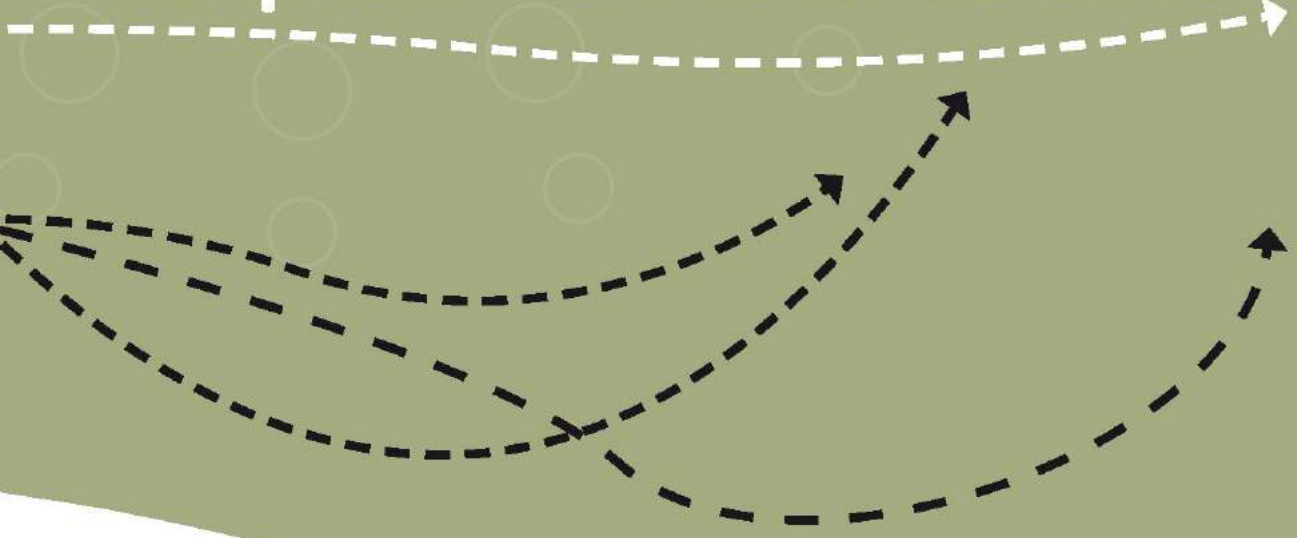
Serie de Cuadernillos Técnicos

Limpieza de bases de datos



Gobierno de Guatemala
Ministerio de Educación

Dirección General de Evaluación
e Investigación Educativa
-Digeduca-

Licenciada Cynthia del Aguila Mendizábal
Ministra de Educación

Licenciada Evelyn Amado de Segura
Viceministra Técnica de Educación

Licenciado Alfredo Gustavo García Archila
Viceministro Administrativo de Educación

Doctor Gutberto Nicolás Leiva Alvarez
Viceministro de Educación Bilingüe e Intercultural

Licenciado Eligio Sic Ixpancoc
Viceministro de Diseño y Verificación de la Calidad Educativa

Directora Lcda. Luisa Fernanda Müller Durán

Subdirección de Análisis de Datos
Autoría
Lcda. María José Castillo Noguera
M.A. José Adolfo Santos Solares

Revisión de texto y diagramación
Lcda. María Teresa Marroquín Yurrita
Diseño de portada
Lic. Roberto Franco Arias

Dirección General de Evaluación e Investigación Educativa

Para efectos de auditoría, este material está sujeto a caducidad.
Para citarlo: Castillo, M. y Santos, J. (2015). *Limpieza de bases de datos.* Guatemala: Dirección General de Evaluación e Investigación Educativa, Ministerio de Educación.
Disponible en red: http://www.mineduc.gob.gt/Digeduca
Impreso en Guatemala
divulgacion_digeduca@mineduc.gob.gt
Guatemala, 2015

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

I. Resumen

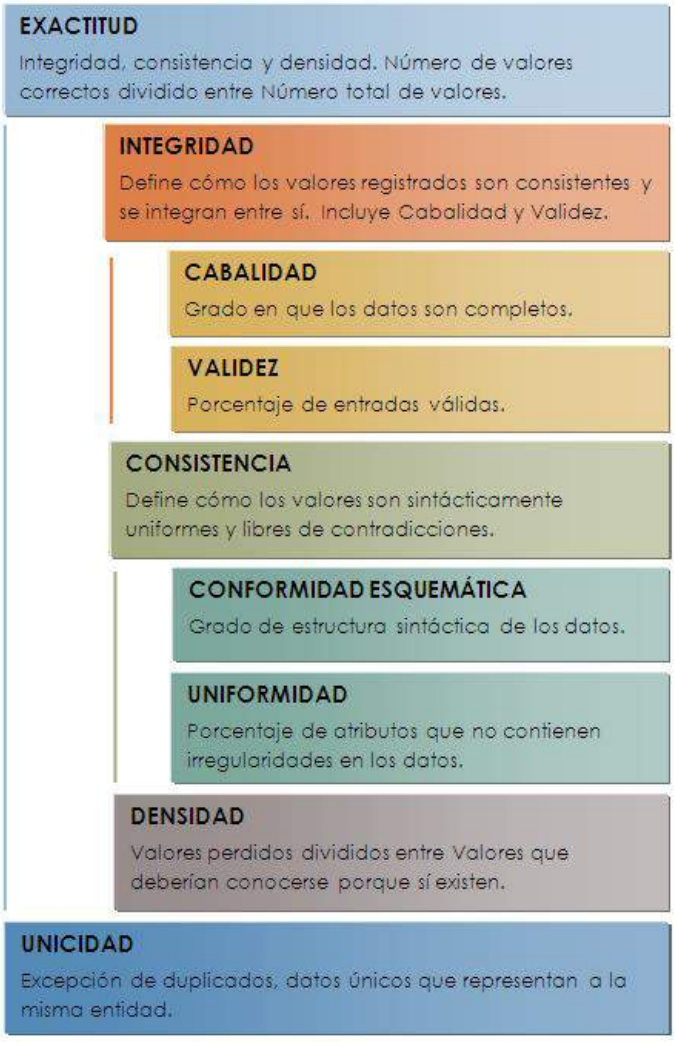
La limpieza de bases de datos constituye una fase fundamental para el posterior análisis y estudio de la información. La comprensión y la generación de conocimiento se verán afectadas si la calidad de los datos no se asegura inicialmente mediante este proceso. El presente documento se llevó a cabo con el objetivo de facilitar una orientación para la realización de la limpieza de las distintas bases de datos que se manejan en la Dirección General de Investigación y Evaluación Educativa –Digeduca–. Se plantea un panorama general sobre cómo se define este proceso, cuál es su importancia y algunos elementos a considerar. Si bien en el día a día se presentarán situaciones específicas, se ofrecen ejemplos que, en términos generales, pueden facilitar lineamientos o guías de acción.

II. Bases de datos y calidad de la información

Para que las bases de datos sean útiles y puedan ser procesadas, analizadas e interpretadas de manera eficaz y eficiente, requieren que los datos que las conforman sean los correctos.

Más que hablar de datos precisos o exactos, se habla de datos de calidad, los cuales cumplen con una serie de criterios que dan un valor agregado a la información. Entre estos se encuentran: exactitud, integridad, cabalidad, validez, consistencia, uniformidad, densidad y unicidad –ver Figura 1– (Ahmed & Aziz, 2010); (Müller & Freytag, 2003). Pipino, Lee y Wang (2002) agregan que la calidad de los datos es un concepto multidimensional que abarca accesibilidad, cantidad apropiada de información, credibilidad, datos completos, representación concisa, representación consistente, facilidad de manipulación, libertad de error, interpretabilidad, objetividad, relevancia, reputación, seguridad, datos oportunos, facilidad de comprensión, grado en que los datos son benéficos y grado en que la utilización de los datos provee ventajas.

Figura 1. Criterios de calidad de los datos



Modificado de: Ahmed & Aziz, 2010; Müller y Freytag, 2003.

Cuando los datos no cumplen con criterios de calidad, se consideran "sucios" debido a que cualquier información inválida, inconsistente, incompleta o incorrecta, atenta contra la integridad de cualquier tipo de base de datos (Kitlas, 2012). Aunque son difíciles de medir, los costos de datos erróneos o inexactos son generalmente mayores a los costos de ingreso de la información, sobre todo cuando se procesan y convierten en conocimiento para la toma de decisiones (Ahmed & Aziz, 2010).

Es un hecho que en los conjuntos de datos, particularmente en los que son de gran tamaño, se presentan errores frecuentemente. Aun cuando se toman medidas para evitar al máximo datos erróneos en el registro y en la adquisición de la información, las tasas de error en las bases de datos se encuentran generalmente entre el 5 % o más (Maletic & Marcus, 2000). Se ha encontrado que hasta el 40 %

de los datos recolectados pueden estar sucios de una u otra manera (Maletic & Marcus, 2010). La evaluación de la calidad de los datos puede ser utilizada para cuantificar la necesidad del proceso de limpieza para una base específica. Asimismo, puede ser utilizada para la optimización del proceso de limpieza, ayudando a determinar criterios prioritarios y facilitando información para que, una vez finalizado, se pueda estimar el éxito del proceso en sí (Ahmed & Aziz, 2010).

III. Proceso de limpieza en las bases de datos

Se denomina limpieza de datos al conjunto de operaciones que se llevan a cabo para determinar información inexacta e incompleta, eliminar anomalías, corregir errores detectados y omisiones en las bases de datos(Ahmed & Aziz, 2010); (Müller & Freytag, 2003). La Figura 2 hace referencia al lenguaje técnico que se utiliza en la preparación de bases de datos.

