un nuevo método para la obtención de indicadores.

Método Propuesto:

1. Objetivo. Se define el objetivo con una descripción breve y concisa del propósito de la encuesta en términos de la información que se pretende obtener.
2. Población objeto de estudio. Definición clara y precisa del conjunto de entes o individuos de los que se pretende obtener información.
3. Cobertura. Límites geográficos o sectoriales del universo de estudio.
4. Elección de Dominios de interés. Se refiere a los subconjuntos del universo de estudio para los que se pretende obtener las estimaciones, basándose en la matriz de armonización de variables (Tabla 5).
5. Elección de Variables principales. Aquellas variables que son relevantes para la investigación y por tanto deben tomarse en como punto de referencia para fijar la precisión de las estimaciones de la encuesta, basándose en la matriz de armonización de variables (Tabla 5).
6. Programación. Creación de cubos necesarios para obtener los indicadores solicitados.
7. Procesamiento. Es el tiempo que tarda la computadora en realizar los cálculos hechos anteriormente por el programador.
8. Obtener indicadores principales. Obtención de Reportes con los indicadores solicitados.

	TRADICIONAL		PROPUESTO	
1	OBJETIVO	Se define el objetivo con una descripción breve y concisa del propósito de la encuesta en términos de la información que se pretende obtener.	OBJETIVO	Se define el objetivo con una descripción breve y concisa del propósito de la encuesta en términos de la información que se pretende obtener.
2	REVISIÓN	El muestrista y el programador revisan el cuestionario perteneciente a la encuesta para definir las variables que forman parte de ella y así definir las variables principales.		NO ES NECESARIO REVISAR EL CUESTIONARIO. LA ARMONIZACIÓN DE VARIABLES NOS PERMITE OMITIR ESTE PASO.
3	POBLACIÓN OBJETO DE ESTUDIO	Definición clara y precisa del conjunto de entes o individuos de los que se pretende obtener información.	POBLACIÓN OBJETO DE ESTUDIO	Definición clara y precisa del conjunto de entes o individuos de los que se pretende obtener información.
4	COBERTURA	Límites geográficos o sectoriales del universo de estudio.	COBERTURA	Límites geográficos o sectoriales del universo de estudio.
5	DOMINIOS DE INTERÉS	Se refiere a los subconjuntos del universo de estudio para los que se pretende obtener las estimaciones.	ELECCIÓN DE DOMINIOS DE INTERÉS	Se refiere a los subconjuntos del universo de estudio para los que se pretende obtener las estimaciones.
6	VARIABLES PRINCIPALES	Aquellas variables que son relevantes para la investigación y por tanto deben tomarse en como punto de referencia para fijar la precisión de las estimaciones de la encuesta.	ELECCIÓN DE VARIABLES PRINCIPALES	LAS VARIABLES YA ESTAN DEFINIDAS EN EL DATA MART ESQUEMA CONSTELACIÒN, POR LO TANTO SOLO DEBEN SER ELEGIDAS
7	PROGRAMACIÓN	El programador revisa las bases de datos necesarias para la obtención de las variables principales. Revisa la cobertura, el dominio de interés y la población objeto de estudio. Realiza la programación necesaria para obtener las variables principales y pasa la información al muestrista	PROGRAMACIÓN	EL DATA MART ESQUEMA CONSTELACIÓN CONTIENE LAS VARIABLES PRINCIPALES ARMONIZADAS, POR LO CUAL NO ES NECESARIA UNA REVISIÓN, SOLO REALIZA LA PROGRAMACIÓN NECESARIA PARA LA CREACIÓN DEL CUBO QUE SIRVA PARA OBTENER LOS INDICADORES SOLICITADOS.
8	PROCESAMIENTO	Es el tiempo que tarda la computadora en realizar los cálculos hechos anteriormente por el programador.	PROCESAMIENTO	Es el tiempo que tarda la computadora en realizar los cálculos hechos anteriormente por el programador.
9	OBTENER INDICADORES PRINCIPALES	Obtención de Reportes con los indicadores solicitados.	OBTENER INDICADORES PRINCIPALES	Obtención de Reportes con los indicadores solicitados.

Tabla 10; Comparativo de Métodos para la Obtención de Indicadores Principales

Para obtener los resultados finales con la Metodología Propuesta a partir de los cubos generados, se usó una aplicación en Excell que maneja tablas dinámicas y se obtuvieron reportes a diferentes niveles de detalle. (Tabla 10)

CUBO	No.	REPORTE	PROYECTO
POBLACIÓN Y VIVIENDAS EN MÉXICO	1	Reporte de Población y Viviendas por Entidad.	CENSO DE POBLACIÓN Y VIVIENDA 2000
	2	Reporte de Población por Entidad y Tamaño de Localidad.	
	3	Reporte de Viviendas por Entidad y Tamaño de Localidad.	
	4	Reporte de Viviendas de los Municipios de Aguascalientes, Baja California y Baja California Sur	
	5	Reporte de Población y Viviendas por Entidad.	CENSO DE POBLACIÓN Y VIVIENDA 2000 y CONTEO DE POBLACIÓN Y VIVIENDA 2005
SERVICIOS DE SALUD EN MÉXICO	1	Reporte de Servicios de Salud en el Estado de Aguascalientes	CONTEO DE POBLACIÓN Y VIVIENDA 2005
	7	Reporte de Población con Derechohabiencia por Tamaño de Localidad en el Estado de Aguascalientes	
	8	Reporte de Población Con Derechohabiencia y Sin Derechohabiencia por Género en el Estado de Aguascalientes .	
BIENES DE LAS VIVIENDAS EN MÉXICO	11	Reporte de Bienes de las Viviendas de los Municipios de Aguascalientes	CENSO DE POBLACIÓN Y VIVIENDA 2000
	12	Reporte de Viviendas con Drenaje por Tamaño de Localidad en los Municipios de Aguascalientes .	
	13	Reporte de Viviendas con Computadora en los Municipios de Aguascalientes .	

Tabla 11; Reportes Generados.

REPORTE DE POBLACIÓN Y VIVIENDAS POR ENTIDAD
DEL CENSO DE POBLACIÓN Y VIVIENDA 2000

ENTIDAD	POBLACIÓN	VIVIENDAS
Aguascalientes	944,285	200,673
Baja California	2,487,367	610,057
Baja California Sur	424,041	105,229
Campeche	690,689	157,172
Chiapas	3,920,892	806,551
Chihuahua	3,052,907	755,959
Coahuila	2,298,070	544,660
Colima	542,627	132,330
Distrito Federal	8,605,239	2,132,413
Durango	1,448,661	325,309
Guanajuato	4,663,032	926,284
Guerrero	3,079,649	657,989
Hidalgo	2,235,591	494,317
Jalisco	6,322,002	1,394,026
Mexico	13,096,686	2,893,357
Michoacan	3,985,667	855,512
Morelos	1,555,296	367,399
Nayarit	920,185	220,118
Nuevo Leon	3,834,141	888,552
Oaxaca	3,438,765	741,005
Puebla	5,076,686	1,065,882
Queretaro	1,404,306	298,372
Quintana Roo	874,963	213,566
San Luis Potosi	2,299,360	492,914
Sinaloa	2,536,844	575,292
Sonora	2,216,969	530,435
Tabasco	1,891,829	412,634
Tamaulipas	2,753,222	683,068
Tlaxcala	962,646	194,549
Veracruz	6,908,975	1,606,194
Yucatan	1,658,210	373,432
Zacatecas	1,353,610	299,483
Total general	97,483,412	21,954,733

