# Recomendaciones a considerar en el diseño e integración de Bases de Datos en Ciencias Sociales.

José Ramón Gil-García
María Gabriela Martínez Tiburcio

Reportes de investigación

México estatal

Calidad de Gobierno y Rendición de Cuentas en las Entidades Federativas

México estatal

Calidad de Gobierno y Rendición de Cuentas en las Entidades Federativas

CIDE 35 años

PNUD

# Recomendaciones a considerar en el diseño e integración de Bases de Datos en Ciencias Sociales.

José Ramón Gil-García
María Gabriela Martínez Tiburcio

Núm. 9

Año 2010

Este documento forma parte del proyecto ***Calidad de gobierno y rendición de cuentas en las entidades federativas de México*, coordinado por Guillermo M. Cejudo y Alejandra Ríos Cázares del Centro de Investigación y Docencia Económicas (CIDE)** y financiado con recursos del **Programa de las Naciones Unidas para el Desarrollo México (PNUD)**, provenientes del Centro de Gobernabilidad de Oslo. El proyecto busca contribuir a la generación de capacidades y al fortalecimiento del gobierno democrático en las 32 entidades federativas mexicanas mediante la construcción de indicadores objetivos que sirvan como insumos para evaluar la gestión pública y los mecanismos de rendición de cuentas. Las opiniones, hallazgos y conclusiones son responsabilidad de los autores y no reflejan necesariamente el punto de vista del Programa de las Naciones Unidas para el Desarrollo.

# Recomendaciones a considerar en el diseño e integración de Bases de Datos en Ciencias Sociales

José Ramón Gil-García
María Gabriela Martínez Tiburcio
*CIDE-División de Administración Pública*

## 1. Introducción

Ante la necesidad de racionalizar el uso de los recursos tanto en el ámbito público como en el privado, la toma de decisiones requiere de información oportuna y confiable para establecer líneas de acción que permitan cumplir con sus funciones y objetivos, así como resolver problemas sociales, económicos u organizacionales y rendir cuentas de forma transparente y oportuna. Las bases de datos representan un recurso valioso para la toma de decisiones, pues éstas se encuentran disponibles en el momento en que se requieran utilizar y para distintos usos; además de que el sistema que las almacena puede brindarle aplicaciones muy útiles, según el fin para el cual se estén utilizando esas bases de datos.

Tanto el sector privado como el público, incluyendo a la academia, generan una enorme cantidad de información cuantitativa y cualitativa, que en muchos de los casos se encuentra dispersa o es de difícil acceso, fundamentalmente porque los archivos que la resguardan centran su función en la custodia y preservación y no están orientados a servir las necesidades de los usuarios. Por otra parte, el carácter descentralizado de la operación de quienes producen y preservan la información provoca que su almacenamiento se encuentre en una compleja diversidad de formatos que, además, frecuentemente se hacen obsoletos ante el desarrollo de las tecnologías de manejo de información, lo que complica

su utilización y provoca que mucha de esta información se pueda incluso perder ya que se hace inaccesible.

Las bases de datos también permiten integrar información de distintos sectores y niveles jerárquicos, con lo que se logra establecer un mismo formato y se estandariza el procesamiento de la información; además de que evita que se almacene un dato en varios archivos (lo que se denomina redundancia). Con las bases de datos se puede permitir que el acceso a la información sea total o parcial (diferenciado) de acuerdo al tipo de usuarios, y según lo establecido por la propia organización que las resguarda.

El presente documento tiene como finalidad describir los principales elementos que se deben considerar en el diseño e integración de bases de datos, principalmente en el área de las Ciencias Sociales. Este documento está dirigido principalmente a estudiantes, profesores, investigadores y analistas en Ciencias Sociales; no obstante cualquier persona interesada en el tema encontrará información y recomendaciones útiles.

El documento se encuentra organizado en seis secciones. La primera presenta una breve introducción sobre la relevancia de las bases de datos en las ciencias sociales; la segunda sección del documento, señala los aspectos que se deben considerar en la recolección de datos primarios para la integración de bases de datos, indicando las diversas peculiaridades según el instrumento de investigación. La tercera parte describe algunos aspectos que deben tomarse en cuenta al utilizar datos secundarios para crear una base de datos.

La cuarta sección expone concisamente los conceptos y los elementos básicos de una base de datos; así como los tipos, los niveles de la estructura, los pasos para crear una base de datos, el diseño y las características básicas que deben contemplar las bases de datos. La penúltima parte de este documento presenta algunos aspectos relevantes a

considerar en el desarrollo de indicadores. Finalmente, el último apartado proporciona algunos comentarios finales a manera de conclusión.

## 2.- Aspectos a considerar en la Recolección de Datos Primarios para integrar Bases de Datos

Festinger y Katz (1988: 233) establecen que hay sólo tres métodos para obtener datos en una investigación social: el primero consiste en preguntar a los sujetos; el segundo, en observar la conducta de los individuos, grupos, organizaciones, así como sus resultados o productos; y el tercero, en utilizar información ya existente de documentos, registros, índices y censos. En este apartado se abordarán los principales aspectos que deben considerarse en la creación de bases de datos con información recabada a través del primer método: preguntar a los sujetos. En el siguiente apartado se abordará lo referente a la creación de bases de datos usando el tercer método (información secundaria ó ya existente).

### 2.1 Estrategias, técnicas o instrumentos para recolectar información primaria.

Las estrategias o técnicas para recolectar datos primarios en una investigación social son diversas, pero destacan: los cuestionarios, las encuestas, las entrevistas, y desde luego la observación del sujeto u objeto de investigación. El cuestionario es un instrumento orientado a conseguir respuestas a preguntas, utilizando un impreso ó formulario que el encuestado llena por sí mismo; éste debe tener una extensión y ámbito limitados, por lo que se recomienda que el tiempo que éstos van a requerir para ser contestados no exceda de treinta minutos (Goode y Hatt, 1980:164). Esta herramienta es recomendable para obtener información de un número grande de personas de alguna población objetivo, por lo que las

bases de datos originadas por este método son muy grandes, principalmente en lo que respecta al número de observaciones o de encuestados.

Por lo que se refiere a la investigación por este método de recolección de datos, se obtiene información directamente de las personas que fueron seleccionadas (muestra) para establecer una base sobre la que se pueda inferir sobre una población más amplia (Manheim y Rich, 1988: 145), pero que el instrumento es un cuestionario. De acuerdo con Manheim y Rich (1988) este método es idóneo para estudios cuyas unidades de análisis son los individuos y donde los principales conceptos pertenecen a éstos (opiniones, actitudes o percepciones); pero se recomienda utilizarla solo en aquellos casos donde la información requerida no puede obtenerse con más facilidad y con menos costos (Campbell y Katona, 1990: 32). Este tipo de instrumento de investigación permite recabar información que puede ser más fácil de codificar y que es más susceptible a la cuantificación o medición, lo que facilita hasta cierto punto la captura de la información y el diseño de la estructura de la base de datos.

Mientras tanto, la entrevista se realiza típicamente cara a cara, donde el entrevistador hace preguntas al entrevistado verbalmente y va realizando sus anotaciones (Babbie, 1988: 210). La información que se obtiene mediante este método es más de tipo cualitativo y más complejo de medir y codificar; lo que hace más compleja la tarea de elaborar la base de datos y de la captura de la información. Esta técnica, al igual que la encuesta, se recomienda utilizar sí los datos fundamentales de un proyecto de investigación son las percepciones, actitudes y opiniones del individuo, los cuales no pueden inferirse solamente con la observación (Cannell y Kanh, 1990: 314).

En diversas investigaciones sociales, con la finalidad de alcanzar una mayor calidad en la recolección de información, así como de cumplir con los criterios de confiabilidad y

validez, se aplican todas estas técnicas de recolección de información para un mismo estudio, de tal forma que se puedan complementar, y en su caso verificar la información que se obtiene con una técnica mediante la realización de otra.

Considerando el objetivo que persigue este documento, es necesario señalar la relevancia y el impacto de aplicar una u otra técnica de investigación en la integración de bases de datos. Para el caso de la encuesta, cuestionario y entrevista, se pueden considerar que se pueden agrupar en dos rubros: la encuesta y el cuestionario en uno solo, debido a que es el mismo instrumento a través del cual se recaba la información (cuestionario), por tal motivo para fines de este documento se entenderá que cuestionario y encuesta es lo mismo; y en el caso de la entrevista en otro, por ser un instrumento más cualitativo.

Otro elemento que permite diferenciar a estos dos grupos de técnicas, reside en el tamaño de la población que será sujeta a dicha técnica de investigación, es decir, el número de entidades o personas que serán cuestionadas, pues esto determina el tamaño de la base de datos. En este caso, es mucho más probable que el número de individuos sea mucho más grande cuando se realiza investigación mediante el cuestionario que cuando se llevan a cabo entrevistas, por la inversión que se tendría que hacer en recursos económicos y humanos, así como del tiempo de que dispone el sujeto que responderá a dichos cuestionamientos.

Otro aspecto que las distingue a estas técnicas (entrevista y cuestionario) y que también las separa en dos grupos, es el número y tipo de preguntas que utilizan, pues éstas suelen ser más numerosas en el cuestionario que en una entrevista, debido también a los mismos factores que se mencionaron anteriormente y al tipo de pregunta que predomina

más en determinado instrumento: cerrada o abierta[1]. Ambos tipos de preguntas se encuentran presentes en las dos técnicas, no obstante, en la entrevista las preguntas, por lo general, son abiertas dejando al entrevistado que responda con mayor detalle y proporcione más información. Esto influye en la elaboración de la base de datos, pues mientras la codificación de las respuestas cerradas tienen la opción de elegir: una respuesta o varias respuesta, éstas son cortas y es más fácil codificar e identificar como se capturará en una base de datos; además de que pueden ser más susceptibles a cuantificarse. En lo que se refiere a las respuestas a preguntas abiertas, éstas por lo general pueden ser muy grandes lo que implica un mayor número de caracteres, con una mayor complejidad para intentar codificar y cuantificar, y al analizar la información puede originar que una respuesta origine más de una variable.

Lo antes señalado origina que, cuando en un instrumento de recolección de información utiliza ambos tipos de preguntas (abiertas y cerradas) en proporciones relativamente iguales, se recomienda que se elaboren dos bases de datos, una para cada tipo de preguntas, con la finalidad de facilitar el uso de sus datos, en el desarrollo y aplicación de técnicas estadísticas y en el análisis mismo de dicha información.

[1] La pregunta abierta es aquella en la que se estructura el tema para el sujeto a entrevistar pero se le da la libertad de que responda con sus propias palabras y hable todo lo que desee. Mientras que la pregunta cerrada contiene ya las posibles respuestas, de tal forma que sólo el sujeto debe limitarse a seleccionar la opción que se aproxime más a su opinión (Cannell y Kahn, 1990: 330).

## 2.2 Pasos a seguir para el diseño e integración de bases de datos de información primaria, según instrumento o técnica de recolección de información seleccionada.

### 2.2.1 Para el diseño e integración de las bases de datos de información primaria a través de cuestionario se deben seguir los siguientes pasos.

**Primero:** se selecciona la(s) estrategia(s) de recolección de información que responda(n) a las necesidades de la investigación, es decir, se establece que se aplicará cuestionario.