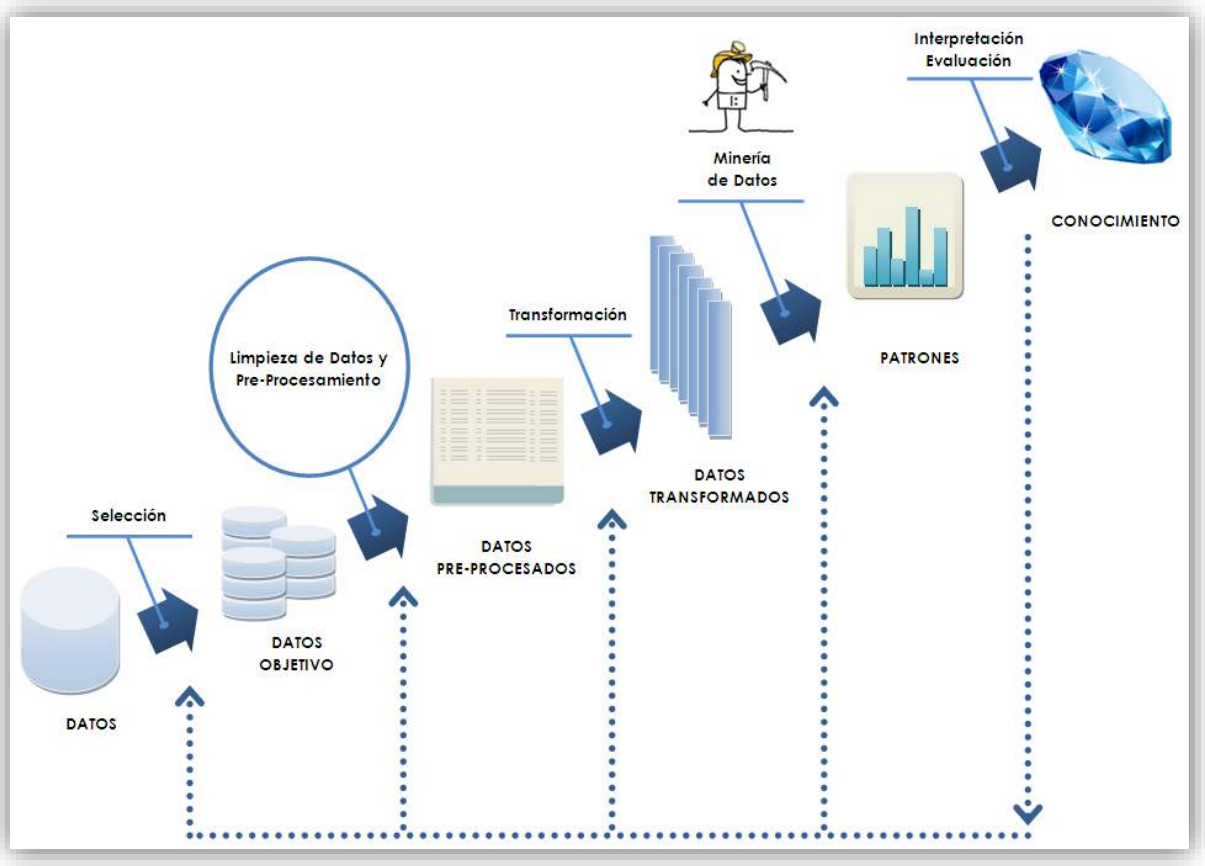
Figura 2. Limpieza de datos

Fuente: Information Builders, 2011

Las rutinas de limpieza, pueden considerarse procesos de mejora de calidad, ya que intentan completar valores faltantes y disminuir, mediante la corrección de inconsistencias, datos sucios o valores que generan ruido(Han, Kamber, & Pei, 2012). Por otro lado, estas rutinas también se definen como el primer paso del proceso de descubrimiento de conocimiento en las bases de datos o proceso KDD –por sus siglas en inglés, *Knowledge Discovery in Database*–. Bajo esta perspectiva se enfatiza la importancia de la limpieza de datos según el principio «basura entra, basura sale», ya que cualquier información incorrecta, generará resultados y estadísticas erróneas e imprecisas(Maletic & Marcus, 2010).

Normalmente, la limpieza y la integración de datos se llevan a cabo como una fase de preprocesamiento. La integración, que consiste en la combinación de datos procedentes de distintos almacenes, ayuda también a reducir y evitar redundancias e inconsistencias en el conjunto de datos resultante (Han et al., 2012).

Figura 3. Proceso de descubrimiento de conocimientos en bases de datos



Modificado de: Fayyad, Piatetsky-Shapiro, & Smyth, 1996

La limpieza de datos es mucho más que una actualización de la base con elementos correctos; se debe observar el objetivo por el cual se hace esto (se muestra en la Figura 3). Implica también descomponer y volver a ensamblar cierta información (Maletic & Marcus, 2000). Van den Broeck, Argeseanu Cunningham, Eeckels y Herbst (2005) plantean que tres fases componen el proceso de limpieza de datos: exploración, diagnóstico y edición. Estas fases se retroalimentan entre sí en ciclos que pueden repetirse. En la exploración o cribado se procura identificar falta o exceso de información, inconsistencias, datos extremos, patrones extraños y resultados sospechosos. Luego se busca identificar la naturaleza inquietante de puntos, patrones y estadísticas. Se diagnostican errores, datos perdidos, datos extremos verdaderos, datos normales verdaderos y se señala o marca información a observar. A partir de ello, se editan las bases de datos corrigiendo, borrando o dejando sin cambio los datos; se retroalimenta el sistema.

El marco general del proceso de limpieza de datos implica:

- ✓ Definir y determinar errores.
- ✓ Buscar e identificar casos de error.
- ✓ Corregir los errores.
- ✓ Documentar casos y tipos de error.
- ✓ Actualizar mecanismos de entrada para evitar errores futuros (Ahmed & Aziz, 2010).

Es importante recordar que la limpieza trata con problemas de datos una vez que estos han ocurrido. No compensa problemas de muestreo, diseños metodológicos inadecuados, ni pobres objetivos de análisis. Si se identifica que parte de las deficiencias de la base de datos se relacionan con procesos anteriores al ingreso de la información, también será importante retroalimentar a ese nivel y documentarlo para que se tome en cuenta en las siguientes fases del proceso de generación de conocimiento.

La documentación de las rutinas de limpieza es parte del manejo transparente y apropiado de datos. Idealmente debería contarse con un registro de herramientas de tamizaje, procedimientos y diagnósticos utilizados para distinguir errores de valores verdaderos, formas de señalar o marcar elementos a revisar, reglas de decisión aplicadas para la edición de los datos, tipos y tasas de error encontrados, tasas de eliminación de valores y de corrección –al menos para las principales variables–, justificación de cambios, impacto en los resultados al quitar valores extremos o fuera del patrón esperado, fechas de trabajo y personal involucrado (Van den Broeck et al., 2005).

La limpieza de datos requiere al menos a una persona o a un equipo para que lea y verifique la exactitud de la información. Estas personas deberían ser expertas en el campo específico del conjunto de datos, para que puedan realizar con mayor confiabilidad correcciones en bases de gran tamaño. Hacerlo manualmente toma un tiempo considerable. Algunas rutinas pueden hacerse de modo semiautomático utilizando paquetes estadísticos. Se han desarrollado herramientas como Centrus Merge/Purge, Data Tools Twins, Data Cleansing DataBlade, DataSet V, DATASTAGE, DECISIONBASE, DeDuce, DeDupe, dfPower, DoubleTake, ETI Data Cleanse, Holmes, i.d. Centric, Integrity, matchIT, matchmaker, NADIS Merge/Purge Plus, NoDupes, POWERMART, Pure Integrate, PureName PureAddress, Quick Address Batch, reUnion and MasterMerge, SAGENT SOLUTION PLATFORM, SSA-Name/Data Clustering Engine, Trillium Software System, TwinFinder, Ultra Address Management y WAREHOUSE ADMINISTRATOR, que dependiendo del

tipo de datos, facilitan la validación de información personal y la identificación de casos duplicados(Ahmed & Aziz, 2010 ; Maletic & Marcus, 2010 ; Maydanchik, 2007 ; Müller & Freytag, 2003 ; Rahm & Do, 2000).

También algunos procedimientos de limpieza pueden llevarse a cabo utilizando *software* que permite la programación de reglas y configuración de rutinas. Con un algoritmo simple pueden corregirse errores similares simultáneamente. Aunque esto ahorra tiempo, el reto está en el esfuerzo analítico requerido para entender los patrones y determinar las soluciones, así como en el hecho de que las reglas de calidad de datos son interdependientes. Corregir un elemento de un conjunto de valores, puede resultar en la violación de otras reglas establecidas. En un intento por corregir un error original pueden inducirse nuevos errores, si se falla en ver las distintas relaciones entre los valores. Entre los programas utilizados para corregir anomalías y realizar procesos de limpieza en las bases de datos están AJAX, FraQL, Potter'sWheel, ARKTOS e IntelliClean (Ahmed & Aziz, 2010 ; Maletic & Marcus, 2010 ; Maydanchik, 2007 ; Müller & Freytag, 2003).

IV. Métodos generales para limpieza de datos

Una limpieza completa incluye un agregado de operaciones que pueden resumirse como se muestra en la Tabla 1:

Tabla 1. Desarrollo de la limpieza de datos

AUDITORÍA DE DATOS	Análisis de atributos de los datos y de la colección de datos de la cual se genera información como mínimos y máximos, rangos, frecuencia de valores, varianza, unicidad, ocurrencia de valores nulos, patrones típicos específicos y generales. Los resultados contribuyen a la definición de criterios de consistencia, formatos de dominio e indicadores de posibles errores y las características con las que se presentan.
ESPECIFICACIÓN DEL FLUJO DE TRABAJO	Determinación de los procedimientos específicos a aplicar para modificar errores, así como para detectar y eliminar anormalidades en los datos. La especificación de métodos de limpieza y corrección debe considerar las posibles causas de los valores erróneos.
EJECUCIÓN DEL FLUJO DE TRABAJO	Verificación e implementación del flujo de trabajo; el objetivo es alcanzar la mayor exactitud posible. La aplicación eficiente de las distintas operaciones y reglas, debe ser factible independientemente del tamaño del conjunto de datos. La interacción con expertos resulta fundamental, puesto que el criterio profesional determinará ciertas decisiones de consistencia y validez de la información.
POSTPROCESAMIENTO /CONTROL	Comprobación de los resultados obtenidos para verificar la exactitud de las operaciones ejecutadas en la base de datos. Se llevan a cabo procedimientos de control para localizar información no corregida. El conjunto de datos resultante es objeto de un nuevo ciclo de limpieza.

Fuente: Ahmed & Aziz, 2010 ; Müller & Freytag, 2003

Müller y Freytag (2003) señalan que la limpieza de datos nunca termina por completo, algunas anomalías son difíciles de localizar y muchas veces se identifican hasta que las bases de datos se encuentran en procesos posteriores de análisis. Según las aplicaciones previstas de la información y según los recursos disponibles, se debe determinar la cantidad de tiempo y esfuerzo que se invertirá en este proceso, considerando los criterios mínimos de calidad esperados y los criterios mínimos para asegurar la compatibilidad con otros conjuntos de información en el sistema.