Tabla 12; Reporte No. 1 del Cubo "Población y Viviendas en México"

REPORTE DE POBLACIÓN POR ENTIDAD Y TAMAÑO DE LOCALIDAD

DEL CENSO DE POBLACIÓN Y VIVIENDA 2000

Tamaño de Localidad					
Entidad	1	2	3	4	Total general
Aguascalientes	594,092	93,895	69,592	186,706	944,285
Baja California	1,922,046	166,634	186,069	212,618	2,487,367
Baja California Sur	162,954	104,675	77,106	79,306	424,041
Campeche	316,837	48,946	124,526	200,380	690,689
Chiapas	716,860	404,829	655,705	2,143,498	3,920,892
Chihuahua	1,845,151	457,257	217,039	533,460	3,052,907
Coahuila	1,492,650	444,166	117,937	243,317	2,298,070
Colima	119,639	261,062	83,737	78,189	542,627
Distrito Federal	8,391,517	113,231	80,171	20,320	8,605,239
Durango	637,248	100,643	186,164	524,606	1,448,661
Guanajuato	1,754,716	967,175	409,349	1,531,792	4,663,032
Guerrero	868,161	332,861	496,931	1,381,696	3,079,649
Hidalgo	231,602	452,408	418,684	1,132,897	2,235,591
Jalisco	3,482,257	1,028,826	828,282	982,637	6,322,002
Mexico	8,287,770	1,264,191	1,680,931	1,863,794	13,096,686
Michoacan	898,693	851,745	853,746	1,381,483	3,985,667
Morelos	606,553	317,159	405,010	226,574	1,555,296
Nayarit	265,817	119,013	205,598	329,757	920,185
Nuevo Leon	3,130,475	273,174	177,722	252,770	3,834,141
Oaxaca	251,846	520,730	759,597	1,906,592	3,438,765
Puebla	1,476,271	779,436	1,212,500	1,608,479	5,076,686
Queretaro	536,463	179,380	230,510	457,953	1,404,306
Quintana Roo	518,793	121,383	81,362	153,425	874,963
San Luis Potosi	904,503	180,795	269,376	944,686	2,299,360
Sinaloa	1,069,721	265,908	369,340	831,875	2,536,844
Sonora	1,080,217	512,363	249,537	374,852	2,216,969
Tabasco	330,846	316,623	347,152	897,208	1,891,829
Tamaulipas	1,815,621	348,929	187,379	401,293	2,753,222
Tlaxcala		371,511	383,752	207,383	962,646
Veracruz	1,647,517	1,233,332	1,193,647	2,834,479	6,908,975
Yucatan	662,530	313,286	373,401	308,993	1,658,210
Zacatecas	113,947	340,334	267,783	631,546	1,353,610
Total general	46,133,313	13,285,900	13,199,635	24,864,564	97,483,412

Tabla 13; Reporte No. 2 del Cubo "Población y Viviendas en México"

REPORTE DE VIVIENDAS POR ENTIDAD Y TAMAÑO DE LOCALIDAD
DEL CENSO DE POBLACIÓN Y VIVIENDA 2000

Tamaño de Localidad					
Entidad	1	2	3	4	Total general
Aguascalientes	132,382	18,499	13,772	36,020	200,673
Baja California	473,842	39,848	44,124	52,243	610,057
Baja California Sur	40,193	26,636	18,646	19,754	105,229
Campeche	78,339	11,199	26,615	41,019	157,172
Chiapas	166,398	90,506	139,398	410,249	806,551
Chihuahua	456,043	115,059	54,770	130,087	755,959
Coahuila	353,173	105,783	28,072	57,632	544,660
Colima	29,750	63,863	19,751	18,966	132,330
Distrito Federal	2,085,354	25,076	17,386	4,597	2,132,413
Durango	148,271	22,494	41,712	112,832	325,309
Guanajuato	354,476	192,261	81,352	298,195	926,284
Guerrero	204,408	73,606	106,474	273,501	657,989
Hidalgo	56,724	104,301	91,757	241,535	494,317
Jalisco	774,404	223,529	183,314	212,779	1,394,026
Mexico	1,899,672	271,892	345,425	376,368	2,893,357
Michoacan	201,157	185,654	181,144	287,557	855,512
Morelos	150,332	73,627	92,704	50,736	367,399
Nayarit	64,227	29,322	50,705	75,864	220,118
Nuevo Leon	712,365	68,424	43,301	64,462	888,552
Oaxaca	59,701	120,112	164,134	397,058	741,005
Puebla	346,054	163,037	237,778	319,013	1,065,882
Queretaro	121,967	40,026	46,086	90,293	298,372
Quintana Roo	131,415	30,791	18,301	33,059	213,566
San Luis Potosi	206,577	39,797	58,048	188,492	492,914
Sinaloa	252,531	59,457	82,788	180,516	575,292
Sonora	262,163	123,833	57,500	86,939	530,435
Tabasco	81,999	74,147	74,404	182,084	412,634
Tamaulipas	451,051	86,552	46,882	98,583	683,068
Tlaxcala		76,675	76,626	41,248	194,549
Veracruz	428,718	299,204	275,404	602,868	1,606,194
Yucatan	163,751	66,568	79,624	63,489	373,432
Zacatecas	26,309	76,292	58,523	138,359	299,483
Total general	10,913,746	2,998,070	2,856,520	5,186,397	21,954,733

Tabla 14; Reporte No. 3 del Cubo "Población y Viviendas en México"

REPORTE DE VIVIENDAS DE LOS MUNICIPIOS DE

AGUASCALIENTES,
BAJA CALIFORNIA Y BAJA CALIFORNIA SUR
DEL CENSO DE POBLACIÓN Y VIVIENDA 2000

Entidad	Municipio	Viviendas
Aguascalientes	AGUASCALIENTES	141,784
	ASIENTOS	7,346
	CALVILLO	10,622
	COSIO	2,454
	JESUS MARIA	12,377
	LLANO, EL	3,009
	PABELLON DE ARTEAGA	6,692
	RINCON DE ROMOS	7,903
	SAN FRANCISCO DE LOS ROMO	3,859
	SAN JOSE DE GRACIA	1,462
	TEPEZALA	3,165
Total Aguascalientes		200,673
Baja California	ENSENADA	92,336
	MEXICALI	190,426
	PLAYAS DE ROSARITO	15,493
	TECATE	19,020
	TIJUANA	292,782
Total Baja California		610,057
Baja California Sur	CABOS, LOS	26,983
	COMONDU	15,482
	LORETO	2,873
	MULEGE	11,531
	PAZ, LA	48,360
Total Baja California Sur		105,229

Tabla 15; Reporte No. 4 del Cubo "Población y Viviendas en México"