**Segundo:** se debe establecer el medio a través del cual se va a recolectar la información, si la encuesta y/o el cuestionario se aplicará vía telefónica y/o haciendo uso del servicio postal, y/o usando las tecnologías de información (correo electrónico o la utilización de un portal electrónico), y/o personalmente (cara a cara). Considerando que los distintos casos influyen en la elaboración de la base de datos, en cuanto al tiempo que se tardara en procesar la información, y acorde con el número de personas que participarán en el levantamiento de la información, así como de quién (es) será (n) responsable (s) de administrar las bases de datos y determinar el programa o *software* que se utilizará para el almacenamiento y manejo de la base de datos.

**Tercero:** en este siguiente paso se debe decidir si se aplicará el cuestionario a toda la población sujeta a investigación ó a una muestra representativa (para este efecto, se debe elegir un método de selección de la muestra y determinar el tamaño de la misma). En esta fase es importante considerar el financiamiento y los recursos que se tienen para dicho proyecto.

**Cuarto:** diseñar el instrumento que se aplicará, considerando las necesidades de su investigación, se tienen que elaborar y diseñar las preguntas (abiertas o cerradas), considerando su posible codificación. En este paso también se debe diseñar la estructura de la tabla en la que se vaciará la información, retomando la recomendación de que si hay un

porcentaje significativo de preguntas abiertas, se diseñen dos tablas: una para preguntas cerradas ó de opción múltiple, y otra para preguntas abiertas.

Por otra parte, las bases de datos basadas en cuestionario suelen ser muy grandes en cuanto al número de casos ó individuos a los que se aplicó el instrumento, es decir, la base de datos es mucho más extensa en cuanto al número de casos ó filas; y sus columnas van a corresponder al mismo número de datos solicitados o preguntados a cada sujeto de investigación. Cabe señalar que es importante tener presente el tamaño de la base de datos (tanto en filas como en columnas) para elegir el programa estadístico o *software* que se va a utilizar para concentrar la base de datos, pues en algunos casos como el de *Excel* (en versiones anteriores a la de 2007) se tiene un número limitado de filas y columnas (*Excel* 256 columnas y 65536 renglones).

Un ejemplo del tamaño de una encuesta y de su cuestionario, que confirma lo que antes se mencionó, se puede encontrar en: *Encuesta a la población en reclusión en el Distrito Federal y Estado de México 2005*, cuyo tamaño de la muestra es 1264 reclusos sentenciados y su cuestionario consta de 236 preguntas[2], estudio realizado por el Dr. Marcelo Bergman, y las Doctoras Elena Azaola y Ana Laura Magaloni del Programa de Estudios para la Seguridad Pública y Estado de Derecho (PESED) y el CIDE.

**Quinto:** elaborar una carta de presentación e invitación que permita establecer un vínculo entre el investigador y los sujetos del estudio, en ésta se debe presentar el investigador señalando a qué institución u organización pertenece, explicando qué está realizando y el porqué del estudio; además de proporcionar argumentos que persuada al encuestado a responder a estos instrumentos, así como también indicar las instrucciones de cómo

[2] Para mayor información sobre esta encuesta y su base de datos, puede consultarla en: http://biiacs-dspace.cide.edu:8080/dspace/handle/10089/16085

responder al cuestionario, y la garantía del anonimato. Cabe señalar que es recomendable que esta carta sea breve para evitar que sea desechada sin ser leída o con efectos negativos hacia la investigación.

**Sexto:** aplicación del instrumento, para este paso es recomendable realizar una prueba previa o estudio piloto antes de iniciar el trabajo de campo concreto, con la finalidad de corroborar que las instrucciones son claras, probar el cuestionario (que las preguntas están correctamente planteadas y que son comprensibles), analizar el instrumento para ver si las respuestas satisfacen los objetivos de la investigación (Cannell y Kanh, 1990: 332), así como si están contemplando todas las respuestas posibles e incluso considerar aquellos casos de no sabe, no contesto o no aplica; y verificar que la codificación o categorización de las respuestas, es la más apropiada para la captura de la información en las bases de datos diseñadas para tal efecto. Tener contemplado este tipo de situaciones va a permitir mejorar la calidad de la información y facilitar la posterior captura de los datos.

Para la siguiente fase de procesamiento de información se debe tener mucho cuidado al codificar las respuestas, ya que se debe tener presente si éstas involucran más de una opción y por tanto son varias las respuestas ó, si se trata de una respuesta que pueden tener dos ó más opciones, pero que sólo se tiene que elegir una (es decir, son excluyentes). Ejemplos: cuando se realiza una pregunta en donde la respuesta es sí o no, ¿Actualmente tiene trabajo? 1) Sí ó No; y cuando la respuesta a un cuestionamiento puede ser más de una: ¿Qué otros ingresos tenía? 1) Salario, 2) Pensiones ó seguro social, 3) Progresa, 4) Dinero de familiares, 5) Dinero de amigos, 6) Dinero por venta de droga (ver la Tabla No. 1 y No. 2).

**Tabla No. 1**

| | Base de datos | |
|---|---|---|
| P59. ¿Usted paga impuestos?<br>1) SI<br>2) NO<br>8) NO SABE N/S<br>9) NO CONTESTO N/C | **No. Folio** | **P59** |
| | 6 | 2 |
| | 7 | 1 |
| | 8 | 1 |
| | 9 | 1 |
| | 10 | 1 |

| | Base de datos | |
|---|---|---|
| P59a. Comparando con los servicios que recibe ¿cómo considera los impuestos que usted paga?<br>1) Muy altos<br>2) Algo altos<br>3) Justos<br>4) Algo bajos<br>5) Muy bajos<br>8) N/S<br>9) N/C<br>0) NO APLICA N/A | **No. Folio** | **P59a** |
| | 6 | 0 |
| | 7 | 2 |
| | 8 | 1 |
| | 9 | 2 |
| | 10 | 3 |

**Base de datos con ejemplos de codificación de preguntas cerradas con una sola respuesta**:

Fuente: esta tabla presenta una pequeña parte de información de la base de datos del CIDE, Estudio Comparativo de los Sistemas Electorales (CSES) – 2003 [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [14 septiembre 2009], http://hdl.handle.net/10089/3687

**Tabla No. 2**
**Base de datos con ejemplo de codificación de pregunta en donde el encuestado puede elegir varias respuestas**

p26 ¿A quién cree que ha apoyado más el gobierno del Presidente Fox?

0) No respondió
1) Los pobres
2) Los empresarios
3) La Iglesia
4) Los sindicatos
5) El PAN
6) Los indígenas
7) Las clases medias
8) La población en general
9) Los inversionistas extranjeros
10) A Estados Unidos
11) Los Bancos
12) Los braceros
13) Los trabajadores

| FOLIO | p26a | p26b | p26c | p26d | p26e | p26f | p26g | p26h | p26i |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 2 | 6 | 7 | 10 | 0 | 0 | 0 | 0 |
| 11 | 1 | 2 | 8 | 11 | 12 | 13 | 0 | 0 | 0 |
| 12 | 2 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 | 10 |
| 14 | 2 | 5 | 7 | 9 | 10 | 11 | 0 | 0 | 0 |

Fuente: esta tabla presenta una pequeña parte de información de la base de datos de Albo, Andrés, Martínez de Velasco, Alberto, BANAMEX y FACTUM MERCADOTÉCNICO, Pulso Sociopolítico - 2003 [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [14 septiembre 2009], http://hdl.handle.net/10089/16075

**Séptimo:** procesamiento de la información recolectada, en la que el investigador enfrenta lo que se considera un problema común, el cual consiste en encontrar la manera de asignar un solo valor representativo o calificación a un dato [comportamiento, actitud, opinión, etc.] (Manheim y Rich, 1988: 193); es decir, en esta fase se traducen los objetos reales a símbolos o nombres, lo que Coombs (1990) denomina: medición.

Esta necesidad de medir variables por parte de un investigador social reside en diversas razones, entre las que destacan: es de interés para el investigador cuantificar el

material simbólico o cualitativo para poder comparar diferentes conjuntos de material y examinar relaciones en una forma precisa (Cartwright, 1990: 410); y en muchas ocasiones el investigador desea emplear una medición refinada de las variables que está utilizando en su investigación, que le permita desarrollar un análisis más sofisticado de su información (Babbie, 1988: 300).

Dos elementos que son utilizados frecuentemente como medidas de variables en la investigación social son los índices y las escalas. Éstos son medidas acumulativas, y se utilizan debido a que aún cuando el cuestionario se haya elaborado lo más cuidadosamente posible, rara vez se logrará alcanzar que una sola pregunta represente una variable compleja, e incluso algunos fenómenos no pueden ser medidos directamente y para tal fin se desarrollan una o más preguntas que de manera indirecta nos puede proporcionar algún dato sobre esa variable. Debido a la relevancia que la medición tiene en la elaboración de bases de datos y en el desarrollo de la investigación en las ciencias sociales se tratará más detalladamente este proceso en el último apartado de este documento.

### 2.2.2 Diseño e integración de bases de datos con información primaria a través de entrevistas

Una vez que se ha elegido como estrategia de recolección de información la entrevista, se tienen que realizar los siguientes pasos:

**Primero:** identificar a las personas que pueden proporcionar información relevante y amplia sobre el objeto de estudio, con base en la literatura y documentos revisados sobre el estudio. Asimismo, se recomienda que una vez identificadas las personas que serán entrevistadas, se les envíe alguna carta u oficio informándoles de la visita de lo entrevistadores, con la finalidad de que se pueda preparar para dicho evento.

**Segundo:** se diseña la guía de entrevista, que consiste en un listado de preguntas, las cuales se caracterizan por permitirle al entrevistado que conteste con sus propias palabras y libremente, por lo que sus respuestas no están limitadas. Sin embargo, a diferencia del cuestionario y de la encuesta donde el número de personas cuestionadas es normalmente grande, en el caso de la entrevista se aplica a un número relativamente más pequeño de personas pero las respuestas de éstas son mucho más amplias. Por lo general, de cada pregunta se puede obtener información muy diversa que al analizarla y registrarla puede dar origen a muchas variables. Este tipo de técnica de investigación tiene como objeto producir material simbólico, verbal, cualitativo, que por su complejidad no se codifica en una sola categoría (Cartwright, 1990: 389).

En este caso las bases de datos cuyo origen reside en una entrevista, el número de filas es pequeño, pues el número de entrevistas (dependiendo del estudio) difícilmente llegan a ser mayor a dos dígitos. Realizar entrevistas a muchas personas implicaría altos costos, así como contar gran disposición de las personas a entrevistar. Pero las columnas pueden ser mucho más amplias y más numerosas, las cuales no necesariamente van a corresponder al mismo número de preguntas, pues de una respuesta pueden desprenderse más de una variable, y no siempre es muy sencillo la codificación o categorización; por lo que la estructura de la tabla estará acabada hasta que se halla concluido el total de entrevistas a aplicar y después de realizar el análisis de la información. Por otra parte las entrevistas también pueden ser resguardadas en una base de datos a través de una lista de archivos, y no necesariamente con la información capturada en una tabla.

Con la finalidad de mostrar las características de una base de datos desarrollada con base en información obtenida de entrevistas, se presentan los datos principales del estudio realizado por José Ramón Gil García en NUEVA YORK e INDIANA, donde se aplicaron

14 entrevistas a organizaciones gubernamentales y privadas. Cabe señalar que la guía de entrevista consta de 18 preguntas[3].

**Tercero:** Es importante que el investigador realice una presentación del proyecto con información referente a la naturaleza de la entrevista, los objetivos y aspectos más relevantes del estudio que está realizando, y tener preparadas explicaciones sobre la guía de entrevista (preguntas), con la finalidad de aclarar las dudas sobre los cuestionamientos y explicar aquellos términos más técnicos que deben ser explicados por un especialista en el tema. Esta presentación será de utilidad para capacitar a las personas que aplicarán la entrevista, en caso de que el (la) investigador(a) no realice las entrevistas personalmente, sino que contraten a otros para llevar a cabo esa actividad. Asimismo, en esta capacitación también se les debe indicar y proporcionar los documentos que llevarán consigo los entrevistadores al aplicar el instrumento (carta de presentación, credencial, guía de entrevista, etc.).