Tener conocimiento de cómo deben observarse los datos –información esperada– facilita la identificación de errores y valores fuera del patrón de normalidad. Sin embargo, es un hecho constatado que los datos de campo son a menudo muy diversos y rara vez se ajustan por completo a una distribución estadística estándar (Maletic & Marcus, 2010).

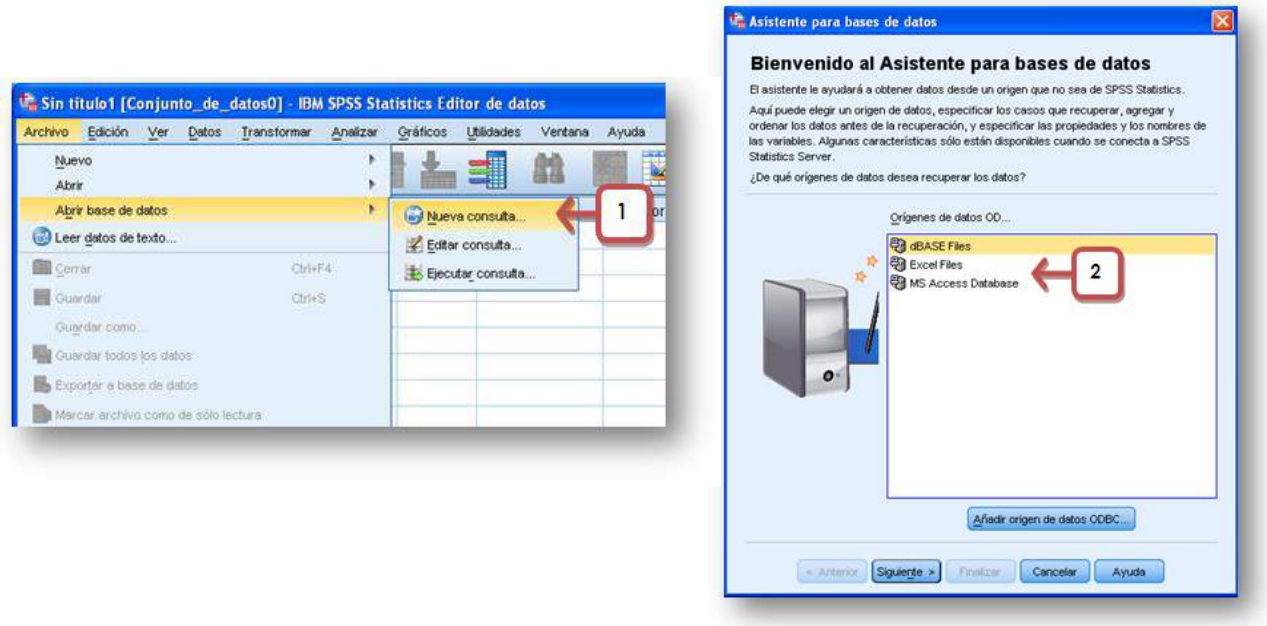
No existe un procedimiento único a implementar. Varias rutinas pueden adaptarse y utilizarse en bases con información similar, pero hay que considerar que cada conjunto de datos es particular. La metodología de limpieza de datos es altamente dependiente del dominio de conocimiento y de naturaleza significativamente exploratoria.

En la Digeduca las bases de datos se manejan en el *software* IBM® SPSS® Statistics. Dado que la digitación e ingreso de la información no se realiza dentro de la Subdirección de Análisis, el primer paso es la recepción y preparación del conjunto de datos.

Los dos formatos típicos en los que puede recibirse la información son: formato de libro de Excel (*.xls) o formato de base de datos de Microsoft Access (*.mdb). Una vez identificado el conjunto de datos con el que se va a trabajar, se importan y se transforman en un archivo de SPSS (*.sav), como se observa en la Figura 4[1]. Si el conjunto de datos está dividido en varios archivos, se sugiere agregar una variable "marcador" para poder identificar de dónde proviene cada caso cuando la información esté integrada.

[1]Las figuras y tablas que se presentan en adelante son elaboración propia de los autores, por lo que no se indicará individualmente su fuente.

Figura 4. Importación de archivos en SPSS



Es importante contar con el libro de códigos de las bases de datos con las que se va a trabajar. Este documento debe describir los ítems que se recolectaron, el proceso de ingreso de datos, la codificación de la información, así como la organización del archivo. También es primordial disponer de los instrumentos utilizados en la evaluación tal y como fueron aplicados. De manera que pueda observarse la estructura de la prueba o cuestionario, los tipos de preguntas y formatos posibles de respuesta.

Una vez que se tienen los datos en un archivo de extensión *.sav, se procede a generar frecuencias para cada una de las variables con el objetivo de conocer qué información nos ha sido trasladada en la base de datos. El contenido de esta información puede editarse desde las variables o desde los casos. Los archivos de IBM® SPSS® Statistics están organizados por filas y columnas. En la vista de datos (ver Figura 5), las filas representan los casos u observaciones, mientras que las columnas representan las variables. En la vista de variables (ver Figura 6), cada fila es una variable y cada columna es un atributo asociado a dicha variable.

Figura 5. Vista de datos en SPSS

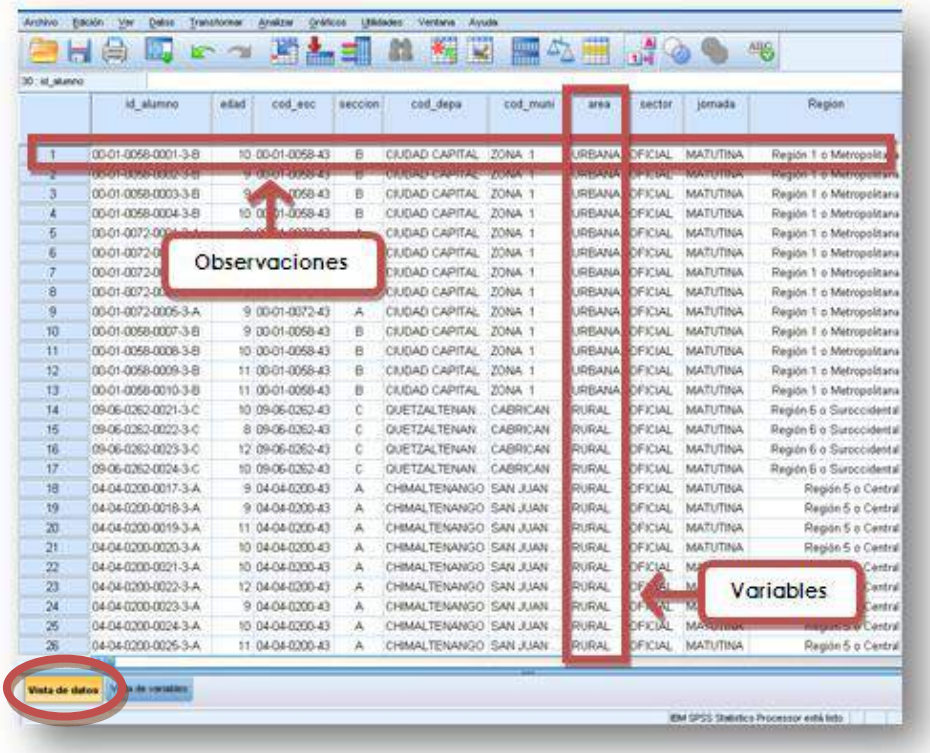
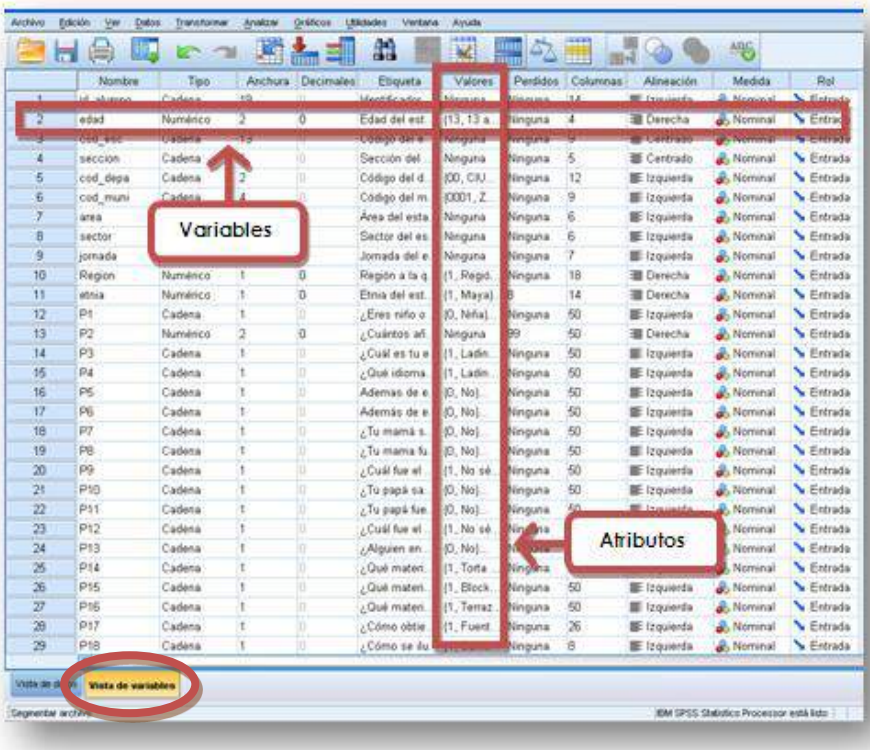
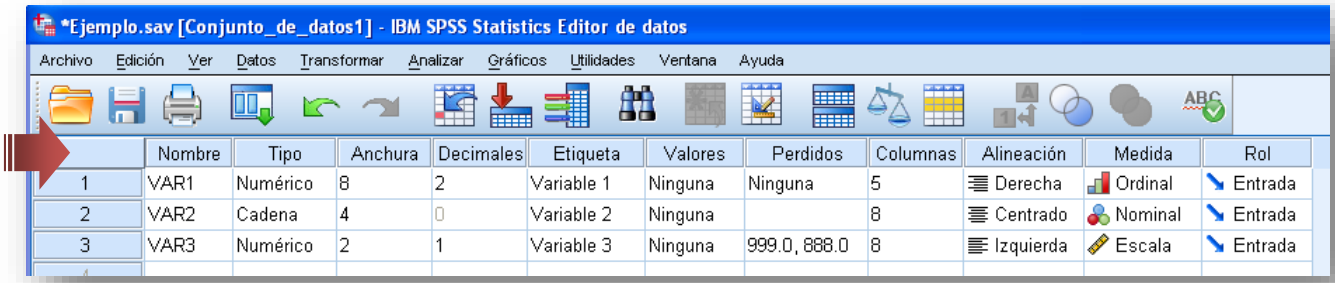


Figura 6. Vista de variables en SPSS



Distintas características pueden definirse para cada una de las variables, como se muestran en la Figura 7: nombre, tipo, anchura, decimales, etiqueta, valores, perdidos, columnas, alineación, medida y rol.