	CENSO 2000		CONTEO 2005	
ENTIDAD	POBLACIÓN	VIVIENDAS	POBLACIÓN	VIVIENDAS
Aguascalientes	944,285	200,673	1,065,416	245,795
Baja California	2,487,367	610,057	2,844,469	738,907
Baja California Sur	424,041	105,229	512,170	136,059
Campeche	690,689	157,172	754,730	184,231
Chiapas	3,920,892	806,551	2,495,200	625,465
Chihuahua	3,052,907	755,959	567,996	149,385
Coahuila	2,298,070	544,660	4,293,459	916,884
Colima	542,627	132,330	3,241,444	853,210
Distrito Federal	8,605,239	2,132,413	8,720,916	2,288,633
Durango	1,448,661	325,309	1,509,117	358,460
Guanajuato	4,663,032	926,284	4,893,812	1,049,234
Guerrero	3,079,649	657,989	3,115,202	702,130
Hidalgo	2,235,591	494,317	2,345,514	558,685
Jalisco	6,322,002	1,394,026	6,752,113	1,583,293
Mexico	13,096,686	2,893,357	14,007,495	3,244,466
Michoacan	3,985,667	855,512	3,966,073	914,103
Morelos	1,555,296	367,399	1,612,899	403,288
Nayarit	920,185	220,118	949,684	244,689
Nuevo Leon	3,834,141	888,552	4,199,292	1,014,452
Oaxaca	3,438,765	741,005	3,506,821	803,355
Puebla	5,076,686	1,065,882	5,383,133	1,207,930
Queretaro	1,404,306	298,372	1,598,139	360,222
Quintana Roo	874,963	213,566	1,135,309	286,019
San Luis Potosi	2,299,360	492,914	2,410,414	557,864
Sinaloa	2,536,844	575,292	2,608,442	642,325
Sonora	2,216,969	530,435	2,394,861	615,002
Tabasco	1,891,829	412,634	1,989,969	473,308
Tamaulipas	2,753,222	683,068	3,024,238	789,466
Tlaxcala	962,646	194,549	1,068,207	233,965
Veracruz	6,908,975	1,606,194	7,110,214	1,778,551
Yucatan	1,658,210	373,432	1,818,948	435,578
Zacatecas	1,353,610	299,483	1,367,692	325,416
Total general	97,483,412	21,954,733	103,263,388	24,720,370

Tabla 16; Reporte No. 5 del Cubo "Población y Viviendas en México"

Reporte de Servicios de Salud en el Estado de Aguascalientes

MUNICIPIOS	PT	CON DERECHOHAB.	IMSS	ISSSTE	PEMEX	OTRA INST.	SIN DERECHOHAB.
AGUASCALIENTES	643,419	410,312	354,448	53,305	1,876	683	228,856
ASIENTOS	37,763	12,085	10,496	1,557	20	12	25,381
CALVILLO	51,291	8,470	6,344	1,950	169	7	42,301
COSIO	12,619	4,912	3,921	985	3	3	7,629
JESUS MARIA	64,097	33,442	31,505	1,893	30	14	30,104
LLANO, EL	15,327	5,897	5,203	679	14	1	9,317
PABELLON DE ARTEAGA	34,296	17,687	13,200	4,462	22	3	16,329
RINCON DE ROMOS	41,655	17,938	13,564	4,308	43	23	23,558
SAN FRANCISCO DE LOS ROMO	20,066	9,656	9,036	612	5	3	10,220
SAN JOSE DE GRACIA	7,244	2,139	1,260	874	5	0	5,017
TEPEZALA	16,508	5,298	4,488	786	22	2	11,115
Total general	944,285	527,836	453,465	71,411	2,209	751	409,827

Tabla 17; Reporte No. 1 del Cubo "Servicios de Salud en México"

Reporte de Población con Derechohabiencia
por Tamaño de Localidad en el Estado de Aguascalientes

	TAMAÑO DE LOCALIDAD				
MUNICIPIOS	1	2	3	4	TOTAL GENERAL
AGUASCALIENTES	388,007		3,504	18,801	410,312
ASIENTOS			5,015	7,070	12,085
CALVILLO		4,248	961	3,261	8,470
COSIO			2,220	2,692	4,912
JESUS MARIA		14,668	5,241	13,533	33,442
LLANO, EL			2,009	3,888	5,897
PABELLON DE ARTEAGA		13,713	1,266	2,708	17,687
RINCON DE ROMOS		11,376	2,734	3,828	17,938
SAN FRANCISCO DE LOS ROMO			4,874	4,782	9,656
SAN JOSE DE GRACIA			1,662	477	2,139
TEPEZALA			2,723	2,575	5,298
Total general	388,007	44,005	32,209	63,615	527,836

Tabla 18; Reporte No. 2 del Cubo "Servicios de Salud en México"

Reporte de Población Con Derechohabiencia y Sin Derechohabiencia por Género en el Estado de Aguascalientes

	HOMBRE			MUJER		
MUNICIPIOS	CON DERECHOHAB.	SIN DERECHOHAB.	TOTAL	CON DERECHOHAB	SIN DERECHOHAB.	TOTAL
AGUASCALIENTES	196190	112171	308361	214122	116685	330807
ASIENTOS	5705	12531	18236	6380	12850	19230
CALVILLO	3818	20365	24183	4652	21936	26588
COSIO	2354	3737	6091	2558	3892	6450
JESUS MARIA	16131	14788	30919	17311	15316	32627
LLANO, EL	2811	4762	7573	3086	4555	7641
PABELLON DE ARTEAGA	8346	8060	16406	9341	8269	17610
RINCON DE ROMOS	8456	11738	20194	9482	11820	21302
SAN FRANCISCO DE LOS ROMO	4597	5020	9617	5059	5200	10259
SAN JOSE DE GRACIA	983	2428	3411	1156	2589	3745
TEPEZALA	2533	5412	7945	2765	5703	8468
Total general	251924	201012	452936	275912	208815	484727

Tabla 19; Reporte No. 3 del Cubo "Servicios de Salud en México"

Reporte de Bienes de las Viviendas de los Municipios de Aguascalientes

MUNICIPIO	TOTAL VIVS	DRENAJE	ELECTRICIDAD	TELEVISIÓN	REFRIGERADOR	LAVADORA	COMPUTADORA
AGUASCALIENTES	141,784	137,075	138,978	136,766	123,740	113,574	23,715
ASIENTOS	7,346	5,782	6,856	6,548	4,209	4,506	99
CALVILLO	10,622	9,882	10,242	9,690	7,786	6,785	263
COSIO	2,454	2,239	2,371	2,257	1,458	1,406	67
JESUS MARIA	12,377	11,448	11,901	11,670	9,395	8,816	648
LLANO, EL	3,009	2,253	2,744	2,678	1,594	1,573	27
PABELLON DE ARTEAGA	6,692	6,253	6,481	6,329	4,745	4,796	445
RINCON DE ROMOS	7,903	7,115	7,573	7,396	5,301	4,998	441
SAN FRANCISCO DE LOS ROMO	3,859	3,618	3,757	3,650	2,826	2,852	127
SAN JOSE DE GRACIA	1,462	1,034	1,361	1,236	809	688	34
TEPEZALA	3,165	2,627	3,020	2,911	1,828	1,960	51
Total general	200,673	189,326	195,284	191,131	163,691	151,954	25,917