**Cuarto:** al igual que el cuestionario, es recomendable realizar una prueba piloto que permita identificar los posibles problemas en el planteamiento de las preguntas y en el desarrollo de la entrevista; esta prueba piloto además ayudará a corregir algunos aspectos de planeación, los cuales pueden evitar gastos innecesarios y mejorar la calidad de la información recolectada.

**Quinto:** aplicación del instrumento a las personas seleccionadas para tal efecto.

**Sexto:** como último paso se realiza el procesamiento de la información, considerando que una gran parte de las preguntas tendrán respuestas que no se podrán clasificar en categorías sencillas, lo que va a requerir de un análisis más profundo; y por lo tanto, se requiere combinar dos o más variables para tener un indicador que permita analizar un aspecto

[3] Para mayor información consulte: http://biiacs-dspace.cide.edu:8080/dspace/handle/10089/16063

complejo del objeto de estudio, el cual comprenda a la mayoría ó en el mejor de los casos a todos sus elementos. El análisis de este tipo de información (cualitativa) se puede llevar a cabo usando diferentes programas que se han desarrollado para tal efecto, tales como: *Atlas.ti*, *Nvivo*, *Ethnograph*, etc.

Además de la información primaria que se puede obtener a través de la aplicación de un cuestionario, encuesta o entrevista, en la mayoría de los estudios se debe realizar un análisis más completo de su objeto de estudio para ubicarlo dentro de sus debidas dimensiones, por lo que es necesario recurrir a información existente sobre éste, ó información que puede complementarlo ó que permita contextualizarlo en su realidad; para alcanzar este fin se recurre a información secundaria, que se puede encontrar en censos, estadísticas de dependencias gubernamentales, profesionales, del sector privado y sociales, entre otras fuentes secundarias. En el siguiente apartado se tratará sobre este tipo de información.

## 3. Recomendaciones para Recolectar Datos Secundarios para integrar Bases de Datos.

### 3.1 ¿Qué son los datos secundarios?

Los datos secundarios son aquellos sobre los cuales el científico social tiene relativamente poco control, pues fueron elaborados por otra persona u organización o institución, y ésta fue la que determinó la forma de almacenamiento e integración de los datos (Angell y Freedman, 1990: 286). Algunas fuentes de información de datos secundarios o agregados son: censos, bibliotecas, investigaciones realizadas por otros investigadores, información de organizaciones de la sociedad civil y de dependencias públicas (educación, salud, seguridad nacional, etc.). Enseguida se señalaran las ventajas y desventajas de utilizar información secundaria en la investigación en general y en el desarrollo de bases de datos en particular.

### 3.2 Utilidad de los datos secundarios

La información obtenida de registros y censos puede ayudar a desarrollar análisis científico de datos de tiempos y lugares remotos, así como de series de hechos de diferentes épocas, a diferencia de los datos que se obtienen del diseño de una investigación particular con un fin específico (información primaria), que se enfoca a un universo ubicado en un tiempo y espacio determinado (Angell y Freedman, 1990: 286).

Angell y Freedman (1990) mencionan que los datos de registro y censos pueden utilizarse para comparar la frecuencia de la presentación de una variable en diferentes lugares y en distintas condiciones temporales; asimismo, esta información se puede usar para estudiar las distintas relaciones (asociación y/o causalidad) que se presentan entre diversas variables que son significativas para un problema de investigación. Otro uso que puede tener este tipo de información es en el diseño de investigación, para seleccionar una muestra con características específicas (similares o contrastantes), y llevar a cabo un estudio que permitirá obtener mayor información sobre el fenómeno estudiado. Finalmente, los índices elaborados por diversas entidades públicas u organizaciones también pueden ser utilizados como medidas aproximadas de una variable (variables *"proxy"*).

Por lo que respecta a la información secundaria obtenida de una encuesta, trae consigo varios beneficios ya que éstas proporcionan información que el investigador original no utilizó debido a que esos datos fueron solo de interés marginal para su estudio; asimismo, esos datos que no fueron utilizados más profundamente por el investigador original pueden servir para responder a otras interrogantes (Manheim y Rich, 1988: 169).

Además, los datos secundarios dan origen a otra nueva base de datos, pues para que puedan ser usados por el investigador, generalmente tendrán que ser transformados

(combinar variables, crear índices) y vaciarlos a otro formato, el cual puede contener otros datos secundarios y/o información primaria (obtenida por el propio investigador a través de técnicas como: observación, cuestionarios, encuestas, entrevistas, etc.), pero este formato será acorde a los fines de su investigación, y muy distinto a aquellos en donde consulto esa información (Manheim y Rich, 1988; Angell y Freedman, 1990).

Un ejemplo de este tipo de base de datos que tuvo su origen en datos secundarios es la desarrollada por Vilalta y Fernández (2009), cuyo título es *"Estadísticas judiciales: homicidio y robo en 60 áreas metropolitanas de México 1997, 2000, 2005 y 2007"*, ésta tuvo su fuente original en un documento del Instituto Nacional de Estadística y Geografía (INEGI) denominado "*Estadísticas Judiciales en Materia Penal"* y se pueden encontrar en el acervo del BIIACS[4].

### 3.3 Aspectos a considerar y/o limitaciones de utilizar datos secundarios

La principal limitación que presentan los datos secundarios consiste en que la definición operativa de los datos y las posibilidades de manipulación experimental se encuentran fuera del control del investigador (Angell y Freedman, 1990: 306). Así como también la información puede aparecer ya resumida o procesada lo cual puede ser no muy útil para los fines que tiene el investigador (Manheim y Rich, 1988).

Además, el investigador al utilizar datos secundarios debe tener presente que éstos tienen limitaciones que, en su mayoría, se refieren al nivel de confianza y validez de esa información. Estas limitaciones se relacionan principalmente con la naturaleza de los datos, la eficacia con la que éstos se registran y los incentivos que tienen las personas para registrar el hecho tratado (Angell y Freedman, 1990: 303).

[4] Más información sobre la base de datos consultar: http://hdl.handle.net/10089/16107

Las limitaciones de los datos secundarios relacionadas con su naturaleza, se refieren principalmente a considerar que la información registrada sobre un hecho que representa información delicada o conflictiva para los encuestados o entrevistados, puede no ser representativa ni ser un indicador de tal hecho, pues los entrevistados pudieron haber falseado la información para no ser señalados o estigmatizados, como puede ser información sobre el pago de impuestos (evasor), víctima de abuso sexual, entre otros.

Asimismo, dentro de la naturaleza del dato se debe considerar que en muchas ocasiones éstos pertenecen a un campo de conocimientos muy sofisticado ó técnico, donde el personal responsable de realizar los registros pudo no haber desarrollado su tarea con precisión debido a la carencia de conocimientos sobre ese tipo de información o porque esta información fue obtenida de personas que no eran las idóneas, influyendo ambas situaciones en la calidad del registro de los datos. Otro factor que puede influir en el registro de los datos, es que el autor del registro haya considerado que los datos no tenían importancia para la dependencia u organización para la cual esta laborando y haya realizado un registro subjetivo o incompleto.

En lo que respecta, a utilizar información obtenida de encuestas por muestreo, ésta debe ser una muestra representativa de la población que es objeto de estudio; además, el instrumento de encuesta debe contener operacionalizaciones apropiadas de las variables que son clave en el estudio que esta desarrollando (Manheim y Rich, 1988: 170, 285).

### 3.4 Estructura de datos y campos individuales (compatibilidad entre dos o más bases de datos).

Manheim y Rich (1988) mencionan algunos de los aspectos que se deben considerar en la utilización de datos secundarios: la forma en que se encuentran los datos, su contenido y

calidad pueden ser diversos, lo que dificultará la posibilidad de realizar comparaciones válidas entre las unidades y generalizaciones.

Por lo que el investigador, al realizar comparaciones de información secundaria entre distintas entidades, organizaciones y/o dependencias, debe observar y revisar que el sistema de registro, clasificación y definición de la información sean compatibles. También al usar series temporales de este tipo de datos, es muy importante que el investigador verifique que las normas que regulan el registro de los datos no hayan sufrido cambios a lo largo del tiempo, pues estos cambios influyen en los resultados y en las interpretaciones del objeto de investigación (Angell y Freedman, 1990: 304).

Ejemplos de las situaciones anteriores se pueden señalar varios: como el caso de utilizar algún indicador construido por alguna dependencia, como puede ser el índice de marginación del Consejo Nacional de Población (CONAPO), se debe verificar que las variables que son consideradas para la construcción de este indicador siempre han sido las mismas, pues si se agregó o eliminó alguna de esas variables de un año a otro, puede alterar este indicador. Otro ejemplo se puede encontrar en la clasificación de los ingresos y egresos de las entidades federativas que hace el Instituto Nacional de Estadística y Geografía (INEGI), verificando que ésta no haya cambiado en el tiempo (para un estudio de serie de tiempo), y en su caso que sea compatible esta clasificación con otras entidades con las cuales se desea realizar alguna comparación.

Otro aspecto que los investigadores deben considerar es, si se desea usar datos agregados de un nivel de análisis distinto del que se está trabajando, utilizar técnicas de análisis que reduzcan los riesgos que conllevan la inferencias entre distintos niveles de análisis, por ejemplo: si utiliza información secundaria a nivel estatal y los resultados que se obtienen del análisis de esta información se quieren aplicar para todo el país, esto no es

posible, porque muy probablemente cuando se realice un análisis con información nacional los resultados serán muy distintos (Manheim y Rich, 1988: 288).

También es importante señalar que si se desea utilizar información que se obtuvo de diferentes encuestas, éstas deben coincidir en los requerimientos metodológicos para que sean compatibles entre ellas (tamaño de la muestra, margen de error, nivel de análisis, etc.), así como con el estudio que se está realizando, para no cometer errores de análisis e interpretación.

### 3.5 Algunas fuentes de información de datos secundarios recomendables.

Entre las fuentes de información de datos secundarios se encuentran: documentos, registros, materiales censales e índices. En lo que respecta a documentos, se refieren a cartas personales, historias de vida, e informes. En datos de registros y censos[5] se tiene información sobre datos de vida, educación, delitos, votaciones, ingresos, estatus migratorio, características de las viviendas, etc.; y en cuanto a índices, éstos se refieren a aquellos desarrollados por alguna institución, como lo es el índice del costo de vida, de capital humano, índice de corrupción y buen gobierno, etc. (Angell y Freedman, 1990: 286).

Por su parte, Manheim y Rich (1988: 283) establecen que hay seis tipos de datos secundarios o agregados. El primer tipo son datos del censo, el segundo son las estadísticas de organizaciones de la sociedad civil, gubernamentales, comerciales y profesionales; el tercero, son las encuestas por muestreo; otro tipo de datos secundarios es el contenido de las publicaciones; el quinto se refiere a los datos sobre sucesos, y finalmente, los datos

[5] De acuerdo con Angell y Freedman (1990: 294), los datos de registro se refieren a informes realizados en el momento en que sucede un hecho según las regulaciones legales o administrativas; y por lo que al censo se refiere, éste consiste en una reunión periódica de datos sobre una población y es llevada a cabo casa por casa.

estimativos, que son utilizados cuando no hay datos que permitan medir con exactitud alguna característica, por ejemplo: la población proyectada por la CONAPO de 2005 hasta 2050[6].