Figura 7. Atributos de las variables en SPSS



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	VAR1	Numérico	8	2	Variable 1	Ninguna	Ninguna	5	Derecha	Ordinal	Entrada
2	VAR2	Cadena	4	0	Variable 2	Ninguna		8	Centrado	Nominal	Entrada
3	VAR3	Numérico	2	1	Variable 3	Ninguna	999.0, 888.0	8	Izquierda	Escala	Entrada

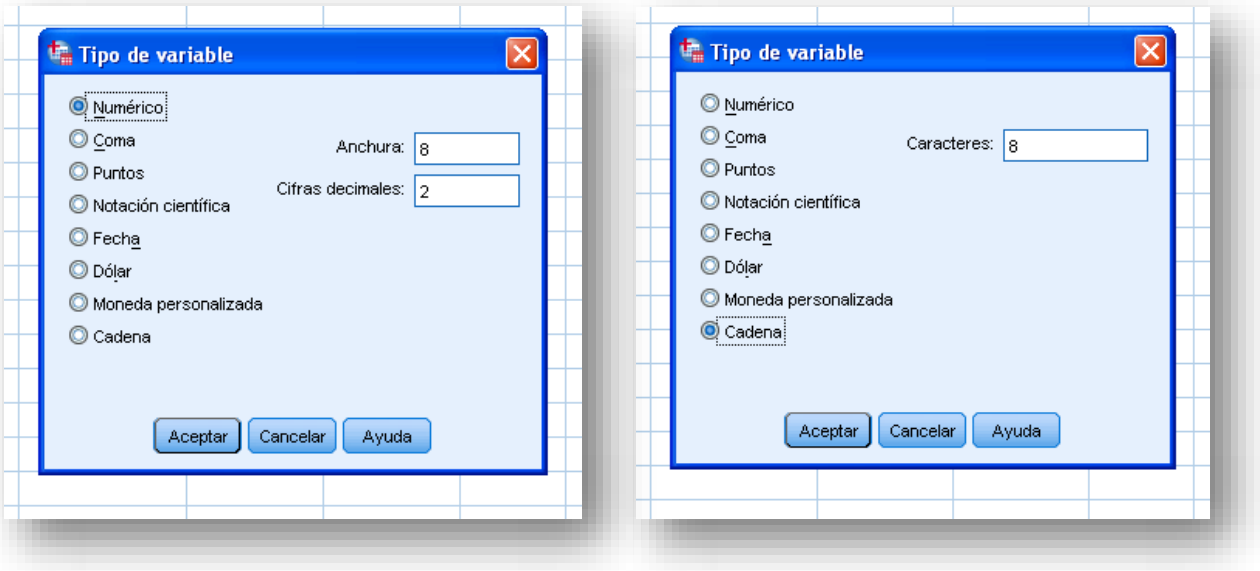
El **nombre** de la variable debe ser claro, conciso y representativo. Se espera también que si se determinan algunas normas de codificación, sea congruente con ellas. Pueden utilizarse caracteres alfanuméricos y guion bajo. No debe incluir tildes, espacios o caracteres especiales. Estas consideraciones son importantes para facilitar la migración y comparación de archivos entre distintas versiones de IBM® SPSS® Statistics. Los nombres de cada variable deben ser únicos dentro de cada conjunto de datos, es decir, en una base no pueden existir dos nombres iguales (ver Tabla 2).

Tabla 2. Ejemplos de variables

Nombre	Explicación de lo que representa la variable
id_alumno	(Identificador único del alumno)
cod_depa	(Código de departamento)
nom_depa	(Nombre de departamento)
forma_CL	(Forma del Cuestionario de Comunicación y Lenguaje)
forma_EE	(Forma del Cuestionario de Estrategias Docentes)
PM1	(Pregunta 1 de Matemáticas)
PL3_C	(Pregunta 3 de Lectura calificada)
P7_rep_primero	(7. Marque los grados que repitió: primero primaria)
P7_rep_tercero	(7. Marque los grados que repitió: tercero primaria)
P7_rep_sexto	(7. Marque los grados que repitió: sexto primaria)

El **tipo** de la variable puede ser numérico, coma, puntos, notación científica, fecha, moneda y cadena. Dada la información que se maneja en la Digeduca, los tipos de variable más utilizados son numéricos y de cadena (ver Figura 8).

Figura 8. Tipo de variable numérico y cadena en SPSS



Las variables numéricas son aquellas que tienen como valores números y por lo tanto permiten la realización de cálculos. Para ellas se definen las cifras decimales y el ancho, que incluye el punto decimal entre los caracteres. Las variables de cadena, también conocidas como variables alfanuméricas, tienen valores no numéricos, de manera que no pueden utilizarse en operaciones matemáticas o estadísticas. Los valores de las variables de cadena pueden contener cualquier caracter siempre que no se exceda la longitud definida. Las mayúsculas y las minúsculas se consideran diferentes.

La **anchura** se refiere al número de dígitos o caracteres de una variable. Este valor no debe ser menor al del dato con mayor longitud observado. Una vez que se reduce el ancho de la variable, los valores que quedan fuera del margen de largo establecido se pierden.

Algunas variables tienen un ancho definido previamente, como por ejemplo, código del establecimiento que tiene un largo de 13 caracteres, código del departamento que tiene una anchura de dos o código de municipio con una anchura de cuatro. Sin embargo, en estos casos se recomienda en un inicio, observar la frecuencia de valores para detectar datos incorrectos y para no crear un error cortando información de un valor.

Para formatos numéricos se pueden ingresar valores con **decimales**. IBM® SPSS® Statistics versión 19 permite hasta 16 dígitos decimales. En la "Vista de datos" se observa solo el número definido de dígitos decimales. Los valores con más decimales de los determinados se redondean para cumplir con el formato; el valor completo se almacena internamente y se utiliza en todos los cálculos.

Las variables también admiten la definición de una **etiqueta**, que identifica y describe el contenido de los valores de cada variable. Las etiquetas de variable pueden tener una longitud de hasta 256 caracteres, contener espacios y caracteres reservados que no se admiten en los nombres de variable.

Tabla 3. Ejemplos de etiquetas de variable

Nombre de la variable	Etiqueta de la variable
nom_escuela	Nombre del establecimiento educativo
cod_jornada	Código de la jornada del establecimiento
PM27	Pregunta 27 de Matemáticas
necesidad_educativa	¿El alumno tiene alguna necesidad educativa especial?
P12_experiencia	Tiempo de experiencia docente (años)
etnia_recode	Etnia del estudiante(re-codificada en Maya/No Maya)
dominio_idioma_materno	¿Qué destreza domina en su idioma materno?

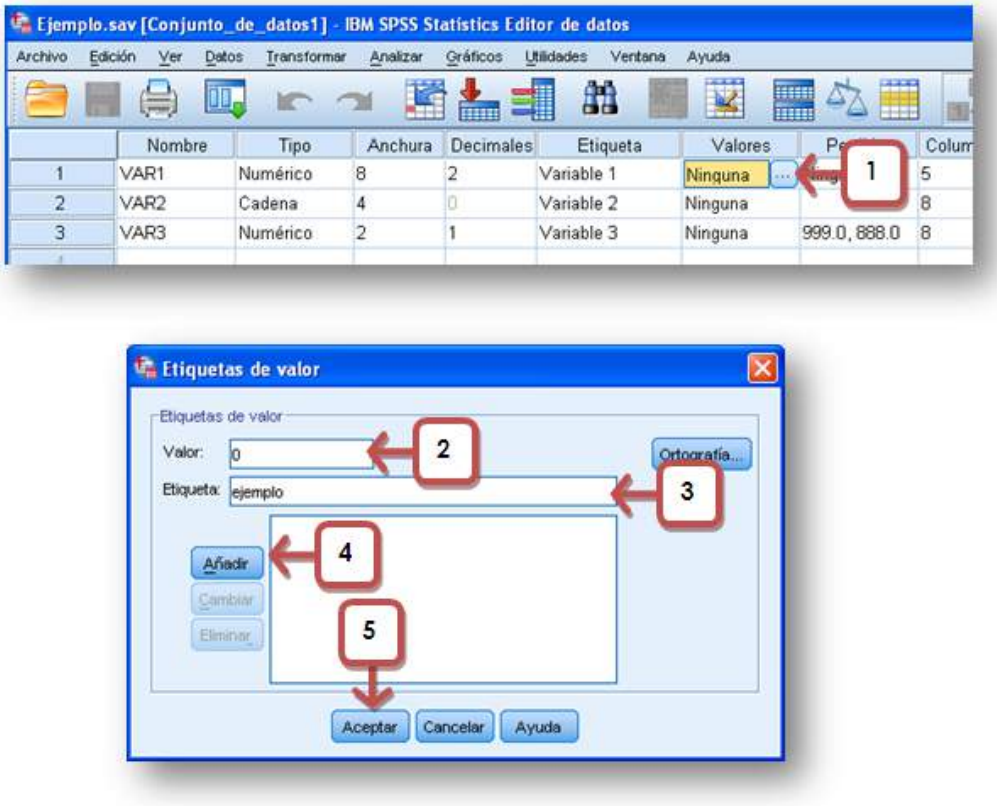
En la característica de **valores**, se pueden asignar etiquetas descriptivas a cada valor de una variable. Esto resulta útil, especialmente cuando una variable utiliza códigos para representar distintas categorías. Una vez definidas las etiquetas de valor para una variable, ya no es necesario volver a indicarlas cada vez que se abre el archivo de datos. Ejemplos de las etiquetas que se asignan a valores de las variables se pueden observar en la Tabla 4.

Tabla 4. Ejemplos de etiquetas de valor

Variable	Valores y etiquetas
sexo	0, Femenino; 1, Masculino
área	11, Urbana; 12, Rural
escalafón salarial	1, Clase A; 2, Clase B; 3, Clase C; 4, Clase D; 5,Clase E; 6, Clase F
idioma materno	1, español; 2, idioma maya; 3, garífuna; 4, xinka; 5, otro
asistencia a preprimaria	0, No; 2, Sí

En el programa SPSS se pueden definir las etiquetas de valor en la vista de variables de la base de datos, según se observa en la Figura 9, lo cual se puede hacer con la propiedad de la variable o con sintaxis de programación de comandos.