Tabla 20; Reporte No. 1 del Cubo "Bienes de las Viviendas en México"

Reporte de Viviendas con Drenaje por Tamaño de Localidad en los Municipios de Aguascalientes

MUNICIPIOS	1	2	3	4	TOTAL GENERAL
AGUASCALIENTES	130,202		1,145	5,728	137,075
ASIENTOS			1,850	3,932	5,782
CALVILLO		3,662	1,303	4,917	9,882
COSIO			836	1,403	2,239
JESUS MARIA		5,407	1,500	4,541	11,448
LLANO, EL			752	1,501	2,253
PABELLON DE ARTEAGA		4,641	457	1,155	6,253
RINCON DE ROMOS		4,293	1,087	1,735	7,115
SAN FRANCISCO DE LOS ROMO			2,062	1,556	3,618
SAN JOSE DE GRACIA			681	353	1,034
TEPEZALA			1,039	1,588	2,627
Total general	130,202	18,003	12,712	28,409	189,326

Tabla 21; Reporte No. 2 del Cubo "Bienes de las Viviendas en México"

Reporte de Viviendas con Computadora en los Municipios de Aguascalientes

MUNICIPIOS	TOTAL VIVS	COMPUTADORA
AGUASCALIENTES	141,784	23,715
ASIENTOS	7,346	99
CALVILLO	10,622	263
COSIO	2,454	67
JESUS MARIA	12,377	648
LLANO, EL	3,009	27
PABELLON DE ARTEAGA	6,692	445
RINCON DE ROMOS	7,903	441
SAN FRANCISCO DE LOS ROMO	3,859	127
SAN JOSE DE GRACIA	1,462	34
TEPEZALA	3,165	51
Total general	200,673	25,917

Tabla 22; Reporte No. 3 del Cubo "Bienes de las Viviendas en México"

Capítulo IV. Conclusiones

4.1 Respuesta a preguntas y proposiciones

Pregunta 1. ¿Es posible realizar un Data Mart con esquema Constelación con variables armonizadas de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2000?

Para hacer posible esto, se realizaron dos procesos principales:

1. Armonizar variables. Como resultados principales de la armonización de variables se obtuvo:

   - Matriz de períodos de Referencia. Nos permitió la comparabilidad en el tiempo.
   - Matriz de comparabilidad de variables principales. Nos permitió vincular las variables principales de los proyectos que estamos estudiando con el proyecto IPUMS-International.
   - Matriz de armonización de variables. Nos permitió la armonización de variables entre proyectos y dominios de estudio.
   - Base de datos armonizada en Oracle. A través de dicha BD se tuvo la fuente idónea para la realización del Data Mart, Esquema Constelación.

2. Crear Data Mart. Para crear el Data Mart Esquema Constelación se siguió la Metodología HEFESTO y se tomó como fuente la Base de Datos armonizada en Oracle.
   En el proceso de creación se obtuvo:

   - Modelo conceptual
   - Modelo conceptual ampliado
   - Tablas de Dimensión: Entidad, Tamaño de Localidad, Tiempo, Edad, Sexo.
   - Tablas de Hechos: Población y Vivienda

Y como resultado final de la Creación del Data Mart se obtuvieron:

- Esquema Constelación (Figura 21)
- Cubos Multidimensionales:
  - " Población y Viviendas en México " (Figura 22)
  - " Servicios de Salud en México " (Figura 23)
  - " Bienes de las Viviendas en México " (Figura 24)

Por lo tanto, sí fue posible realizar el Data Mart Esquema Constelación con variables armonizadas para los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2000.

Pregunta 2. ¿Cuántos procesos intervienen en la metodología tradicional y cuántos procesos intervienen en la metodología propuesta usando el Data Mart con Esquema Constelación para obtener indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005?

En el método tradicional (Tabla 23) se tienen nueve procesos, sin embargo la descripción de este proceso es por proyecto, como se explica posteriormente.

Número	Proceso	Personas involucradas
1	Objetivo	Muestrista
2	Revisión	Muestrista y Programador
3	Población objeto de estudio	Muestrista y Programador
4	Cobertura	Muestrista y Programador
5	Dominios de interés	Muestrista y Programador
6	Variables principales	Muestrista y Programador
7	Programación	Programador
8	Procesamiento	PC
9	Obtener indicadores	Programador

Tabla 23; Método tradicional.

En la metodología propuesta (Tabla 24) intervienen ocho procesos, sin embargo, con realizar una sola vez este proceso se obtendrían los indicadores principales válidos para los tres proyectos a analizar.

Número	Proceso	Personas involucradas
1	Objetivo	Muestrista
2	Elección de Población objeto de estudio	Muestrista
3	Elección de Cobertura	Muestrista
4	Elección de Dominios de interés	Muestrista
5	Elección de Variables principales	Muestrista
6	Programación	Programador
7	Procesamiento	PC
8	Obtener indicadores	Programador

Tabla 24; Método propuesto.

Ejemplo propuesto.

Tomando como ejemplo la obtención de indicadores necesarios para la realización del Reporte No. 5 del Cubo “Población y Viviendas en México” (Tabla 16), en el cuál se están analizando dos proyectos: Censo de PyV 2000 y Conteo de PyV 2005.

El método tradicional se debería realizar dos veces; una vez por cada proyecto y los procesos que intervienen serían:

Método Tradicional = 9 PROCESOS x 2 PROYECTOS = 18 PROCESOS

El método propuesto incluye los dos proyectos analizados, por lo tanto.

Método Propuesto = 8 PROCESOS

Pregunta 3. ¿Cuántas personas se necesitan para obtener los indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 con la metodología tradicional y con la metodología propuesta?

En el método tradicional (Tabla 23) se necesitan dos personas para la obtención de los indicadores principales y comparables por cada proyecto, es decir, continuando con el ejemplo del Reporte No. 5 del Cubo "Población y Viviendas en México" (Tabla 16), las personas involucradas serían seis:

- Censo de PyV 2000: Un muestrista y un programador.
- Conteo de PyV 2005: Un muestrista y un programador.
- ENOE del año 2005: Un muestrista y un programador.

En el método propuesto (Tabla 24) se necesitan dos personas para la obtención de los indicadores principales y comparables por cada proyecto, pero con la diferencia que la armonización de las variables de dichos proyectos, nos permite realizar el método sólo una ocasión y se obtienen resultados para los tres proyectos. Por lo tanto, las personas involucradas serían dos:

- Proyectos: Un muestrista y un programador.

Pregunta 4. ¿Cuántas tablas se necesitan acceder para obtener los indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 con la metodología tradicional y con la metodología propuesta?

El número de tablas que se necesitan acceder para obtener los indicadores principales varía dependiendo del ejercicio que se esté realizando, por tal motivo se tomará como ejemplo la obtención de los indicadores de Población y Viviendas de los proyectos planteados, mismos que se encuentran plasmados en el Reporte No. 5 del Cubo "Población y Viviendas en México" (Tabla 16)

Con el método tradicional, se accedería a:

- 32 tablas para el proyecto “Censo de PyV 2000”
- 32 tablas para el proyecto “Conteo de PyV 2005”
- 4 tablas para el proyecto “ENOE 2005”, una para cada trimestre del año 2005
- 1 tabla de catálogo de Entidades

TOTAL: 69 TABLAS A ACCEDER.