## 4. ¿Cómo integrar una Base de Datos?

Esta sección explica algunos conceptos básicos relacionados con bases de datos y su integración.

### 4.1 ¿Qué es una base de datos?

Una colección de datos usualmente es denominada base de datos y ésta contiene información sobre una organización determinada (Korth y Silberschatz, 1986: 1)[7]. Elmasri y Navathe (2000: 4) señalan que una base de datos tiene la propiedad de representar algunos aspectos del mundo real; además es una colección coherente de datos con significados inherentes (fuente de información); y finalmente ésta se diseña, construye e integra con datos para un fin específico, orientada a un grupo de usuarios y con aplicaciones que pueden ser de interés para estos usuarios (audiencia). Para fines de este documento, esta definición es la que se considera más completa y apropiada para entender los alcances y las limitaciones de las bases de datos.

De acuerdo con Elmasri y Navathe (2000: 5) una base de datos se puede crear y mantener de manera manual o puede crearse mediante la utilización de programas de

[6] Información obtenida el 29 mayo de 2009, de la página: http://www.conapo.gob.mx/index.php?option=com_content&view=article&id=36&Itemid=199

[7] Otros autores como Coll-Vinent (1988: 75) define base de datos como "un depósito de datos de interés y valor para una amplia gama de usuarios". Otra definición de base de datos es la que proporciona Wiederhold (1985: 2) como "un conjunto de datos relacionados". Duffy (2000: 2) amplía el concepto anterior estableciendo que una base de datos es un conjunto de información relacionada con una aplicación específica. Por otra parte, Senn (1989: 385) establece que una base de datos es una compilación de datos almacenados y organizados con base en las relaciones que existen entre ellos mismos, y no tomando como base las estructuras de almacenamiento. Elmasri y Navathe (2000: 4) consideran que la definición de "colección de datos relacionados" es muy general, y amplían este concepto, el cual se considera apropiado para fines de este documento.

aplicación diseñados para dicha tarea o mediante un sistema de gestión de bases de datos. Un sistema de gestión de bases de datos (*DBMS* por sus siglas en inglés, SGBD en español), constará de una colección de datos interrelacionados y un conjunto de programas de acceso a esos datos, cuyo objetivo primordial es proporcionar un ambiente que es conveniente y eficiente para su utilización en la recuperación y almacenamiento de la información en la base de datos[8] (Korth y Silberschatz, 1986: 1).

El sistema de gestión de base de datos es un sistema (*software*) que facilita los procesos de definición, construcción y manipulación de bases de datos para distintas aplicaciones (Elmasri y Navathe, 2000: 5). Estos sistemas deben proporcionar tres funciones para acceder a las bases de datos (Buyens, 2001: 21). La primera es la de selección (también se le llama restringir), esta función mostrará sólo los registros que se hayan elegido, de acuerdo a los valores y campos especificados, es decir, presenta una vista[9] de una tabla con esos registros. La segunda es proyectar, que consiste en presentar una vista de una tabla que no incluye todos sus campos, es decir, extrae las columnas especificadas de una tabla. Y, la tercera es la función de unir o juntar que presenta una vista que combina dos tablas como si fueran una sola.

### 4.2 Tipos de bases de datos

Existen diversos tipos de bases de datos. Se pueden clasificar de acuerdo al tipo de información que contiene o de acuerdo a la estructura que presenta. A continuación se describe brevemente cada uno de estos tipos.

[8] Cabe señalar que algunos expertos en el tema, denominan a está unión de base de datos con el sistema de gestión: banco de datos. Sin embargo, Coll-Vinent (1988: 80) considera que no existe diferencia sustantiva entre banco y base de datos, y en la práctica son consideradas como realidades idénticas, por lo que banco y base de datos se pueden manejar indistintamente.

[9] "Una vista es un subconjunto de la base de datos ó puede contener datos virtuales derivados de los ficheros de las bases de datos pero que no están explícitamente almacenados" (Elmasri y Navathe, 2000: 10).

- Bases de datos según información: cuantitativas y cualitativas.

Las bases de datos se pueden clasificar según el tipo de información que tienen almacenada en cuantitativas o cualitativas. Las primeras son las más comunes y su información está expresada en números. La información cuantitativa describe alguna característica o propiedad del objeto de estudio (persona, cosas, sucesos, etc.) pero expresado o codificado en números; además éste tipo de información tiene una anchura relativamente pequeña en comparación con información expresada en texto.

Por otra parte, las bases de datos cualitativas contienen información mucho más amplia, pues los datos que están almacenados son caracteres alfanuméricos y son enunciados de texto, los cuales ocupan un espacio mayor dentro de cada columna. El tipo de información que almacena una base de datos tiene relevancia sobre su diseño e integración, pues dependiendo de ésta se elegirán el tipo de caracteres y la anchura de cada columna, así como si tendrá o no valores predeterminados; entre otros elementos que se deben considerar.

- Bases de datos según la estructura: planas, relacionales y data *warehouse*

**Bases de datos planas**

Las bases de datos planas son las que se componen de una tabla. Una **tabla,** de acuerdo con Duffy (2000) y Buyens (2001), es la entidad de almacenamiento y es la unidad básica de organización de cualquier base de datos. En una tabla, las columnas representan campos y las filas registros (Buyens, 2001: 20). En una base de datos, un registro es una unidad dentro de una tabla que contiene información relacionada sobre una sola entidad (objeto, persona, lugar, cosa, concepto, o suceso); mientras que un campo es una unidad más pequeña que un registro que contiene un hecho acerca de la entidad (algunos autores le llaman atributo) y es una propiedad que describe algún aspecto del objeto que se esta

almacenando, como por ejemplo: nombre, dirección, teléfono (Duffy, 2000: 2; Connolly y Begg, 2005: 14).

Una tabla representa una entidad y cada fila de la tabla representa una ocurrencia de esa entidad. Las columnas de esa tabla describen los atributos de ésta, por lo que cada una de esas columnas puede almacenar un tipo específico de información o tipo de datos (Waymire y Sawtell, 2000: 276). Por lo que se refiere a la estructura de la tabla, ésta es una serie de instrucciones respecto de la disposición de información dentro de cada registro, el tipo de caracteres utilizados para almacenar cada campo (por ejemplo: numérico, alfanumérico, cadena) y el número de caracteres también requerido para cada campo (Duffy, 2000: 2).

El siguiente ejemplo muestra una tabla de hogares, en donde los hogares son una entidad lógica, cada fila representa una instancia individual de un hogar y las columnas que conforman la tabla describen al hogar; algunas columnas podrían ser: tenencia de la vivienda, número de cuartos, número de recamaras, material de muros, material de techo, material del piso, etc. (ver Tabla 3).

**Tabla No. 3**

| HOGAR (folio) | TENENCIA DE LA TIERRA | CUARTOS | RECA-MARAS | MUROS | TECHO | PISO |
|---|---|---|---|---|---|---|
| 200001110020 | Propia y totalmente pagada en terreno propio | 3 | 2 | Tabique, ladrillo, tabicón, block, | Tabique, ladrillo, tabicón ó loza de concreto | Madera, mosaico, loseta de concreto, loseta de plástico u otros recubrimientos |
| 200001110150 | Prestada | 4 | 3 | Tabique, ladrillo, tabicón, block, | Tabique, ladrillo, tabicón ó loza de concreto | Madera, mosaico, loseta de concreto, loseta de plástico u otros recubrimientos |
| 200001110300 | Rentada o alquilada | 2 | 2 | Tabique, ladrillo, tabicón, block, | Vigueta y poliuretano, vigueta y bovedilla, vigueta y cuña | Cemento ó firme |

Fuente: Elaboración propia, esta tabla presenta una pequeña parte de información de la base de datos del Instituto Nacional de Estadística y Geografía. (2000). Encuesta nacional de ingresos y gastos de los hogares 2000 [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [30 junio 2009], http://hdl.handle.net/10089/16078

## Bases de datos relacionales

El modelo relacional representa la base de datos como un conjunto de relaciones; en donde cada relación se asemeja a una tabla de valores, ó a un fichero plano de registros (Elmasri y Navathe, 2000: 186). Por lo que una base de datos relacional se define como "una colección de relaciones normalizadas en la que cada relación tiene un nombre distintivo" (Connolly y Begg, 2005: 67).

Una base de datos relacional permite unir registros de dos o más tablas basadas en los contenidos de un campo común ó identificador único; por lo que ésta puede contener cualquier número de tablas (Buyens, 2001: 22). Se pueden usar uno o más campos para definir la llave, la cual se usará para ordenar, identificar y recuperar los registros en la base

de datos (Duffy, 2000: 2), así como para integrar dos o más tablas. Según Senn (1989: 390), la base de datos relacional utiliza una estructura de datos que es una tabla de dos dimensiones, en donde los renglones de la tabla corresponden a los registros y las columnas a los datos; éstas tablas son lógicas no físicas, por lo que los datos se almacenan en la forma analizada, y cuando existe un requerimiento de información, el sistema produce una tabla como respuesta.

Según Elmasri y Navathe (2000), en la terminología formal del modelo relacional, una fila se denomina tupla, una cabecera de columnas es un atributo y la tabla se denomina relación. El tipo de datos que describe los tipos de valores que pueden aparecer en cada columna se llama dominio. El dominio es el conjunto de valores atómicos; por atómico, se entiende que cada valor del dominio es indivisible en lo que concierne a este modelo relacional. Ejemplo de dominios, podemos mencionar los siguientes: **id_leg** = el conjunto de números válidos de identificación de los legisladores formado por "*n*" número de dígitos; y **Nombre** = el conjunto de nombres de una persona.

A continuación se presenta un ejemplo de una base de datos relacional en donde cada uno de los cuadros representa una entidad con diferentes campos, en este caso se refiere a información de legisladores, y en donde se encuentran los siguientes campos:

| Legislador | |
|---|---|
| id_leg | Número identificador del legislador |
| Nombre | Nombre del legislador |
| Apellido | Apellido del legislador |
| id_partido | Numero identificador del partido político |
| id_entidad | Número identificador de la entidad representada |
| Eleccion | Tipo de elección en el congreso |
| dist_circ | Número de distrito |
| Camara | Cámara de diputados o de senadores |
| Cargo | Cargo en el congreso |