Figura 9. Definición de etiquetas de valor en SPSS

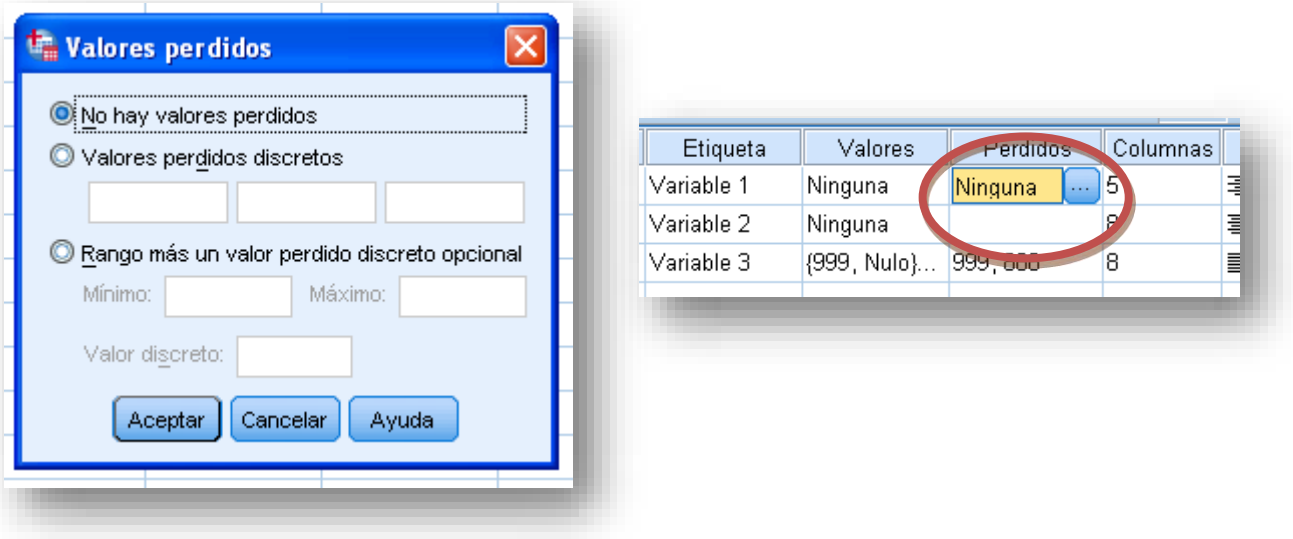


Como se ha mencionado antes, los datos perdidos o no válidos son comunes en las bases de datos. Es probable que las personas que hayan respondido a un cuestionario o participado en alguna evaluación, se nieguen a responder ciertos ítems, no conozcan la respuesta a determinadas preguntas o contesten de forma inesperada. Por ello es necesario identificar y filtrar estos datos **perdidos** en cada variable, para que los análisis proporcionen resultados exactos. IBM® SPSS® Statistics no utiliza los valores que se definen como perdidos en los cálculos realizados con la variable.

Se pueden señalar hasta tres valores perdidos de tipo discreto, un rango de valores perdidos o un rango más un valor de tipo discreto. Los rangos únicamente se pueden especificar para variables numéricas. Los valores perdidos de las variables de cadena distinguen mayúsculas y minúsculas.

El *software* considera como válidos todos los valores de cadena incluidos los valores vacíos o nulos, a menos que se definan explícitamente como perdidos. Para definir como perdidos los valores nulos o vacíos de una variable de cadena, se ingresa un espacio en blanco en uno de los campos de la sección valores perdidos discretos (ver Figura 10).

Figura 10. Ventana para definir valores perdidos para las variables en SPSS



Los valores perdidos también pueden tener etiquetas de valor, por ejemplo "Z" puede representar "Respuesta en blanco"; "Y", "Respuestas múltiples; "E", "Respuesta sin sentido"; 999, "Valor Nulo"; 888, "Doble Respuesta".

En la propiedad **columnas** se especifica el número de caracteres para el ancho de la columna en la "Vista de datos". Esta propiedad puede cambiarse también directamente desde la "Vista de datos", pulsando y arrastrando los bordes de las columnas. El ancho de columna afecta solo la presentación pero no los valores en el "Editor de datos" de IBM® SPSS® Statistics. Al cambiar el ancho de columna no se cambia el ancho definido de una variable.

Debe procurarse que la vista de los valores de las variables sea cómoda y fácil de manejar para el posterior estudio y análisis de la base de datos. La definición de esta propiedad dependerá tanto del largo de los valores de la variable como del aspecto estético y funcional deseado. La percepción de orden y claridad en la "Vista de variables" se favorece al homogenizar anchos de columna para variables similares.

La definición de formato de variables también considera la **alineación**. Los valores y etiquetas de datos pueden presentarse alineados a la derecha, a la izquierda o al centro. Por defecto, las variables numéricas tienen alineación hacia la derecha y las variables de cadena hacia a la izquierda. Al ajustar esta propiedad se afecta únicamente la "Vista de datos".

IBM® SPSS® Statistics permite especificar el nivel de **medida** de cada variable como Nominal, Ordinal o Escala; esta especificación se puede ver en la Tabla 5.

Tabla 5. Nivel de medida de la variable en SPSS

	ESCALA	ORDINAL	NOMINAL
TIPO DE DATOS	Datos numéricos de una escala de intervalo o de razón (continuos).	Datos de cadena (alfanuméricos) o numéricos.	Datos de cadena (alfanuméricos) o numéricos.
ÍCONO			
CARACTERÍSTICAS DE LA VARIABLE	Los valores representan categorías ordenadas con una métrica con significado, por lo que tienen sentido las comparaciones de distancia entre valores.	Los valores de la variable representan categorías con alguna ordenación intrínseca.	Los valores de la variable representan categorías que no obedecen a una ordenación intrínseca.
EJEMPLOS	Edad, cantidad de secciones abiertas, estudiantes inscritos, años de experiencia docente, duración en minutos de un período de clase, habilidad estimada.	Nivel de satisfacción, escalafón salarial, nivel educativo alcanzado, ítems de opinión/actitud con categorías de respuesta en escala tipo Likert.	Departamento, jornada, área, sector del establecimiento, región, nombre del estudiante, rama de enseñanza, ciclo, programas en funcionamiento.

Fuente: IBM® SPSS® Statistics, manual de usuario.

Para variables de cadena ordinales, el *software* asume que el orden alfabético de los valores indica el orden correcto de las categorías, por lo que para evitar errores y representar datos ordinales, es mejor utilizar códigos numéricos con etiquetas de valor.

Por último está la determinación del **rol** o papel de la variable (ver Tabla 6). Algunos cuadros de diálogo de IBM® SPSS® Statistics admiten roles para preseleccionar variables para el análisis, de manera que cuando las variables cumplen los requisitos se muestran automáticamente en las ventanas de selección. Esta propiedad no tiene ningún efecto en la sintaxis de comandos, solo en cuadros de diálogo que admiten asignaciones de papeles. Por defecto todas las variables tienen asignado rol de "entrada".

Tabla 6. Rol/papel de la variable en SPSS

ENTRADA	Predictor, variable independiente
OBJETIVO	Salida u objetivo, variable dependiente
AMBOS	Variable de entrada y salida
NINGUNO	Sin asignación de función
PARTICIÓN	Variable dividirá los datos en muestras diferentes

También se incluye el rol "segmentar" pero únicamente por compatibilidad con IBM® SPSS® Modeler, no tiene alguna función en IBM® SPSS® Statistics.

Una vez que se han definido los distintos atributos de una variable, es posible copiarlos y aplicarlos a otras variables. Se utilizan operaciones básicas de copiar y pegar para aplicar las características establecidas a una o más variables existentes, y para crear variables nuevas con atributos de una variable copiada.

En el caso de variables categóricas (nominales, ordinales) es posible asignar etiquetas y otras características, mediante la función "Definir propiedades de variables". Este proceso explora los datos reales, enumera todos los valores de datos únicos para cada variable seleccionada, identifica valores sin etiquetas, ofrece etiquetas automáticas, permite copiar etiquetas de valor definidas de otra variable en la variable seleccionada o de la variable seleccionada a variables adicionales.

El orden de las variables en la base de datos debe ser congruente con la estructura del instrumento; de igual manera que las características y etiquetas de información deben tener un orden consistente con el catálogo de códigos utilizado en la Digeduca. Es necesario realizar análisis de frecuencias para cada variable, observando valores representados, codificación, categorías, mínimos, máximos, extremos y estadísticos descriptivos posibles.

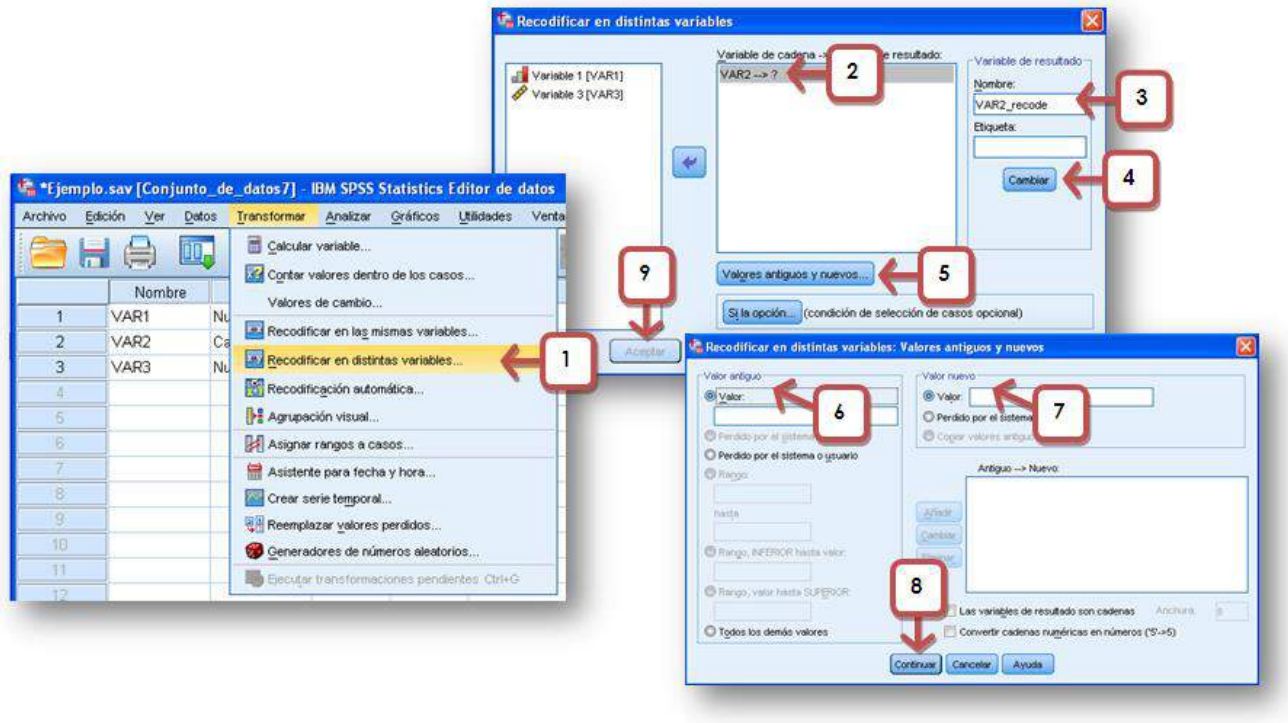
Mediante el contraste de la información en la base de datos, el libro de códigos de dicho archivo y los ítems en la prueba o cuestionario aplicado, se valida y ajusta la estructura de los datos.