Con el método propuesto, se accede a:

- Tabla de Hechos POBLACIÓN
- Tabla de Hechos VIVIENDA
- Tabla de Dimensión ENTIDAD
- Tabla de Dimensión TIEMPO

TOTAL: 4 TABLAS A ACCEDER.

Pregunta 5. ¿Se obtiene una disminución de tiempo de respuesta para la obtención de indicadores principales de los proyectos ENOE del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 a través de la Metodología Propuesta que hace uso del Data Mart con Esquema Constelación?

Para dar respuesta a esta pregunta, se realiza una comparación entre ambos métodos (tradicional y propuesto) a través del ejercicio que hemos trabajado y dio por resultado el Reporte No. 5 del Cubo “Población y Viviendas en México” (Tabla 16). Los resultados obtenidos fueron:

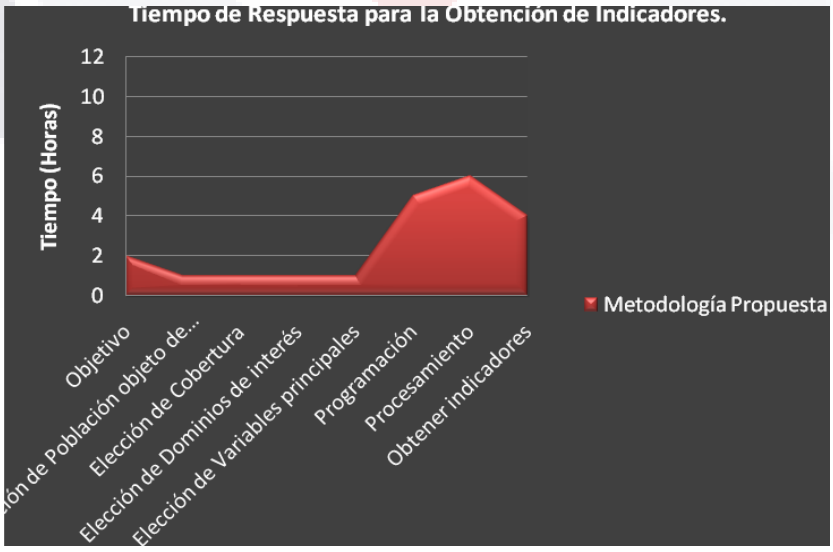
Metodología Tradicional		**Metodología Propuesta**	
Proceso	**Tiempo (Horas)**	**Proceso**	**Tiempo (Horas)**
Objetivo	2	Objetivo	2
Revisión	8		
Población objeto de estudio	1	Elección de Población objeto de estudio	1
Cobertura	1	Elección de Cobertura	1
Dominios de interés	2	Elección de Dominios de interés	1
Variables principales	3	Elección de Variables principales	1
Programación	10	Programación	5
Procesamiento	9	Procesamiento	6
Obtener indicadores	4	Obtener indicadores	4
Total de horas	40	Total de horas	21

Tabla 25; Esquema de Selección Balanceado, Método Tradicional y Propuesto

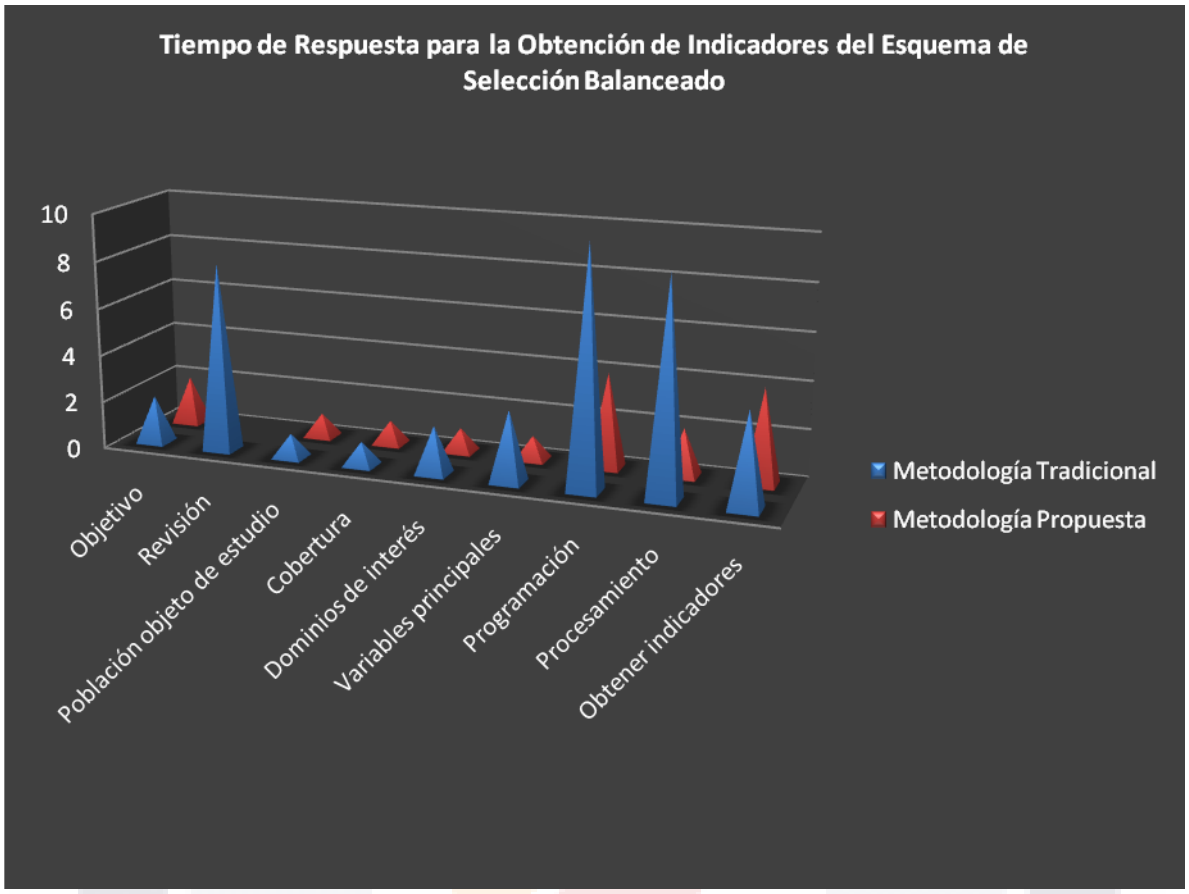
Se puede observar que efectivamente se disminuyó el tiempo de respuesta para la obtención de indicadores principales de los proyectos ENOE del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 a través del Data Mart con esquema Constelación.

Se realizaron una serie de gráficas que representan los tiempos de respuesta para ambas metodologías (Gráfica 1 y 2); así como un comparativo entre ambas gráficas (Gráfica 3) y el total de horas para cada metodología (Gráfica 4).