| Organización |
|---|

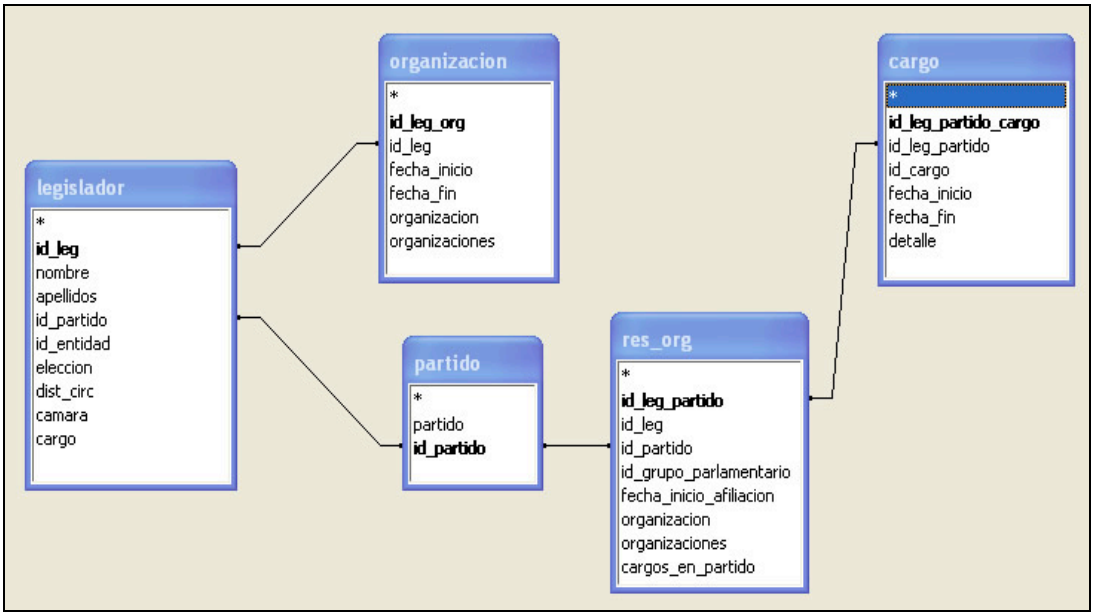
| | |
|---|---|
| id_leg_org | Número identificador del legislador en la organización |
| id_leg | Numero identificador del legislador |
| Fecha_inicio | Fecha de inicio en la organización |
| fecha_fin | Fecha de finalización en la organización |
| Organizacion | Tipo de organización |
| organizaciones | Cargo dentro de la organización |

| Partido | |
|---|---|
| Partido | Nombre del partido político |
| id_partido | Numero identificador del partido político |

| Cargo | |
|---|---|
| id_leg_partido_cargo | Numero identificador del cargo en el partido |
| id_leg_partido | Numero identificador del legislador en el partido |
| id_cargo | Número identificador del cargo |
| fecha_inicio | Fecha de inicio en el cargo |
| fecha_fin | Fecha de finalización en el cargo |
| Detalle | Cargo dentro del partido |

| Res_org= unión de legislador, organización y partido | |
|---|---|
| id_leg_partido | Numero identificador del legislador en el partido |
| id_leg | Numero identificador del legislador |
| id_partido | Numero identificador del partido politico |
| id_grupo_parlamentario | Número identificador del grupo parlamentario |
| fecha_inicio_afiliacion | Fecha de afiliación |
| organización | Tipo de organización |
| organizaciones | Cargo dentro de la organización |
| cargo_en_partido | Número de cargos en el partido |

**Figura 1**




De esta información que se tiene en la base de datos, se integro la siguiente tabla,

**Tabla No. 4**

| id_leg | nombre | apellidos | id_leg_partido_cargo | fecha_inicio | cargo_fecha_fin | detalle |
|---|---|---|---|---|---|---|
| 176 | Blanca Judith | Díaz Delgado | 1006 | 1995-01-01 | 1995-01-01 | Representante del PAN Coahuila ante el Consejo Estatal Electoral (1995). |
| 176 | Blanca Judith | Díaz Delgado | 8004 | 2004-01-01 | 2010-01-01 | Consejero Político Nacional (2004-2010). |
| 176 | Blanca Judith | Díaz Delgado | 8005 | 1995-01-01 | 1997-01-01 | Asesor jurídico en el Grupo Parlamentario del PAN en la LIII Legislatura del H. Congreso de Coahuila (1995-1997). |
| 451 | Ramón | Galindo Noriega | 8004 | 2004-01-01 | 2010-01-01 | Consejero Político Nacional (2004-2010). |
| 451 | Ramón | Galindo Noriega | 8005 | 1995-01-01 | 1997-01-01 | Asesor jurídico en el Grupo Parlamentario del PAN en la LIII Legislatura del H. Congreso de Coahuila (1995-1997). |
| 451 | Ramón | Galindo Noriega | 8227 | 2007-12-10 | 2010-01-01 | Secretario General del Comité Ejecutivo Nacional (CEN) (2007-2010). |
| 393 | René | Arce Islas | 191 | 1996-01-01 | 1999-01-01 | Secretario General del PRD en el Distrito Federal (1996-1999). |
| 393 | René | Arce Islas | 194 | 1990-01-01 | 1993-01-01 | Presidente del Distrito XXVI del PRD en Iztapalapa, (1990-1993). |
| 393 | René | Arce Islas | 1814 | 1989-01-01 | 1989-01-01 | Miembro fundador del PRD (1989). |

Fuente: Elaboración propia, esta tabla presenta una pequeña parte de información de la base de datos de Monitor Legislativo, http://www.monitorlegislativo.org.mx/, consultada el 24 de septiembre de 2009

**Bases de Datos Tipo *Data Warehouses*.**

El término *Data Warehouse* es utilizado por primera vez por Inmon en 1988[10], quien se considera el "padre de los almacenes de datos" (Connolly y Begg, 2005: 1039). Date (2001) señala que los *Data Warehouses* surgen por la necesidad de proporcionar una fuente única de datos, la cual sea consistente y se encuentre lista para la toma de decisiones, y por la necesidad de hacerlo sin afectar los demás sistemas operativos.

Inmon define un *Data Warehouse* como un almacén de datos orientado a un tema, integrado, no volátil, es decir, la información no se modifica ni elimina, se mantiene para futuras consultas; pero los cambios en los datos a lo largo del tiempo se registran para que los reportes puedan mostrar las variaciones, por lo que se considera que es variable en el tiempo; y se utiliza como ayuda en el proceso de toma de decisiones por parte de quien dirige una organización (Date, 2001: 709; Connolly y Begg, 2005: 1039).

Asimismo, Connolly y Begg (2005) explican que los datos son integrados en el *Data Warehouse*, pues reúne información de distintas áreas de aplicación de una organización, con la finalidad de presentarla de forma homogénea, aunque esta información proceda de distintos sistemas de aplicación. Estos autores, también señalan que los datos se consideran no volátiles ya que éstos no se actualizan en forma real sino en forma periódica a partir de sistemas operacionales, por lo que los datos nuevos se añaden para aumentar la base de datos pero no sustituyen la información ya existente.

Por otra parte, se considera que un *Data Warehouse* está orientado a temas por que se organiza con base en aquellos que más relevancia tienen para esa organización o dependencia y no por áreas de aplicación. Por ejemplo en el caso de salud, un *Data*

[10] Date (2001: 727), menciona que el término "data Warehouse" apareció por primera vez en Data Architecture: the information paradigma de W. H. Inmon (1988).

*Warehouse* puede estar organizado por pacientes, personal médico o medicamentos; pero no por el área de aplicación como solicitud de traslados de enfermos, ingresos a pediatría, control de almacén de medicinas, etc.

Sin embargo, muchos tomadores de decisiones no utilizan toda la información que presenta un *Data Warehouse*, sino sólo una parte de ésta, por lo que se consideró necesario construir un tipo de almacén de datos pero limitado a un propósito más específico, donde los usuarios pueden actualizar los datos e incluso crear nuevos datos según su propósito, a esto se le conoce como ***Data Mart***. Éste se define como un "almacén de datos especializados, orientado a un tema, integrado, volátil y variante en el tiempo para apoyar un subconjunto específico de decisiones de administración" (Date, 2001: 710).

Considerando las características y la utilidad que tienen los distintos tipos de bases de datos, se considera que **las bases de datos planas** son las más recomendables para presentar información tanto cualitativa como cuantitativa de cualquier estudio o investigación en ciencias sociales, debido a que éstas permiten capturar la información en cualquier hoja de cálculo (*Excel*, *Calc*, *Lotus 1-2-3*) sin necesidad de tener algún *software* especializado. Asimismo este tipo de bases de datos permite mostrar todas las variables que la integran y en el orden que el instrumento de investigación tiene establecido (cuestionario o entrevista); y finalmente proporciona la ventaja de que ya están listas para ser utilizadas por cualquier paquete estadístico como *SPSS* o *Stata*, y obtener análisis estadísticos simples o más sofisticados: regresiones lineales, correlaciones, tablas de contingencia, frecuencias, etc.

Por otra parte, la base de datos plana, de forma similar a la relacional, le proporciona al investigador flexibilidad al utilizar la información, pues si éste sólo requiere una parte de la información que contiene la base de datos, puede muy fácilmente

seleccionar sólo esas variables para estudiarlas y no el total. También con este tipo de base de datos el usuario puede agregar a dicha vista otras variables, las cuales pueden ser, por ejemplo, las mismas variables pero de otro periodo de tiempo o variables adicionales que pueden ayudarle a realizar análisis de causalidad o asociación como pueden ser: población, producto interno bruto, marginación, etc., las cuales pueden ser usadas tanto como variable independientes como de control.

### 4.3 Pasos para crear una base de datos

De acuerdo con Duffy (2000: 12) para crear una base de datos, primero se debe definir la estructura de la tabla, que consiste en especificar el nombre, los tipos de datos (texto, numérico, fecha, notación científica, etc.) y su longitud, sus valores, así como las estructuras y restricciones para los datos que se van a almacenar en dicha base. Una vez que la estructura de la tabla esta creada, entonces ya se pueden ingresar los datos, es decir, se construye la base de datos, y se almacenan los datos concretos sobre algún medio de almacenamiento controlado por el SGBD. Posteriormente, se lleva a cabo el proceso de manipulación de la base de datos, el cual incluye funciones tales como: consultar la base de datos para recuperar unos datos específicos, en su caso, actualizar la base de datos para reflejar los cambios ocurridos, y generar informes a partir de los datos (Elmasri y Navathe; 2000: 5).

Elmasri y Navathe, (2000: 8) señalan que una característica fundamental del enfoque de base de datos es que éste no solo contiene la base de datos sino también una definición o descripción completa de la estructura de ésta y de sus restricciones; esta definición se almacena en un catálogo del sistema y se llama meta-datos.

Para fines de este documento, se tratará sobre aquellas bases de datos planas que llamaremos *"*sencillas ó básicas*",* que cualquier investigador puede desarrollar, ya sea utilizando un paquete estadístico (*software* como *SPSS, Excel, Access*) o de forma manual. Debido a la gran necesidad que se tiene para utilizar y analizar la información obtenida mediante cualquier técnica de investigación, enseguida se expondrá las distintas etapas a través de las cuales se va ir desarrollando una base de datos de acuerdo a la técnica de recopilación de información.

Para diseñar las bases de datos se encuentran dos técnicas principales, las cuales se denominan "de abajo a arriba" y "de arriba a abajo". La primera técnica resulta apropiada para el diseño de bases de datos simples que tienen un número relativamente pequeño de atributos. Esta técnica empieza en el nivel fundamental de atributos o propiedades de las entidades y relaciones, que mediante el análisis de las asociaciones entre esas propiedades se agrupan para formar relaciones que representan tipos de entidades y relaciones entre éstas de un nivel superior. La técnica "de arriba a abajo" es más recomendable para bases de datos complejas, y ésta comienza con el desarrollo de modelos de datos que contienen pocas entidades y relaciones de alto nivel, posteriormente se identifican los atributos o propiedades de estas entidades y relaciones de nivel inferior y así sucesivamente (Connolly y Begg, 2005: 266).

Los objetivos que se pretenden alcanzar con el diseño de las bases de datos están dirigidos, por un lado, a satisfacer los requisitos de contenido de información de los usuarios y aplicaciones especificadas, así como los de procesamiento y de rendimiento; y por el otro, a proporcionar una estructuración de la información de forma natural y fácil de entender (Elmasri y Navathe, 2000: 12).