Es preferible trabajar con códigos numéricos a trabajar con texto en las variables. Por ejemplo, en la variable sexo, es mejor definir "0", "1", con etiquetas "femenino", "masculino"; a tener una variable de cadena. Si se trabaja de la

segunda forma, es más probable incurrir en error; diferencias en la forma de escribir el contenido o faltas ortográficas pueden crear diferentes categorías. En IBM® SPSS® Statistics "FEMENINO", "Femenino", "femenino", "F" y cualquier otra variación de texto, se reconocerá como diferentes valores incluso cuando hacen referencia a la misma categoría. Por lo general, las variables de cadena se reservan para características de identificación como código de estudiante, código de establecimiento y nombres –del establecimiento, director, docente, estudiante, departamento, municipio–.

En algunos casos será necesario volver a codificar las variables existentes para asignarles las categorías numéricas deseadas. Para ello se utiliza la función "**Recodificar en distintas variables"**, en la que se señalan los valores existentes o valores antiguos y los valores deseados o valores nuevos (ver Figura 11). Mientras el proceso de limpieza se finaliza, se conserva la variable original y a la variable transformada se le agrega al final del nombre el sufijo "_recode".

Figura 11. Transformación de variables en SPSS



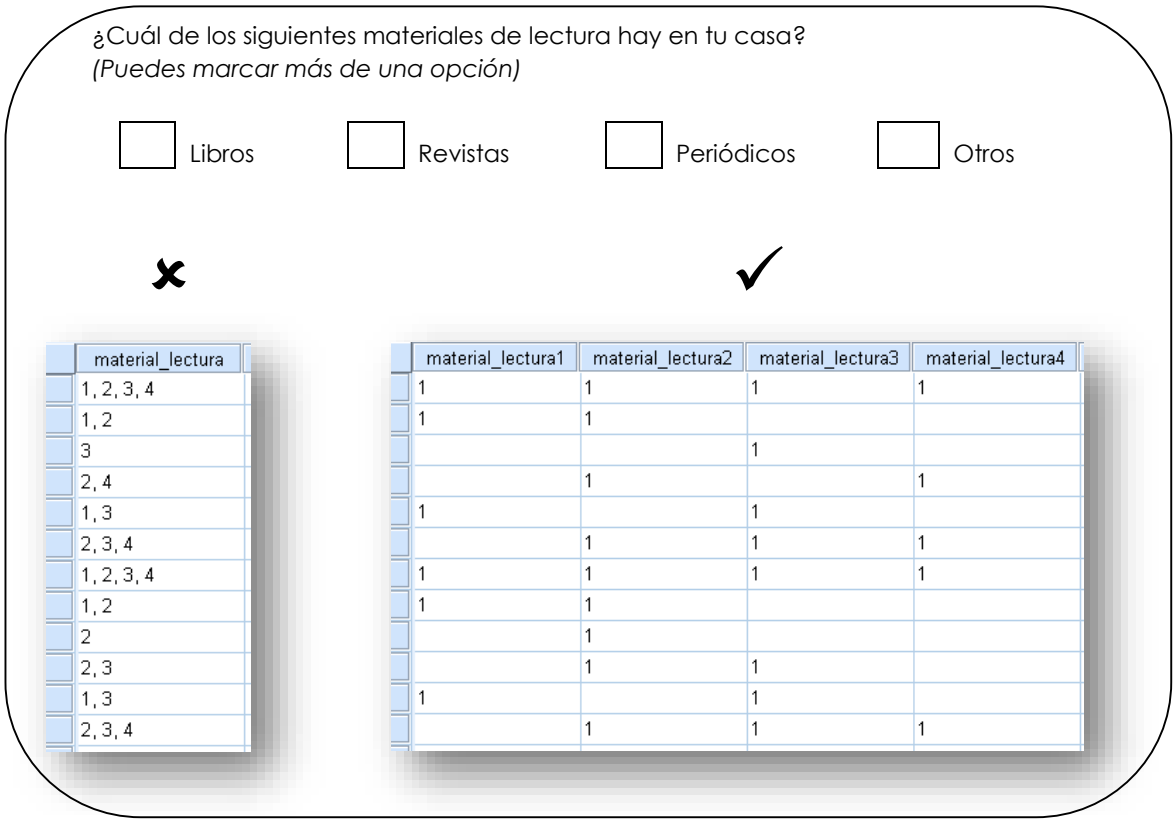
Con esta función también pueden convertirse valores específicos en valores perdidos o asignar a los datos perdidos un determinado valor. Además es posible definir un valor para un rango de datos.

Encontrar edades por ejemplo de 0, 55 o de 99 años en estudiantes de primero primaria de establecimientos de jornada regular, no resulta una información probable real, por lo que estos datos pueden ser sujetos a verificación en los

documentos originales o a falta de poder hacerlo, transformarse en datos perdidos. Para este tipo de variables también es posible agrupar datos sin perder información de interés. Por ejemplo, al observar las frecuencias de la variable "años de trabajar como director del establecimiento", se identifica que por encima de 25, hay muy pocos datos, por lo que puede establecerse una categoría de "25 años o más" que resuma la información en esta parte de la distribución de datos en una variable recodificada.

Por lo general, los instrumentos utilizados en la Digeduca tienen en su mayoría preguntas con respuesta de opción múltiple. En el caso de ítems que admiten más de una respuesta, se segmenta la variable en la cantidad de opciones que tenga definidas, una variable por opción, como se muestra en la Figura 12.

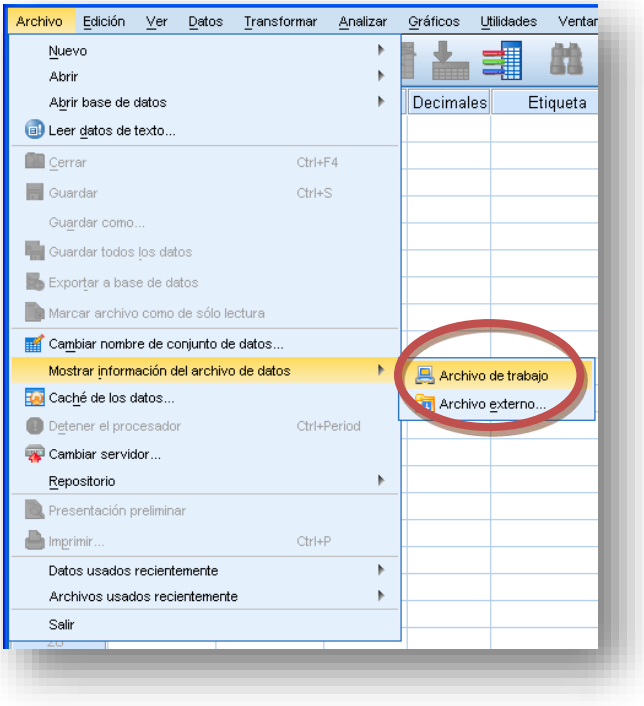
Figura 12. Ejemplo de pregunta que genera una variable por opción de respuesta



Este tipo de ítems tienen restringidas las categorías de respuesta por diseño y defecto. Sin embargo, en preguntas de respuesta abierta, como edad, años de experiencia docente, metros cuadrados de construcción del establecimiento, secciones abiertas, alumnos inscritos o docentes laborando en el establecimiento, es necesario establecer rangos de valores lógicos y válidos.

La versión final de la base de datos debe contener únicamente las variables limpias y se genera un libro de especificación de códigos, que luego puede copiarse a un documento de Word o Excel para su edición; una forma de obtenerlo se muestra en la Figura 13. Este archivo contiene información de las variables y los valores o etiquetas.

Figura 13. Generación del libro de especificación de códigos en SPSS



V. Problemas en la limpieza de datos

Existen diversos problemas que pueden ser resueltos mediante la limpieza y transformación de datos. Las inconsistencias descubiertas pueden deberse a discordancias con definiciones de datos, fallas en la entrada de información y corrupción en la transmisión o en el almacenaje, y afectar distintos tipos de variables. Algunos errores son prioritarios, pero la prioridad es definida específicamente por la evaluación, proyecto o estudio que se esté llevando a cabo (Van den Broeck et al., 2005).

Tabla 7. Ejemplos de problemas a nivel de esquema

PROBLEMA	DATOS SUCIOS	OBSERVACIONES
Atributo: Valores ilegales	Edad del docente: 8 Fecha de nacimiento: 25/13/2000	Valores fuera del dominio
Registro: Violación de características dependientes	Edad: 18 Fecha de nacimiento:01/08/1998 - Nombre: JUAN RONAL TZUL ZET Sexo: NIÑA	Edad=Año actual-Año de nacimiento, debería ser consistente. Nombre y sexo deberían ser congruentes.
Tipo de registro: Violación de unicidad	Nombre1: JOSE MARIO LOPEZ Código de estudiante1:0169450031C - Nombre2: ANGEL GABRIEL AYALA Código de estudiante2:0169450031C	No se cumple el código de estudiante como identificador único.
Fuente: Violación de integridad referencial	Nombre del Establecimiento: Escuela Oficial Rural Mixta Código de Municipio: 2020	La referencia del código de municipio no está definido (2020 no existe).