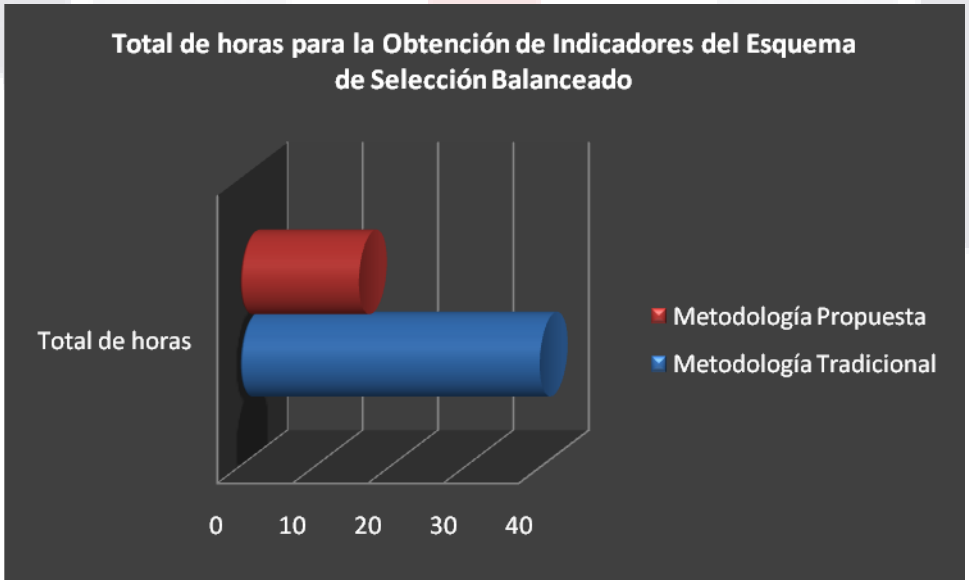
Gráfica 1; Tiempo de Respuesta para la Obtenciòn de Indicadores: Método Tradicional



Gráfica 2; Tiempo de Respuesta para la Obtenciòn de Indicadores: Método Propuesto



Gráfica 3; Comparativo de Tiempos de Respuesta para la Obtención de Indicadores



Gráfica 4; Comparativo del Total de horas para la Obtención de Indicadores.

4.2 Logro de objetivos propuestos

OBJETIVO GENERAL. Generar el prototipo de un Data Mart con esquema Constelación de proyectos con variables armonizadas para la obtención de sus indicadores principales y comparables.

Se logró satisfactoriamente el Objetivo General a través de la construcción del Data Mart con Esquema Constelación (Figura 21), a partir del cual se pudieron generar los indicadores principales y obteniendo resultados comparables como los mostrados en el Reporte No. 5 del Cubo “Población y Viviendas en México” (Tabla 16)

OE1. Determinar si es posible realizar un Data Mart esquema Constelación con variables armonizadas de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

Para lograr el OE1 satisfactoriamente se realizó la armonización de variables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 y posteriormente se construyó el Data Mart Esquema Constelación para dichos proyectos, tal como se muestra en la Figura 21 de este documento.

OE2. Facilitar la obtención de indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005.

El OE2 se cumplió satisfactoriamente, pues se facilitó la obtención de indicadores principales y comparables de los proyectos ENOE a partir del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 a través de:

El número de procesos en el ejemplo de la Pregunta 2.
Método Tradicional = 18 procesos - Método Propuesto = 8 procesos
DISMINUCIÓN = 10 Procesos.

El número de personal involucrado en el ejemplo de la Pregunta 3.
Método Tradicional = 6 personas - Método Propuesto = 2 personas
DISMINUCIÓN = 4 Personas.

El número de tablas a acceder en el ejemplo de la Pregunta 4.
Método Tradicional = 69 tablas - Método Propuesto = 4 tablas
DISMINUCIÓN = 65 Tablas.

OE3. Comprobar la disminución del tiempo de respuesta para la obtención de indicadores principales de los proyectos ENOE del año 2005, Censo de Población y Vivienda 2000 y Conteo de Población y Vivienda 2005 a través de la Metodología Propuesta que hace uso del Data Mart con Esquema Constelación.

Se cumplió satisfactoriamente con el OE3, pues se realizó el ejercicio de Esquema de Selección Balanceado con ambas metodologías y se obtuvo una disminución de 19 horas de trabajo (Tabla 25 y Gráfica 4). Además se realizaron una serie de gráficas en donde se observa que la mayor parte de la disminución del tiempo se da en los procesos de Revisión, Programación y Procesamiento (Gráfica 3) por lo cual se concluye que la disminución en la revisión se debe a la armonización de variables realizada y la disminución en la programación y el procesamiento se debe tanto a la armonización como a la construcción del Data Mart Esquema Constelación.

4.3 Áreas del conocimiento utilizadas

Dentro de la Maestría en Informática y Tecnologías Computacionales se trataron diversas áreas del conocimiento informático, las cuáles fueron base y guía fundamental para el desarrollo de este trabajo, entre ellas se encuentran:

Programación e Ingeniería de Software.
En donde se incluyen las metodologías para construir programas y sistemas informáticos, considerando su análisis y diseño, mantenimiento y otros aspectos relacionados.

Tratamiento de Información.
Área del conocimiento que incluye el manejo de información y tópicos tales como:

- Base de datos.
- Data Warehouse.
- Modelado y Diseño.
- Lenguajes de consulta: SQL.

Por otra parte se tocaron temas relacionadas con Estadística, como lo es el Cálculo de Tamaños de Muestra de Encuestas.

Además de lo mencionado anteriormente, se hizo uso de las teorías de armonización de variables; expuestas fundamentalmente dentro del proyecto IPUMS – International en donde se incluye la recopilación de información, determinar disponibilidad de variables, armonizar variables a través de un esquema de codificación.

Capítulo V. Recomendaciones.

Para trabajos futuros las recomendaciones son las siguientes:

1. La armonización de variables siguiendo el método planteado por IPUMS es válido al cien por ciento; sin embargo si el requerimiento es armonizar variables de Encuestas pequeñas que incluyen variables las cuales no se incluyen dentro la codificación IPUMS, la recomendación es seguir el método IPUMS y crear la codificación inexistente sólo para las variables que así lo requieran.

2. El Data Mart con Esquema Constelación es un trabajo que se debe seguir alimentando con al mayor número de variables principales posibles, esto con el fin de poder aumentar nuestro Data Mart con nuevas encuestas de manera que se mantenga al día y actualizado, también se pueden encontrar nuevas formas y necesidades de consulta.

3. El manejo de los cubos multidimensionales y su aplicación a través de Excel puede ser mejorada con aplicaciones web dedicadas exclusivamente para dicho fin.

ANEXO.