Las personas que diseñan bases de datos tienen que identificar los datos que almacenará en éstas y elegir las estructuras para presentar y almacenar esos datos. Por tanto los diseñadores tienen la responsabilidad de comunicarse con todos los posibles usuarios[11] de la base de datos, con la finalidad de comprender sus necesidades y de presentar un diseño que las satisfaga (Elmasri y Navathe, 2000: 12). Para diseñar una base de datos se tienen que realizar las siguientes actividades (Elmasri y Navathe, 2000; Connolly y Begg, 2005):

- **Obtención y análisis de requisitos.** En esta fase del proceso de diseño se identifican principalmente las áreas de aplicación y grupos de usuarios que utilizarán la base de datos; esto permitirá obtener un conjunto de requisitos del usuario de manera más detallada. También se debe especificar los requisitos funcionales de las aplicaciones que incluye el análisis del tipo de transacciones y del flujo de información dentro del sistema.
- **Diseño conceptual.** Esta etapa tiene como objetivo entender completamente la estructura, el significado (semántica), las relaciones y las restricciones de la base de datos; por lo cual este esquema conceptual es valioso para tener una descripción estable del contenido de la base de datos.
- **Diseño lógico.** Esta fase consiste en implementar la base de datos usando un sistema de gestión de bases de datos (SGBD). La mayoría de los SGBD disponibles en el mercado utilizan un modelo de datos de implementación, como el relacional o el

[11] Existen cuatro categorías de usuarios finales: los ocasionales, que como su nombre lo indica acceden de vez en cuando a la base de datos, pero pueden requerir información distinta en cada ocasión; simples, que son los usuarios más constantes de la base de datos, y sus consultas son más estandarizadas o programadas; usuarios finales avanzados y usuarios autónomos (Elmasri y Navathe, 2000: 12).

orientado a objetos, así que el esquema conceptual se transforma del modelo de datos de alto nivel en el modelo de datos de implementación.

- **Diseño físico.** En este paso se especifican las estructuras de almacenamiento internas, el acceso eficiente a los datos y las organizaciones de los archivos de la base de datos. Para realizar el diseño físico se utilizan criterios como el tiempo de respuesta a una transacción realizada de la bases de datos; así como aprovechamiento del espacio, y a la productividad de las transacciones que el sistema pueda procesar.

### 4.4 Características y requisitos básicos de las bases de datos

Una base de datos debe contener al menos los siguientes elementos: nombre de la base de datos; cuestionario(s) con las preguntas originales ó con la guía de entrevista (sí es el caso); resumen de la base de datos: que contenga la información necesaria acerca del conjunto de datos, tales como: número de variables, número de registros, etc. Por ejemplo, en el caso de entrevistas se debe incorporar un listado de las personas que fueron entrevistadas.

También es necesario el libro de códigos de la base de datos, éste es una relación de todas las variables que se encuentran en la tabla, con una descripción (explicación de cada una de las variables), incluso se mencionan sus sinónimos; también incluye un listado de todos los valores que podría tomar cada una de las variables, y de las puntuaciones numéricas vinculadas a cada uno de esos valores (Manheim y Rich; 1988).

Por ejemplo, si la variable es estatus social, en la columna aparece abreviada como "estasoc", para identificarla se debe establecer que ésta se refiere a estatus social, y qué comprende las diferentes categorías de los miembros de una sociedad, las cuales se definen en términos de variaciones socioeconómicas como puede ser el ingreso familiar; sus posibles valores son 2 = Alto, 1= Medio, 0= Bajo (ver Tabla No. 5).

**Tabla No. 5**
**Ejemplos de libro de códigos.**

| NOMBRE DE LA VARIABLE O PREGUNTA | VALORES DE LAS POSIBLES RESPUESTAS QUE PUEDEN TENER CADA VARIABLE O PREGUNTA |
|---|---|
| FOLIO = Número de cuestionario | 1, 2, 3… / 0001, 0002, 0003; según la numeración que hayan estipulado para el folio del cuestionario, encuesta ó entrevista; y/o dependiendo del tamaño de la muestra. |
| SEXO | (1.00) Masculino<br>(2.00) Femenino |
| P26 = Cuando era niño(a), ¿su madre trabajaba? | (1) Sí<br>(2) No<br>(3) A veces<br>(99) No sabe/No contesto |
| JUST8 = ¿Qué tan honrados cree usted que son los funcionarios de la Justicia en México? | (0) No aplica<br>(1) Nada honrados<br>(2) Poco honrados<br>(3) Algo honrados<br>(4) Muy honrados<br>(98) NS (No sabe)<br>(99) NC (No contesto) |
| R24= ¿Con qué periodicidad se "actualiza" la Información en la página de Internet? | Este reactivo se debe responder como catálogo, es decir, los valores que se asignarán serán los siguientes:<br>1 = Cada semana<br>2 = Cada 2 semanas<br>3 = Cada mes<br>4 = Cada 3 meses<br>5 = No existe un periodo determinado<br>96 = comentario: es muy variada la periodicidad<br>97 = no respondió |
| P19= Durante lo que va del gobierno del Presidente Fox. ¿Qué tanto cree que han cambiado las cosas en el País? | (.00) No sabe / No responde<br>(1.00) Nada<br>(2.00) Poco<br>(3.00) Regular<br>(4.00) Mucho |
| VOTOS EN CONTRA | En algunas bases de datos como ésta, lo que se da como valor es el concepto de dicha variable, en este |

| | |
|---|---|
| | caso:<br>Votos en contra: Es la decisión de los ministros para no apoyar el proyecto propuesto por el ministro ponente. |

Fuente: Elaboración propia con base en distintos libros de códigos de las bases de datos que se encuentran en el acervo del Banco de Información para la Investigación Aplicada en Ciencias Sociales (BIIACS) del CIDE.

Además, con la intención de que la información de la base de datos sea lo más clara y fácil de utilizar, cada fila de la base de datos debe corresponder a un registro ó caso y cada columna a una variable. También se deben de proporcionar las instrucciones adicionales o aclaraciones sobre las particularidades o los problemas que presenta la base de datos y que se considera es muy relevante que el usuario potencial las conozca. Y, finalmente, se sugiere que el *software* en el que se desarrolle la base de datos sea de los más comunes en el medio académico y que se pueda tener acceso a él fácilmente, como pueden ser los "*open source*".

## 5.- Aspectos a considerar en la Conceptualización de Indicadores.

Como se mencionó en la recolección de información de la necesidad de asignar un valor ó calificación a un dato en el procesamiento de la información, y considerando que éste es el insumo primordial de una base de datos, se expondrán en este apartado algunas consideraciones sobre indicadores.

### 5.1 Concepto y tipos de conceptos.

Los conceptos son los componentes básicos de la teoría y representan los puntos alrededor del cual se lleva a cabo la investigación (Bryman, 2001: 65). Bunge (2000: 627) distingue cuatro géneros de conceptos: individuales que se aplican a los conceptos cuyos referentes

son individuos y no son cuantitativos (cualitativos) a menos que consistan en individuos numéricos; de clases se aplican a conjuntos de individuos, y pueden atribuirse números de un modo convencional, por ejemplo: un concepto de clase como "mujer" da origen al predicado "es una mujer", si un determinado individuo es efectivamente una mujer se puede simbolizar el hecho escribiendo la cifra 1 y si no es mujer escribir 0, es decir, que pueden tomar uno de dos valores, presencia o ausencia de la propiedad correspondiente.

Los conceptos relacionales se aplica a relaciones entre objetos (individuos o conjuntos) de algún género; y se dividen en conceptos no comparativos y comparativos, donde los primero solo pueden cuantificarse de modo nominal (convencional) y los comparativos permiten ordenar conjuntos. Los conceptos cuantitativos (magnitudes) implican una cuantificación en sentido estricto.

Lazarsfeld (1979: 35) señala que ninguna ciencia aborda su objeto en su plenitud sino que seleccionan determinadas propiedades de su objeto e intenta establecer entre ellas relaciones recíprocas. Estas propiedades reciben el nombre de aspectos, atributos, componentes, dimensiones o el término matemático de "variables". Cea (1998) menciona que la generalidad sobre los conceptos es que constituyen variables latentes, hipotéticas, que no se observan directamente; por lo que es necesario traducir el concepto teórico a indicadores o variables empíricas que permitan medir las dimensiones o aspectos enmarcados en el concepto. Bunge (2000) establece que se llamará cuantificación numérica al procedimiento a través del cual ciertos conceptos se asocian con variables numéricas. Destacando que hay varias clases de conceptos por lo que es factible que se encuentren también varias clases de cuantificación.

### 5.2 ¿Porqué cuantificar?

Por su parte Piaget (1970), señala que la medida consiste en una aplicación del número a los datos que han de evaluarse. El autor argumenta que si se recurre al número no es en virtud del prestigio de las matemáticas o a algún perjuicio en favor de la cantidad, sino que "el valor instrumental del número proviene del hecho de que constituye una estructura mucho más rica que la de las propiedades lógicas de que se compone: por una parte, la inclusión de clases, que preside los sistemas de clasificación, y, por otra, el orden, que caracteriza las seriaciones". Por lo que el número presenta una riqueza y una movilidad que hacen que sus estructuras sean particularmente útiles en todas las cuestiones de comparación (Piaget, 1970: 80).

Bunge (2000: 634) enumera una lista de ventajas de la cuantificación: 1) permite el afinamiento de los conceptos, 2) proporciona una descripción precisa, 3) precisa la clasificación del sujeto u objeto, 4) permite la formación de hipótesis y teorías exactas estableciendo interrelación precisa entre variables y formulas, y 5) contrastar hipótesis y teorías. Además, este mismo autor señala que no es posible cuantificar todo concepto como son los individuales y las relaciones no comparativas. Sin embargo, si es deseable cuantificar lo que es aún cualitativo pero que no cae dentro de los conceptos anteriores (individuos y relaciones no comparativas). Para tal fin se debe realizar un planteamiento y descripción del fenómeno u objeto de investigación de manera profunda y detallada, e intentar asociarlo con alguna propiedad o conjunto de propiedades cuantitativas; concluyendo que no es el tema, la materia, sino las ideas sobre ellos las que son objeto de cuantificación numérica, con lo que se pueden superar la mayoría de las barreras o resistencias a la cuantificación (Bunge, 2000: 635).

Por otra parte, Cea (1998: 126) establece que la medición de una variable consiste en asignar valores o categorías a las distintas características del objeto de estudio; pero para

que esta medición se realice correctamente debe cumplir con tres requisitos: 1) Exhaustividad: que la medición comprenda el mayor número posible de atributos (dimensiones, aspectos, etc), de tal forma que ninguna observación quede sin clasificarse, como es el caso de las opciones de respuesta para los cuestionarios tales como: "otros", "no sabe" y "no contesto". 2) Exclusividad: que se refiere a que los atributos que componen la variable deben ser mutuamente excluyentes, para que se pueda clasificar cualquier observación en un solo atributo; y 3) Precisión: realizar el mayor número de distinciones posibles lo que contribuye a la generación de información más precisa.

**Niveles de medición.**

Cortés y Rubalcava (1985: 3) y Babbie (1988: 174) señalan que para medir los conceptos, se encuentran cuatro niveles de medición, que se clasifican en tres clases: comparativo sin orden (la escala nominal), comparativo con orden (la escala ordinal) y el cuantitativo (escala de intervalo y de razón).

Las mediciones nominales simplemente distinguen las categorías de una variable determinada y estas categorías deben ser mutuamente excluyentes. Por lo que se refiere a las mediciones ordinales, éstas establecen un orden de rango entre las categorías que integran una variable, aunque estas mediciones son representadas por números, éstos solo indican el orden entre las variables (Babbie, 1988: 175). Por lo que se refiere a las mediciones de intervalo, éstas utilizan números para describir condiciones pero a diferencia de las ordinales, las distancia entre un punto a otro si tiene un significado real; es decir no sólo se usa el nombre y el orden de los números sino también su magnitud. Finalmente, las

mediciones de razón se diferencian de las de intervalo por la característica adicional de contar con un "verdadero cero" que representa ausencia de la variable (Cortés y Rubalcava, 1985: 12; Babbie, 1988: 176).