Según el origen de los datos, los problemas de calidad pueden ser problemas de una sola fuente o de múltiples fuentes y presentarse a nivel de esquema o a nivel de instancia. Los problemas a nivel de esquema se refieren a pobre diseño de la base y a falta de restricciones de integridad que controlen los datos válidos y permitidos (ver Tabla 7). Por otro lado, los problemas a nivel de instancia se refieren a errores e inconsistencias en el contenido actual (ver Tabla 8), los cuales no son visibles desde el esquema de la base de datos (Rahm & Do, 2000).

Tabla 8. Ejemplos de problemas a nivel de instancia

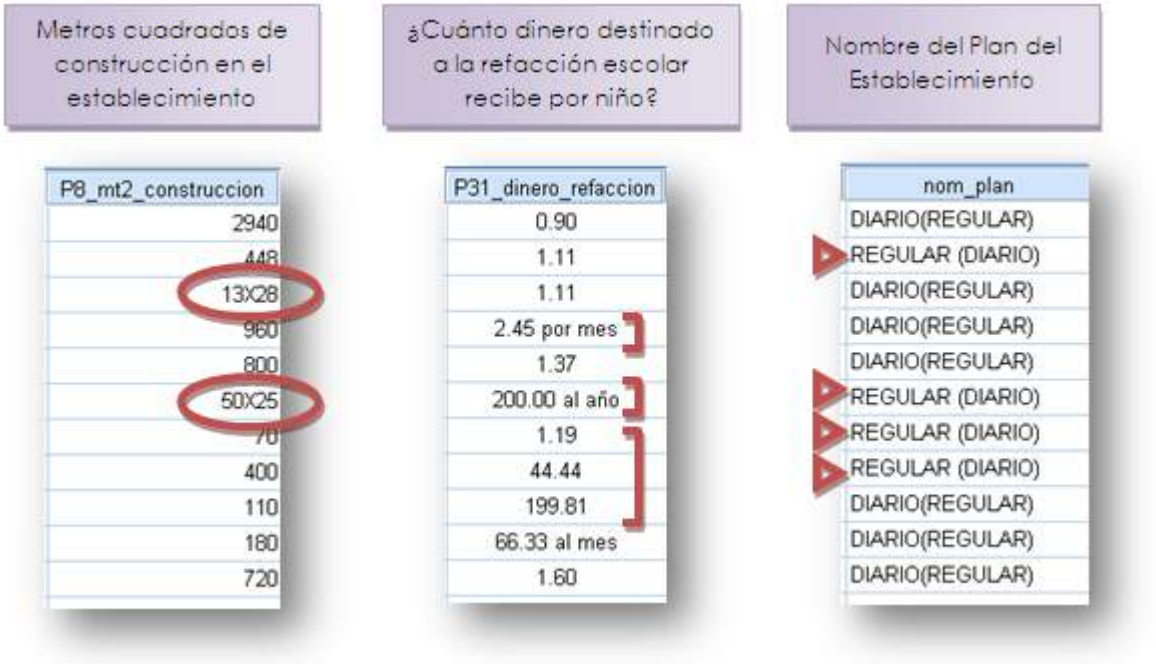
PROBLEMA	DATOS SUCIOS	OBSERVACIONES
Atributo: Valores perdidos	-	Valores no disponibles durante la entrada de datos.
Atributo: Errores de ortografía	Departamento: QUETZALTATENANGO Idioma: AXI	Usualmente errores de tipografía o errores fonéticos.
Atributo: Valores incrustados o mal ingresados	Nombre: CATHERINE YASMIN AJU IXAN 8 1A - Área: 31	Valores múltiples ingresados en un solo campo o ingresados en otra variable (confusión de columnas a la hora de la digitación).
Registro: Violación de características dependientes	Código de departamento: 20 Nombre de departamento: EL PROGRESO	Código y nombre de departamento deben corresponder.
Tipo de registro: Casos duplicados	Código de prueba1: E0509-6124730 Nombre del docente 1: SAQUIC TOL, TERESA - Código de prueba 2: E0509-3157796 Nombre del docente 2: SAQUIC TOL, TERESA	Mismo sujeto representado dos veces.
Tipo de registro: Contradicciones	Código de evaluación1: 11059654 Rama de enseñanza: BACHILLERATO - Código de evaluación2: 11059654 Rama de enseñanza: MAGISTERIO	El mismo sujeto descrito por dos valores diferentes.
Fuente: Referencias erróneas	Código del establecimiento: 00-18-1170-43 Código de departamento: 00 Código de municipio: 0015	La referencia del código de municipio (0015) existe, pero es errónea.

Problemáticas como las que se acaban de ejemplificar se agravan cuando se vuelve necesario integrar varias bases de información. Cada uno de los archivos puede contener datos erróneos y los datos pueden ser representados de diferente manera en cada una, coincidiendo o contradiciéndose. Las principales dificultades a nivel de esquema en estos casos son diferencias de etiquetas y de estructura de las variables(Rahm & Do, 2000).

Müller y Freytag (2003) señalan que también es posible clasificar las anomalías de las bases de datos en problemas sintácticos, semánticos y de cobertura. Las anomalías sintácticas describen características concernientes al formato y valores usados para representar a los sujetos. Las anomalías semánticas impiden que la colección de datos sea una representación exhaustiva y no redundante de la población. Y las anomalías de cobertura disminuyen la representación de la población en la muestra.

Discrepancias entre la estructura de los datos y el formato específico de las variables, se denominan **errores léxicos**. Los **errores de formato del dominio** se presentan cuando un determinado atributo del valor de una variable no va de acuerdo al formato de dominio anticipado. Las **irregularidades** se definen con el uso no uniforme de valores, unidades y abreviaciones. Esta utilización inconsistente de valores puede impedir la comparación entre casos y la interpretación adecuada de información, incluso cuando los datos representen datos correctos. Dado que estos inconvenientes están relacionados directamente con el formato general de los datos, se consideran problemáticas sintácticas(Müller & Freytag, 2003). Para ver ejemplos de estas anomalías véase la Figura 14.

Figura 14. Ejemplos de anomalías sintácticas en las bases de datos



Los problemas semánticos pueden observarse de distintas maneras. Las **violaciones a la restricción de integridad** describen casos o conjuntos de casos, que no satisfacen una o más condiciones de validez de los datos. Cada restricción de integridad es una regla que representa el conocimiento sobre el dominio y los valores permitidos para representar cierta variable. Por ejemplo: edad>0, años de experiencia docente<100, rama de enseñanza a nivel diversificado=bachillerato|magisterio|secretariado|perito|técnico.

Otra anomalía observada son las **contradicciones**, las cuales representan valores en una o más entradas, que violan algún tipo de dependencia entre datos. Las contradicciones incluyen tanto violaciones de dependencia funcional que pueden ser representadas como restricciones de integridad, como casos duplicados que contengan información inexacta.

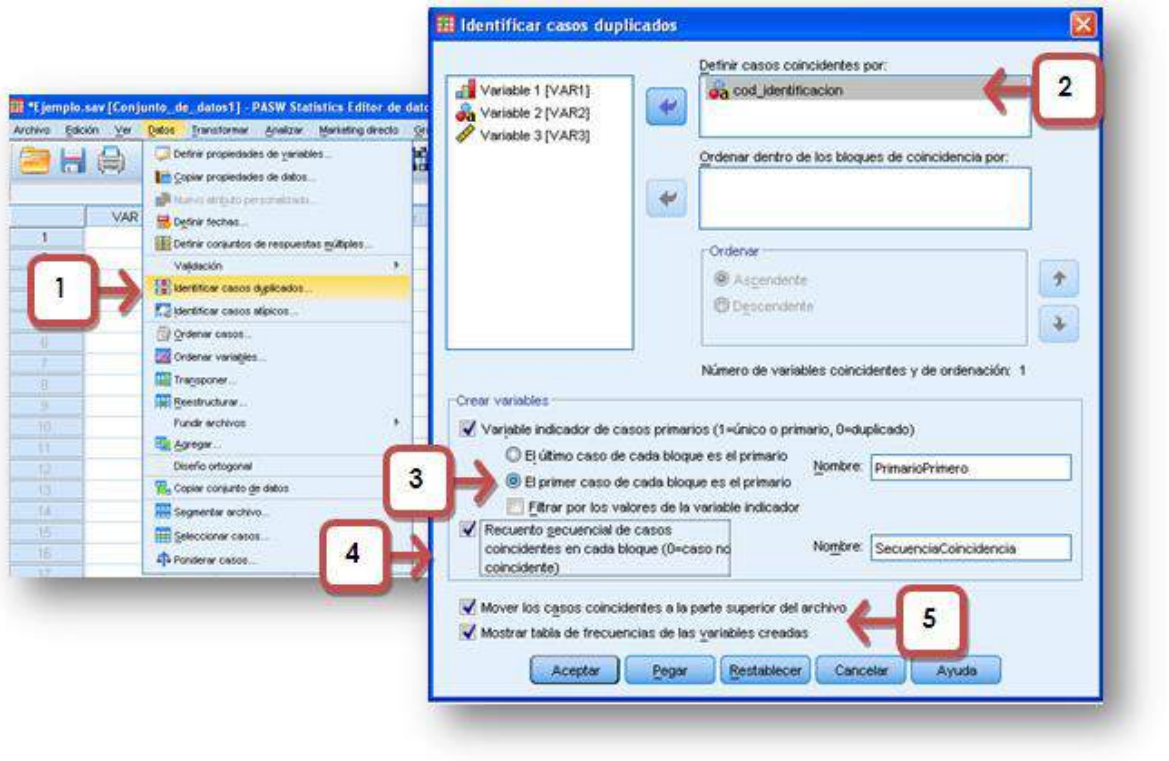
Los **duplicados** se definen como dos o más casos que representan al mismo sujeto. En ocasiones, los datos son idénticos completamente, lo cual facilita la depuración por eliminación o integración de información. Sin embargo, la limpieza de datos se dificulta cuando los casos repetidos representan al mismo sujeto, pero con distintos valores para algunas o todas las propiedades/variables. Por último, están también las **entradas inválidas**, que conforman la clase más complicada de anomalía en las colecciones de datos. Resultan de la incapacidad para describir la realidad mediante condiciones de integridad dentro de un modelo formal. No muestran problemáticas como las descritas con anterioridad, pero no representan información válida para la muestra. Las entradas inválidas son difíciles de detectar y cuando se identifican, la corrección de los datos puede resultar compleja (Müller & Freytag, 2003).