VARIABLES DEL CENSO DE POBLACIÓN Y VIVIENDA 2000

Procedure VARS_00

** P O B L A D O R

```
Select Strtran(Str(Entidad,2),' ','0') As Ent,;
Municipio As Mun,Localidad As Loc,Ageb,Manzana As Mza,;
Strtran(Str(Id_viviend,10),' ','0') As id_viv,;
Strtran(Str(Id_hogar,3),' ','0') As id_hog,;
Strtran(Str(Id_poblado,6),' ','0') As id_pob,;
zona,unidad_pri As upm, '  ' As AM,;
' ' As TL, ' ' As EstNal, ' ' As EstEst,;
'001' as Id_per,;
edad As AGE,;
SUM(Iif(Sexo='1',1,0)) As SEX1,;
SUM(Iif(Sexo='2',1,0)) As SEX2,;
SUM(Iif(Estado_con='8',1,0)) As MARST1,;
SUM(Iif(Inlist(Estado_con,'1','5','6','7'),1,0)) As MARST2,;
SUM(Iif(Inlist(Estado_con,'2','3'),1,0)) As MARST3,;
SUM(Iif(Estado_con='4',1,0)) As MARST4,;
SUM(Iif(Condicion_='1',1,0)) As HLTHCOV10,;
SUM(Iif(Condicion2='2',1,0)) As HLTHCOV20,;
SUM(Iif(Condicion3='3',1,0)) As HLTHCOV30,;
SUM(Iif(Cond_derha='4',1,0)) As HLTHCOV40,;
SUM(Iif(No_tiene_d='5',1,0)) As HLTHCOV60,;
SUM(Iif(Condicion4='1',1,0)) As DISUPPR,;
SUM(Iif(Condicion5='2',1,0)) As DISMOBL,;
SUM(Iif(Condicion6='3',1,0)) As DISDEAF,;
SUM(Iif(Condicion7='4',1,0)) As DISMUTE,;
SUM(Iif(Condicion8='5',1,0)) As DISBLND,;
SUM(Iif(Condicion9='6',1,0)) As DISMNTL,;
SUM(Iif(Habla_leng='1',1,0)) As SPKIND,;
SUM(Iif(Alfabetism='1',1,0)) As LIT,;
SUM(Iif(Asistencia='1',1,0)) As SCHOOL;
FROM (wtr_pob);
GROUP By Ent,Mun,Loc,Ageb,Mza,id_viv,id_hog,id_pob,zona,upm,AGE;
Into Cursor C1
```

```
** V I V I E N D A
        Select Strtran(Str(Ent,2),' ','0') As Ent,;
        Mun,Loc,Ageb,Mza,'  ' As AM,;
        Strtran(Str(id_viv,10),' ','0') As id_viv,;
        zona,unidad_pri As upm,' ' As TL,;
        ' ' As EstNal, ' ' As EstEst,;
        '001' as Id_per,;
        SUM(Iif(Material_p='1',1,0)) As WALL201,;
        SUM(Iif(Material_p='2',1,0)) As WALL204,;
        SUM(Iif(Material_p='3',1,0)) As WALL546,;
        SUM(Iif(Material_p='4',1,0)) As WALL405,;
        SUM(Iif(Material_p='5',1,0)) As WALL532,;
        SUM(Iif(Material_p='6',1,0)) As WALL300,;
        SUM(Iif(Material_p='7',1,0)) As WALL523,;
        SUM(Iif(Material_p='8',1,0)) As WALL501,;
        SUM(Iif(Material_t='1',1,0)) As ROOF64,;
        SUM(Iif(Material_t='2',1,0)) As ROOF65,;
        SUM(Iif(Material_t='3',1,0)) As ROOF34,;
        SUM(Iif(Material_t='4',1,0)) As ROOF40,;
        SUM(Iif(Material_t='5',1,0)) As ROOF12,;
        SUM(Iif(Material_t='6',1,0)) As ROOF10,;
        SUM(Iif(Material_2='1',1,0)) As FLOOR100,;
        SUM(Iif(Material_2='2',1,0)) As FLOOR202,;
        SUM(Iif(Material_2='3',1,0)) As FLOOR231,;
        SUM(Iif(Tiene_coci='1',1,0)) As KITCHEN20,;
        SUM(Iif(Tiene_coci='2',1,0)) As KITCHEN10,;
        SUM(Iif(Dispone_el='1',1,0)) As ELECTRC20,;
        SUM(Iif(Dispone_el='2',1,0)) As ELECTRC10,;
        SUM(Iif(Dispone_te='3',1,0)) As TV20,;
        SUM(Iif(Dispone_te='4',1,0)) As TV10,;
        SUM(Iif(Dispone_re='1',1,0)) As REFRIG20,;
        SUM(Iif(Dispone_re='2',1,0)) As REFRIG10,;
        SUM(Iif(Dispone_la='3',1,0)) As WASHER20,;
        SUM(Iif(Dispone_la='4',1,0)) As WASHER10,;
        SUM(Iif(Dispone_co='3',1,0)) As COMPUTR20,;
        SUM(Iif(Dispone_co='4',1,0)) As COMPUTR10,;
        SUM(Iif(Dispone_sa='1',1,0)) As TOILET20,;
        SUM(Iif(Dispone_sa='2',1,0)) As TOILET10,;
        SUM(Iif(Inlist(Dispone_dr,'1','2','3','4'),1,0)) As SEWAGE20,;
        SUM(Iif(Dispone_dr='5',1,0)) As SEWAGE10,;
        SUM(Iif(Inlist(Dispone_ag,'1','2','3','4'),1,0)) As WATSUP11,;
        SUM(Iif(Inlist(Dispone_ag,'5','6'),1,0)) As WATSUP20,;
        SUM(Iif(Dispone_t2='5',1,0)) As PHONE20,;
        SUM(Iif(Dispone_t2='6',1,0)) As PHONE10,;
        SUM(Iif(Dispone_au='1',1,0)) As AUTOS20,;
        SUM(Iif(Dispone_au='2',1,0)) As AUTOS10;
        FROM (wtr_viv);
        GROUP By Ent,Mun,Loc,Ageb,Mza,id_viv,zona,upm,tamano_loc;
        into CURSOR C2
Return
```