**5.3 Proceso de operacionalización del concepto.**

Lazarsfeld (1979: 36) establece un proceso de cuatro fases, a través del cual se permite expresar los conceptos en términos de indicadores empíricos o asignar mediciones a los conceptos[12]: 1) la representación literaria del concepto; 2) la especificación de las dimensiones o atributos que se consideran partes constitutivas de un concepto (Munck y Verkuilen, 2002) y que dichos componentes pueden ser deducidos analíticamente a partir del concepto que las engloba o empíricamente a partir de la estructura de sus intercorrelaciones; 3) la elección de los indicadores observables de las dimensiones definidas en la etapa anterior; y 4) la síntesis de los indicadores o elaboración de índices.

Con respecto a la segunda fase del proceso de operacionalización del concepto, Lazarsfeld, Munck y Verkuilen (2002: 413) consideran que es importante señalar cómo se relacionan entre sí los diferentes atributos ó dimensiones del concepto, y más específicamente, establecer los pasos a seguir para garantizar la organización vertical de los atributos según el nivel de abstracción (atributos ó dimensiones ó variables, componentes de los atributos, y subcomponentes de atributos, etc). Para tal fin, estos autores señalan que para la agregación de atributos se presupone que éstos están organizados cumpliendo dos reglas básicas de la lógica conceptual: primero, que los atributos menos abstractos (componentes de los atributos) se ubiquen inmediatamente subordinados del atributo más abstracto (atributo) al cual le da cuerpo y hacen más concreto; y segundo, los atributos que

[12] Algunos autores como Cea (1998) define este proceso como operacionalización de los conceptos teóricos.

están en el mismo nivel de abstracción debe ser mutuamente excluyentes del atributo en el nivel de abstracción inmediato superior.

**Indicador.**

El indicador es algo que es ideado o ya existe, y se emplea como una medida de un concepto (Bryman, 2004: 67). Estos indicadores pueden ser directos ó indirectos, dependiendo de su relación con el concepto que miden, por ejemplo: los datos de una pregunta en una encuesta sobre la cantidad que una persona gana al mes podría ser una medida directa del ingreso personal, pero si lo tratamos como un indicador de clase social, éste sería una medida indirecta.

Hay diversas maneras en las que un indicador puede ser ideado o construido. Una manera es a través de una ó varias preguntas que son parte de un cuestionario ó de una guía de entrevista estructurada, que podrían ser una actitud o alguna situación social. Otra es mediante el registro de las conductas de los individuos usando un programa de observación estructurado. Una más es utilizando información de estadísticas oficiales tales como las del INEGI o mediante la observación de los medios de comunicación a través del análisis de contenido.

Un cuestionamiento que siempre se encuentra presente en los investigadores sociales es sí un indicador es suficiente para medir un concepto. Existen varios motivos para sugerir que se utilicen múltiples indicadores, pues se presentarían varios problemas al usar un solo indicador. En primer lugar, es posible que un solo indicador clasifique incorrectamente muchos individuos, pero si existen varios indicadores y si la gente u objeto de estudio fue clasificado erróneamente a través de una pregunta en particular, esto puede contrarrestar sus efectos. En segundo lugar, un indicador puede capturar solo una parte o

aspecto del concepto o ser muy general, en este caso una sola pregunta puede necesitar ser de un nivel excesivamente alto de generalidad y por lo tanto no puede reflejar la verdadera situación de las personas que respondieron a la misma. Y por último, con la utilización de múltiples indicadores se pueden realizar distinciones más finas y análisis más sofisticados (Bryman, 2001: 69).

Los indicadores pueden ser de insumos (unidades de trabajo, capital, bienes y servicios invertidos para la producción, en el caso de gobierno, de servicios públicos), de procesos (estructuras, procedimientos y arreglos administrativos), de resultados (número de beneficiarios de algún programa público) o de impacto (el impacto que tienen los servicios en el usuario final).

**Integración del índice.**

Por lo que corresponde a la integración del índice o el resumen de indicadores, esta etapa se lleva a cabo después de que se tienen los diferentes valores que miden las distintas dimensiones que integran el concepto, y para desarrollarlo se realizan algunas operaciones aritméticas (Cea, 1998; Neupert, 1977). Pero para tal fin es necesario que las distintas medidas se transformen en una escala de medición común, con la finalidad de facilitar su agregación. Para tal efecto se debe realizar una ponderación, es decir, asignar "pesos" a los distintos valores que presentan los indicadores con la finalidad de expresar diferencias, en caso de existir, en la importancia relativa de cada uno de los indicadores que van a integrar el índice (Cea, 1998: 139).

Un ejemplo de lo que se acaba de mencionar es el siguiente: si consideramos que los componentes (indicadores) de un concepto tienen el mismo peso, lo que podemos hacer es realizar la operación aritmética de sumar; pero si consideramos que cada uno de los

componentes tiene un peso mayor a otro, la operación que se realiza es la de multiplicar (Muck y Verkuilen, 2002: 430). Cabe mencionar que Babbie (1988: 313) sugiere que se deben ponderar por igual todos los componentes de un concepto a menos que existan razones poderosas para establecer ponderaciones diferenciales.

Los procedimientos más usados para integrar variables operacionales para constituir un índice son: índices sumatorios e índices aritméticos. El primer método se puede utilizar con cualquier tipo de variables y consiste en asignarle valores numéricos (generalmente de manera arbitraria) a las categorías que presentan, entre los valores de las variables operacionales identificadas en el constructo teórico. Después se suman los valores de cada combinación posible de categorías para obtener una escala numérica que a su vez se vuelve a categorizar; en el caso de que existan variables con valores cuantitativos, éstos se van a categorizar para darle a las categoría formadas los mencionados valores arbitrarios, como en el caso de ingresos se establecen intervalo que van de "x" a "y" monto pero que se denomina alto, de "y" a "z" medio, etc. (Neupert, 1977: 56).

Ejemplo: suponemos que los valores de todas las variables operacionales (ocupación, educación e ingreso) que miden el concepto de Nivel Socioeconómico, se han reducido a tres categorías: alto, medio y bajo; en donde la escala numérica puede variar entre 0 (la posición más baja) y 6 (el nivel más alto).

**Tabla 6**

| Nivel ocupacional | Nivel educacional | Ingresos |
|---|---|---|
| Alto (2) | Alto (2) | Alto (2) |
| Medio (1) | Medio (1) | Medio (1) |
| Bajo (0) | Bajo (0) | Bajo (0) |

Fuente: elaborado con base en Neupert, Ricardo (1977). **Manual de investigación social**, Editorial Universitaria, Honduras.

El método de índices aritméticos se utiliza cuando las variables operacionales presentan valores en término de magnitudes numéricas; y se usa principalmente cuando el significado de cada una de las variables identificadas, por sí sola, no aporta nada al significado de la variable teórica, esto sólo se puede conseguir con la combinación de dichas variables. Generalmente se obtiene el índice realizando una división de una variable entre otra. Por ejemplo: Grado de hacinamiento = Número de camas / número de miembros de la familia.

Cea (1998) agrega un paso adicional a los establecidos por Lazarsfeld, el que se refiere a comprobar hasta qué punto la medición de los conceptos teóricos reúne las condiciones mínimas de validez y fiabilidad. De acuerdo con esta autora, los indicadores antes de ser fiables deben ser válidos pues deben de proporcionar una representación adecuada del concepto que miden y esto es independiente de si reúne o no las condiciones de fiabilidad. Enseguida se expone más detalladamente estos aspectos de la medición.

**5.4 Validez y fiabilidad en la medición del concepto.**

De acuerdo con Bryman (2001) nos dice que la validez se refiere a que tanto un indicador (o serie de indicadores) que esta ideado para medir un concepto, realmente lo mide. Algunos autores que han escrito sobre la validez de la medición distinguen entre distintos tipos de validez**: validez de criterio** que se comprueba comparándola con algún criterio ó medida que se haya utilizado anteriormente para medir el mismo concepto; demostrando que la nueva medida clasifica u ordena el mismo concepto de igual forma que otros indicadores alternativos de éste. Se encuentran dos variedades: **concurrente** (se correlaciona el nuevo indicador con un criterio adoptado en un mismo momento) y **predictiva** (usa un criterio futuro que esta correlacionado con la medida).

Otro procedimiento es el de **validez convergente** el cual establece que una medida debe ser comparada con las medidas del mismo concepto desarrolladas a través de otros métodos; y la **validez aparente** es decir que la medida refleja de forma fehaciente el contenido del concepto (Bryman, 2001: 73). Cea (1998: 151) agrega otros dos tipos de validez: **de contenido** (se refiere al grado en que una medición empírica cubre la variedad de significados incluidos en un concepto) **y de constructo** (que compara una medida particular con aquella que teóricamente se habría de esperar).

Por lo que respecta a la fiabilidad de la medición, se refiere a la consistencia de la medida de un concepto. Hay tres factores importantes involucrados para considerar si una medida es fiable: la estabilidad en el tiempo (que los resultados obtenidos en un indicador no cambia con el tiempo ó es muy pequeña la variación), la fiabilidad interna (que los indicadores, escala ó índice son consistentes, es decir, si la puntuación de los respondientes en cualquier indicador tiende a estar relacionada con sus puntuaciones en otros indicadores), y la consistencia inter-observador (*inter-observer consistency*) que se refiere a situaciones en donde el investigador esta muy involucrado en el registro de observaciones o en la traducción de datos a categorías, sobre todo cuando hay más de un observador que participa en tales actividades, existe la posibilidad de que se tenga poca consistencia en sus decisiones (Bryman, 2001: 71).

### 5.5 Información necesaria para juzgar la fiabilidad y validez de la medición.

Por otra parte, Wright *et al*. (2004)[13] señalan que si bien existen varias normas que pueden ser usadas para determinar si los investigadores reportan adecuadamente la información

[13] Estas recomendaciones las realiza Wrigth *et al* (2004) con base en un estudio realizado sobre una muestra de 143 artículos seleccionados de las principales revistas de administración pública que describen

referente a sus medidas de investigación, se asume que las investigaciones cuantitativas publicadas deberán proveer al menos dos tipos de información para cada medida de investigación. El primer tipo de información consiste en proporcionar información de carácter general como la fuente de sus medidas y el método de recolección de los datos. El segundo tipo de información que debe proporcionarse es la que describe y apoya cómo se realizó la operacionalización de cada variable del estudio.

Muck y Verkuilen (2002: 422) también sugieren que los analistas deben hacer pública información referente a la formación de medidas: las reglas de medición (indicadores, niveles de medición para cada indicador e información detallada), el proceso de medición (fuentes usadas, número de analistas que participan, y pruebas de convergencia de las medidas) y datos desagregados de todo los indicadores.

Además de proporcionar esta información formal o directa de evidencia de validez, los estudios deben proveer información básica que permita a sus lectores o audiencia juzgar indirectamente (validez aparente) qué tan bien las medidas de investigación representan fielmente las variables de estudio. Tres tipos de información pueden ser útiles en este sentido (Wright *et al*., 2004: 752):

- Los investigadores pueden fortalecer la confianza de sus lectores en sus medidas proporcionando una clara explicación de porqué cada medida fue considerada para cada una de sus respectivas variables de estudio.
- También pueden ayudar a los lectores a estar más seguros de que los aspectos más importantes de un constructo u objeto conceptual en particular se abordaron, si se

investigación cuantitativa; donde se encontró que tenían serias debilidades, pues en estas investigaciones no se especificaba claramente la fuente de la medida ó éstas se basaban en conjuntos de datos preexistentes o cuestionarios auto-administrados.

ofrecen descripciones y ejemplos de las propias medidas, pues esto permitirá que los lectores evalúen si los elementos representan la variable de estudio.