VARIABLES DEL CONTEO DE POBLACIÓN Y VIVIENDA 2000

```
Procedure VARS_05
** P O B L A D O R
        Select Strtran(Str(Entidad,2),' ','0') As Ent,;
        Strtran(Str(Municipio,3),' ','0') As Mun,;
        Strtran(Str(Localidad,4),' ','0') As Loc,;
        Ageb,Strtran(Str(Manzana,3),' ','0') As Mza,;
        Strtran(Str(Consecutiv,10),' ','0') As id_viv,;
        Strtran(Str(Consecuti2,3),' ','0') As id_hog,;
        Strtran(Str(Consecuti3,6),' ','0') As id_pob,;
        '  ' As Zona, '     ' As UPM, '  ' As AM,;
        ' ' As TL, ' ' As EstNal, ' ' As EstEst,;
        '002' As Id_Per,;
        Strtran(Str(Edad,3),' ','0') As AGE,;
        SUM(Iif(Sexo=1,1,0)) As SEX1,;
        SUM(Iif(Sexo=2,1,0)) As SEX2,;
        SUM(Iif(Derechohab='1',1,0)) As HLTHCOV10,;
        SUM(Iif(Derechoha2='1',1,0)) As HLTHCOV20,;
        SUM(Iif(Derechoha3='1',1,0)) As HLTHCOV30,;
        SUM(Iif(Inlist(Derechoha6,'1','2','3'),1,0)) As HLTHCOV40,;
        SUM(Iif(Sin_derech='6',1,0)) As HLTHCOV60,;
        SUM(Iif(Habla_leng='1',1,0)) As SPKIND,;
        SUM(Iif(Alfabetism='1',1,0)) As LIT,;
        SUM(Iif(Asistencia='1',1,0)) As SCHOOL;
        FROM (wtr_pob);
        GROUP By Ent,Mun,Loc,Ageb,Mza,id_viv,id_hog
        into Cursor c1

** V I V I E N D A
        Select Strtran(Str(Entidad,2),' ','0') As Ent,;
        Strtran(Str(Municipio,3),' ','0') As Mun,;
        Strtran(Str(Localidad,4),' ','0') As Loc,;
        Ageb,Strtran(Str(Manzana,3),' ','0') As Mza,;
        Strtran(Str(Consecutiv,10),' ','0') As id_viv,;
        '  ' As Zona, '     ' As UPM, '  ' As AM,;
        ' ' As TL, ' ' As EstNal, ' ' As EstEst,;
        '002' As Id_Per,;
        SUM(Iif(Material_p='1',1,0)) As FLOOR100,;
        Sum(Iif(Material_p='2',1,0)) As FLOOR202,;
        Sum(Iif(Material_p='3',1,0)) As FLOOR231,;
        Sum(Iif(Disponibil='1',1,0)) As ELECTRC20,;
        SUM(Iif(Disponibil='2',1,0)) As ELECTRC10,;
        SUM(Iif(Disponibi2='1',1,0)) As TV20,;
        SUM(Iif(Disponibi3='1',1,0)) As REFRIG20,;
        SUM(Iif(Disponibi4='1',1,0)) As WASHER20,;
        SUM(Iif(Disponibi5='1',1,0)) As COMPUTR20,;
        SUM(Iif(Disponibi7='1',1,0)) As TOILET20,;
        SUM(Iif(Disponibi7='2',1,0)) As TOILET10,;
        SUM(Iif(Inlist(Disponibi8,'1','2','3','4'),1,0)) As SEWAGE20,;
        SUM(Iif(Disponibi8='5',1,0)) As SEWAGE10;
        FROM (wtr_viv);
        GROUP By Ent,Mun,Loc,Ageb,Mza,id_viv;
        into Cursor C2
Return
```

GLOSARIO.

Microdatos. Datos originados por la aplicación de métodos estadísticos tales como: totales, promedios, frecuencias, sobre grupos o agregaciones de microdatos.

Microdato. Cualquier dato, referido a una unidad estadística individual.

Cálculo de tamaño de muestra. Se refiere a la definición de unidades que deben conformar la muestra en función de la confianza y precisión fijadas.

Tamaño de Muestra. Es el número de elementos u observaciones que tomamos.

Muestreo. Técnica empleada en el análisis parcial de un grupo de casos o eventos, a efecto de obtener cierta probabilidad o certidumbre en relación a las características del universo analizado.

ENOE. Encuesta Nacional de Ocupación y Empleo.

Indicador. Se le da el nombre de indicador a los datos que se seleccionan por su utilidad e importancia en el estudio de determinado tema. De hecho cualquier dato adquiere el dato de indicador cuando es seleccionado para un estudio en particular.

Indicadores:

Población Total. Personas captadas por la encuesta, nacionales y extranjeras, que residen habitualmente en las viviendas seleccionadas en el momento de la entrevista.

Históricamente a lo largo del quehacer estadístico del INEGI los datos absolutos de las encuestas en hogares se han venido ajustando a proyecciones demográficas, no sólo con la finalidad de tener un referente poblacional en periodos inter censales, sino también para eliminar las fluctuaciones en los datos estimados que son inherentes a los esquemas de muestreo probabilístico propios de estas encuestas y, de ese modo, facilitar las comparaciones con periodos previos. Desde un punto de vista metodológico las proyecciones siempre se deben actualizar cada vez que se tengan nuevas evidencias sobre la magnitud y distribución de la población. Los datos que aquí se presentan ya corresponden a una estimación actualizada de las poblaciones totales para cada

trimestre, en función de la evidencia proporcionada por el Segundo Conteo de Población y vivienda (2005).

Modelo conceptual. Descripción de alto nivel de la estructura de la base de datos, en la cual la información es representada a través de objetos, relaciones y atributos.

Diagrama de Entidad Relación: representa la información a través de entidades, relaciones, cardinalidades, claves, atributos y jerarquías de generalización.

Modelo Lógico: representación de una estructura de datos, que puede procesarse y almacenarse en algún SGBD.

BIBLIOGRAFÍA

Bernabeu, R. D. “Investigación y Sistematización de Conceptos-HEFESTO Metodología propia para la Construcción de un DW”, Córdoba, Argentina 2007.

Chaudhuri, S., & Dayal, U. “An Overview of Data Warehousing and OLAP Technology”, Association for Computing Machinery Vol 26 (1), pp. 65-74, Marzo 1997.
Inmon, W. H. “Building the Data Warehouse.” 2nd ed. John Wiley and Sons, Inc., New York, 1996.

Dane, “Homologación de los microdatos censales colombianos 1964 – 1993” Population Center University of Minnesota & CIDS Universidad Externado de Colombia.

Esteve, Albert, Sobek Matthew. Challenges and Methods of International ensus Harmonization. October 2002

Kimball, R., Reeves, L., Ross, M., Thornthwaite, W. “The Data Warehouse Lifecycle Toolkit. Expert Methods for Designing, Developing, and Deploying Data Warehouses”
John Wiley & Sons Inc, 1998.

Kimball, R. & Ross, M. “The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling” John Wiley & Sons Inc, 2002.

McCaa, R. & Esteve, A. “El proyecto IPUMS-International: Microdatos censales para investigadores y planificadores en Chile, Latino América y el mundo.” IASI Seminario Internacional IPUMS-International. Chile, 2003.

McCaa, R. & Esteve, A. “El proyecto IPUMS-International: Microdatos censales para investigadores argentinos, latinoamericanos y del resto del mundo.” Seminario Internacional de Población y Sociedad en América Latina, 2005 (SEPOSAL 2005), 2 Tomos, Salta, Argentina, Tomo I, pp. 51-74.

McCaa, R., Esteve, A., Sobek M. “La Integración de los microdatos censales América Latina” Estudios Demográficos y Urbanos. Enero – Abril 2005

Morgado, I., Isfan, M., “Documenting Variables” Proceedings of Q2006. European Conference on Quality in Survey Statistics. 2006.

Neil, C. G., & Pons, C. F. “Aplicando MDA al Diseño de un Data Warehouse Temporal” VI Jornada Iberoamericana de Ingeniería de Software e Ingeniería del Conocimiento (JIISIC'07). Lima,Peru, 2007.

Phipps, C., & Davis, K. C. “Automating Data Warehouse Conceptual Schema Design and Evaluation” Proc. of the International Workshop on Design and Management of Data Warehouses, 2002.

Stefanov, V. “Bridging the Gap between Data Warehouses and Organizations” Ninth IEEE International EDOC Enterprise Computing Conference, 2005.