- Una tercera manera de proveer evidencia indirecta de validez es describir el contexto de la investigación de tal manera que se asegure al lector que las fuentes potenciales de sesgo se han abordado durante la recopilación de datos; con esto el investigador sostiene la ausencia de sesgo sistemático en las medidas de su investigación.

## 6.- COMENTARIOS FINALES

Las bases de datos representan un instrumento relevante para la toma de decisiones en los distintos tipos de organizaciones (públicas, privadas y académicas). Pero, para el caso de las ciencias sociales, en la investigación y en el diseño, aplicación y evaluación de políticas públicas, la información sobre algún objeto o sujeto social reviste de gran importancia, por lo que las bases de datos representan un insumo básico para estas actividades.

Como ya se mencionó al principio del documento, su finalidad es señalar los elementos básicos que se deben considerar para elaborar y diseñar bases de datos, enfocándose principalmente a las relacionadas con las ciencias sociales. En este documento se destaca que el diseño e integración de las bases de datos varía de acuerdo al método de recolección de información (cuestionario, encuesta, entrevista, ó utilizar información de fuentes secundarias); así como también influye en éstas el alcance del estudio que se están realizando [numero de objetos/sujetos analizados y el número de variables que se están contemplando en el estudio (número de preguntas)]; y el tipo de información que se está recabando (cualitativa y/o cuantitativa). Sugiriendo que en aquellos casos en donde se solicita ambos tipos de información cualitativa y cuantitativa, y en proporciones significativas, lo mejor es elaborar dos bases de datos, una para cada tipo de información.

Cabe señalar que si bien hay diferentes tipos de bases de datos, un tipo de base básico para poder desarrollar las demás (*Relacional, Data Warehouse, Data Mart*) es la tabla plana. Asimismo, se identificaron algunos aspectos que se deben considerar cuando se utilizan datos secundarios para una determinada investigación, así como para integrarlos a una base de datos, tales como: al comparar información se debe verificar que la forma de registro, clasificación y definición de la información sean compatibles, o que en un estudio de serie de tiempo que las normas que regulan el registro permanezcan iguales en ese periodo de tiempo, que son datos que corresponden al mismo nivel de análisis, y si se retoman datos de una encuesta también sean compatibles los requerimientos metodológicos (tamaño de la muestra, alcance, margen de error, nivel de análisis, etc.).

Por otra parte, en este documento se mencionaron algunos aspectos y reglas que se deben contemplar al elaborar indicadores que permitan medir las distintas dimensiones o variables de un concepto, de tal manera que ésta medición pueda cumplir con las condiciones mínimas de validez y fiabilidad. Y, se señalan algunas recomendaciones sobre información que deben divulgar los investigadores que desarrollen indicadores en sus distintas investigaciones con la finalidad de clarificar la validez de las medidas de las variables utilizadas en sus estudios.

Finalmente, en este documento se busco ser ilustrativo sobre las bases de datos, de tal manera de que en él se pueden encontrar distintos ejemplos ó sugerencias de consultas de bases de datos según la estructura y según el tipo de información, e incluso también se muestran diversos libros de códigos (útil para diseñar los instrumentos de recolección de información: cuestionarios y encuestas). Con base en lo antes expuesto, se espera que este documento permita enriquecer el trabajo de investigación en el diseño y uso de bases de datos en ciencias sociales.

## 7.- AGRADECIMIENTOS.

Los autores agradecen los valiosos comentarios de Alejandra Ríos y Guillermo Cejudo a versiones previas de este documento. Este trabajo fue parcialmente financiado por el Programa de las Naciones Unidas para el Desarrollo (PNUD), mediante el proyecto "México Estatal. Calidad de Gobierno y Rendición de cuentas en las Entidades Federativas". Las opiniones, hallazgos y conclusiones son responsabilidad de los autores y no reflejan necesariamente el punto de vista del PNUD.

## 8.- BIBLIOGRAFÍA.


Albo, Andrés, Martínez de Velasco, Alberto, BANAMEX y FACTUM MERCADOTÉCNICO, *Pulso Sociopolítico - 2003* [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [14 septiembre 2009], http://hdl.handle.net/10089/16075

Angell, Robert C. y Ronald Freedman (1990). "El uso de documentos, registros, materiales censales e índices", en Festinger y Katz, ***Los métodos de investigación en las Ciencias Sociales***, Editorial Paidós, México.

Babbie, Earl R. (1988). *Métodos de investigación por encuesta.* Fondo de Cultura Económica, México.

Bergman, Marcelo et al. (2006). *Encuesta a la población en reclusión en el Distrito Federal y Estado de México - 2005* [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [14 julio 2009] http://hdl.handle.net/10089/16085

Bergman, Marcelo *et al*. (2007). *Encuesta de Victimización y Eficacia Institucional (ENVEI)* [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [13 julio 2009], http://hdl.handle.net/10089/3714

Bryman, Alan (2001). *Social Research Methods.* Editorial Oxford University Prees Inc. New York.

Buyens, Jim (2001). *Aprenda desarrollo de bases de datos Web ya*. McGraw – Hill, España.

Caballero, José Antonio y Meneses, Rodrigo. (2006). *Observatorio judicial* [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [14 julio 2009], http://hdl.handle.net/10089/16065

Campbell, A. Angus y George Katona (1990). "La encuesta por muestreo: una técnica para la investigación en ciencias sociales", en Festinger y Katz, *Los métodos de investigación en las Ciencias Sociales*, Editorial Paidós, México.

Cannell, Charles F. y Robert L. Kahn (1990). "La reunión de datos mediante entrevistas", en Festinger y Katz, *Los métodos de investigación en las Ciencias Sociales*, Editorial Paidós, México.

Cartwright, Dorwin P. (1990). "Análisis del material cualitativo", en Festinger y Katz, *Los métodos de investigación en las Ciencias Sociales*, Editorial Paidós, México.

Cea, María de los Ángeles (1998). *Metodología cuantitativa estrategias y técnicas de investigación social***.** Editorial Síntesis, S.A., España.

CIDE. *Banco de Información para la investigación Aplicada en Ciencias Sociales (BIIACS)*, www.biiacs.cide.edu

CIDE. *Estudio Comparativo de los Sistemas Electorales (CSES) – 2003* [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [14 septiembre 2009], http://hdl.handle.net/10089/3687

Coll-Vinent, Robert (1988). *Información y poder. El futuro de las bases de datos documentales.* Editorial Herder, S. A., España

Connolly, Thomas M. y Carolyn E. Begg (2005). *Sistemas de base de datos. Un enfoque práctico para diseño, implementación y gestión*. Pearson Educación, España.

Consejo Nacional de Población. "De la población en México 2005 – 2050", consultada en http://www.conapo.gob.mx/index.php?option=com_content&view=article&id=36&Itemid=199, el 29 de mayo de 2009.

Coombs, Clyde H. (1990). "Teoría y métodos de la medición social", en Festinger y Katz, ***Los métodos de investigación en las Ciencias Sociales***, Editorial Paidós, México.

Cortés, Fernando y Rosa María Rubalcava (1985). *Metodología y técnicas de investigación***.** Serie C: Número 3**,** Editorial FLACSO, México.

Date, C. J. (2001). *Introducción a los sistemas de bases de datos.* Pearson Educación, México.

Duffy, Tim (2000). *Microsoft Access 2000.* Prentice-Hall, Inc., Estados Unidos.

Elmasri, Ramez A. y Sahmkent B. Navathe (2000). *Fundamentos de Sistemas de Bases de datos.* Pearson Educación, España.

Festinger, L. y D. Katz (1990). *Los métodos de investigación en las Ciencias Sociales.* Editorial Paidós, México.

Gil-García, José Ramón (2006). *Enacting state websites: a mixed method study exploring e-government success in multi-organizational settings - Interviews* [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [16 julio 2009], http://hdl.handle.net/10089/16063

Gil-García, José Ramón et al. (2009). *Conducting Web-based surveys of government practitioners in social sciences: practical lessons for e-government researchers*, hicss, pp.1-10, 42nd Hawaii International Conference on System Sciences, USA.

Gomezjara, Francisco y Nicolás Pérez Ramírez (1980). *El diseño de la investigación social.* Editorial Nueva sociología, México.

Goode, William J. y Paul K. Hatt (1980). **Métodos de investigación social**, Editorial Trillas, S. A., México.

Instituto Nacional de Estadística y Geografía. (2000). *Encuesta nacional de ingresos y gastos de los hogares 2000* [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [30 junio 2009], http://hdl.handle.net/10089/16078

Korth, Henry y Abragam Silbeerschatz (1986). *Database System Concepts*. McGraw-Hill, Estados Unidos.

Lazarsfeld, Paul (1979). "De los conceptos a los índices empíricos", en Boudon, Raymond y Paul Lazarsfeld., *Metodología de las ciencias sociales***,** Editorial Laia, S.A. España.

López-Ayllón, Sergio y Arellano Gault, David. (2005). *Transparencia y acceso a la información en los Otros Sujetos Obligados - Cuestionario* [en línea]. Distribuidos por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales: Centro de Investigación y Docencia Económicas. [13 julio 2009], http://hdl.handle.net/10089/16060

Manheim, Jarol B. y Richard C. Rich (1988). *Análisis político empírico. Métodos de investigación en ciencia política*. Alianza Editorial S. A., España.

Munck, Gerardo L. y Jay Verkuilen (2002). "Conceptualizando y midiendo la democracia: una evaluación de índices alternativos", en *Política y gobierno***,** CIDE, México.

Neupert, Ricardo (1977). *Manual de investigación social***.** Editorial Universitaria Tegucigalpa, Honduras.

Piaget, Jean (1973). "Introducción: La situación de las ciencias del hombre dentro del sistema de las ciencias", en Piaget, Jean et al., *Tendencias de la Investigación en las ciencias sociales*. Editorial Alianza/Unesco, S.A., España.

Senn, James (1988). *Análisis y Diseño de Sistemas de Información*. McGraw-Hill, México.

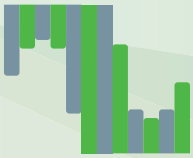
Vilalta Perdomo, Carlos J. y Fernández, Leonel. (2009**).** *Estadísticas Judiciales: Homicidio y Robo en 60 Áreas Metropolitanas de México - 1997, 2000, 2005 y 2007* [en línea]. Distribuido por: México, D.F.: Banco de Información para la Investigación Aplicada en Ciencias Sociales : Centro de Investigación y Docencia Económicas. [9 septiembre 2009], http://hdl.handle.net/10089/16107

Waymire, Richar y Sawtell, Rick (2000). *Aprendiendo Microsoft SQL Server 7.0 en 21 días.* Prentice Hall, México.

Wiederhold, Gio (1985). *Diseño de bases de datos.* McGraw Hill, México.

Wright, Bradley E. *et al* (2009). *Quantitative research measurement in public administration: An assessment of journal publications, Administration & society.* Editorial Sage publications, Estados Unidos.

CARRETERA MÉXICO- TOLUCA 3655
COL. LOMAS DE SANTA FE 01210
MÉXICO, D.F.



www.mexicoestatal.cide.edu