En IBM® SPSS® Statistics existe una función para identificar los casos duplicados (ver Figura 15). Se define una variable para seleccionar los casos coincidentes, la cual debe ser el identificador de los casos; por lo general, el código único del estudiante, código del docente, código del director o código del establecimiento. A partir de este proceso, se crean las variables "PrimeroPrimario", que identifica el primer caso de cada bloque de coincidencias como el caso principal y "SecuenciaCoincidencia" que contiene el recuento secuencial de casos coincidentes en cada bloque. Es importante tomar en cuenta que las variables de identificación también pueden tener algún error (como anomalías de digitación) por lo que este método no debe tomarse como forma única de tratar la problemática de duplicidad.

Según Müller y Freytag (2003), las problemáticas de cobertura pueden clasificarse en dos tipos. Los **valores perdidos**, que resultan de omisiones en la recolección de la información. Para algunas variables, es admisible un valor nulo como en el caso de no respuesta en preguntas de conocimiento, o no respuesta

en ítems que por determinada condición no aplican al sujeto. Pero para todas las demás variables, debe decidirse si el dato nulo en la muestra realmente existe en la población y si la información perdida debe y puede deducirse y recuperarse de alguna forma. También son consideradas como anomalías aquellos valores perdidos que no están en la base de datos, pero que de acuerdo a determinada característica y valor medible deberían presentarse para uno o varios sujetos. Por otro lado están los **casos perdidos**, que resultan de omisiones de sujetos de la muestra y de sujetos que no tienen ninguna entrada de información en la base de datos.

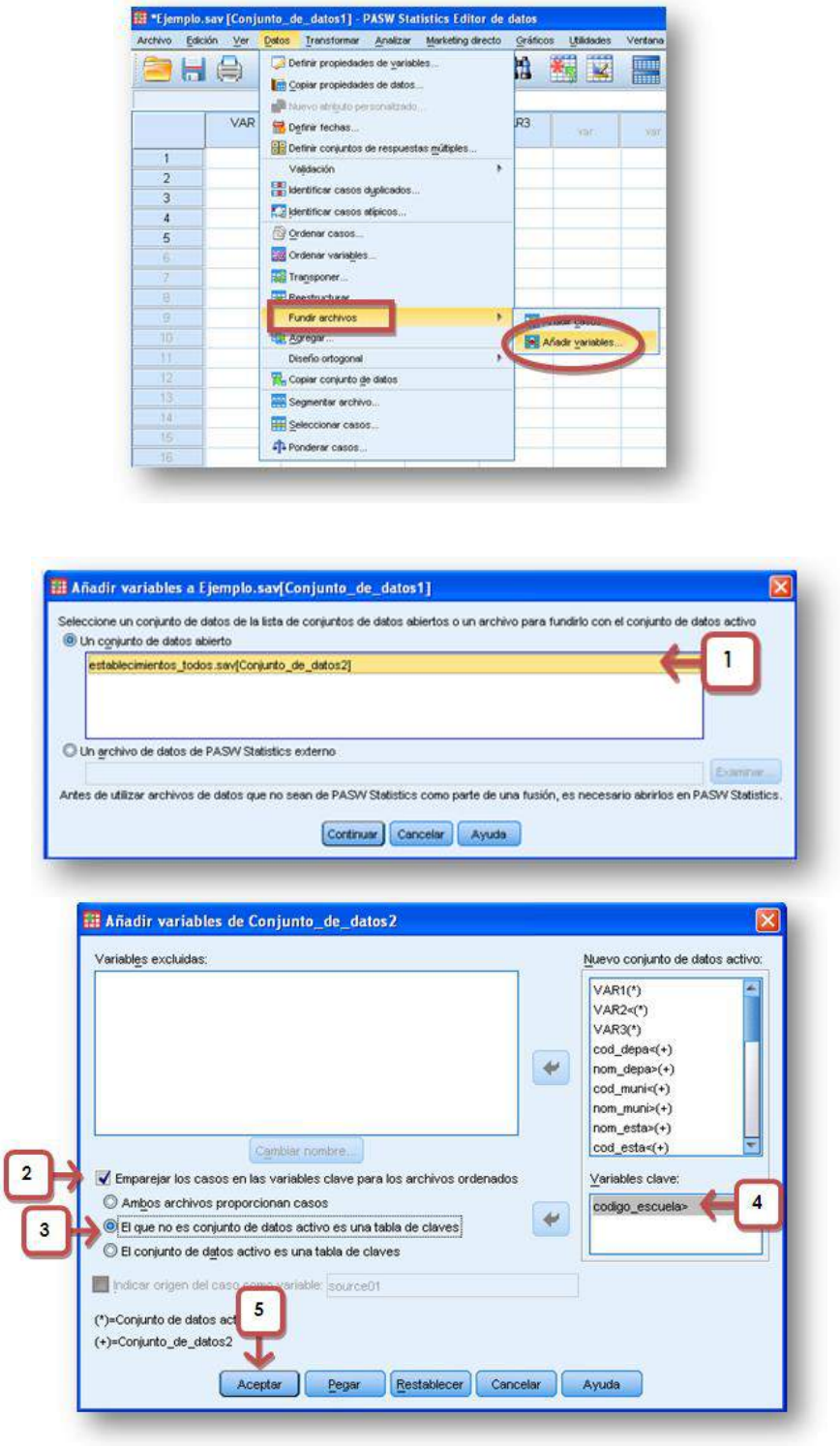
Figura 15. Identificar casos duplicados en SPSS



El tratamiento que se dé a los problemas de cobertura puede variar según las variables o propiedades medidas y según los objetivos de análisis. Algunas características de los sujetos pueden deducirse de información disponible, la cual puede provenir de la misma base de datos o de otras colecciones de información. Por ejemplo, datos de sexo pueden inferirse a través del nombre; el departamento o municipio pueden definirse si se tiene el código correcto del establecimiento. Características del centro educativo como nivel escolar, sector, área, jornada, plan y modalidad, se agregan si no estuvieran incluidas y esto puede llevarse a cabo utilizando otra base de datos como fuente de información complementaria. Los pasos básicos para realizar este proceso se ilustran en la Figura 16; la variable clave

o llave (vincula a los dos archivos) debe tener el mismo nombre y propiedades en ambas bases de datos y estar en orden ascendente.

Figura 16. Agregar variables de información a una base de datos en SPSS



Las diversas problemáticas en las bases de datos también pueden registrarse de acuerdo a la cantidad de datos o información que viola las restricciones de integridad y los criterios de calidad. Como errores simples, cantidad de datos incorrectos por caso, error dentro de un rango de ítems o inexactitudes en conjuntos de casos (Müller & Freytag, 2003). Independientemente, información duplicada debe ser depurada así como información complementaria debe consolidarse e integrarse para alcanzar una perspectiva consistente de los sujetos reales. Se pueden implementar diversas soluciones, aunque problemas relacionados deben tratarse de manera uniforme (Rahm & Do, 2000).

VI. Consideraciones finales

Con base en los elementos teóricos y metodológicos presentados; se esboza una lista de las actividades básicas que deben completarse para la entrega de una base de datos limpia y con criterios mínimos de calidad.

- ☐ Verificar variables de la base de datos con el instrumento y población aplicada.
  - ✓ cantidad de variables
  - ✓ cantidad de casos
  - ✓ incluir variables de opción múltiple donde se elijan más de una opción
  - ✓ casos no duplicados
  - ✓ todos los encuestados con identificación
- ☐ Verificar etiquetas entre libro de códigos del instrumento aplicado y la base de datos.
  - ✓ sintaxis aplicada
  - ✓ cantidad de valores por variable
  - ✓ etiqueta de variable descriptiva
- ☐ Verificar calidad de variables de identificación.
  - ✓ códigos con formato
  - ✓ códigos de registro
  - ✓ nombres y apellidos
  - ✓ identificador personal
- ☐ Verificar coincidencia entre códigos y nombres.
  - ✓ código de departamento con nombre de departamento
  - ✓ código de municipio con nombre de municipio
  - ✓ código de departamento con primeros dos dígitos del código del municipio
  - ✓ código de departamento con primeros dos dígitos del código del establecimiento
  - ✓ código de municipio con primeros cuatro dígitos del código del establecimiento

- ✓ cualquier variable que tenga código y nombre (como nivel escolar, sector, área, jornada, plan y ciclo escolar)

Y para cada una de las variables en la base de datos:

Análisis de frecuencias de la variable. ✓ cantidad y proporción de casos y valores de cada variable	☐
Revisión de propiedades de la variable. ✓ nombre, tipo, ancho, decimales, etiqueta, valores, perdidos, columnas, alineación y medida	☐
Revisión de posición en la base de datos, así como lógica y claridad del nombre de la variable. ✓ ordenada en la forma del instrumento o la disposición del análisis	☐
Verificación de códigos y valores. ✓ dentro de rangos específicos y sin caracteres extraños	☐
Revisión de ortografía de nombres y etiquetas. ✓ además de la revisión visual, la utilización de revisor de ortografía electrónico	☐
Verificación con otras variables relacionadas (si aplica). ✓ si solo debe ser llenada al dar cierto tipo de respuesta anterior debe coincidir con la condición	☐
Registro del proceso de limpieza de la variable. ✓ procesos de transformación, recodificación, corrección de datos, información agregada, datos/casos con sospecha de error y otras observaciones importantes	☐

VII. Referencias

Ahmed, I., & Aziz, A. (2010). Dynamic Approach for Data Scrubbing Process. *International Journal on Computer Science and Engineering, 2* (2), 416-423.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 37-54.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* Estados Unidos: Elsevier.

Information Builders. (2011). *Data Quality: It's no joke.* Obtenido de Insights - Trends & Technologies: http://informationbuilders.co.uk/new/insights/UKNewsletterQ1.html

Kitlas, J. (2012). *Dirty Data*. Obtenido de Subject Guides - Syracuse University Library: http://researchguides.library.syr.edu/content.php?pid=156265&sid=2578897

Maletic, J. I., & Marcus, A. (2000). *Data Cleansing: Beyond Integrity Analysis.* Obtenido de Proceedings of the Conference on Information Quality (IQ2000), Boston, October 2000: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.5212

Maletic, J., & Marcus, A. (2010). Data Cleansing: A prelude to knowledge discovery. En O. Maimon, & L. Rokach, *Data Mining and Knowledge Discovery Handbook* (págs. 21-36). New York: Springer.

Maydanchik, A. (2007). *Data Quality Assessment.* Estados Unidos: Technics Publications, LLC.

Müller, H., & Freytag, J.-C. (2003). *Problems, Methods, and Challenges in Comprehensive Data Cleansing.* Berlin: Humboldt University Berlin.

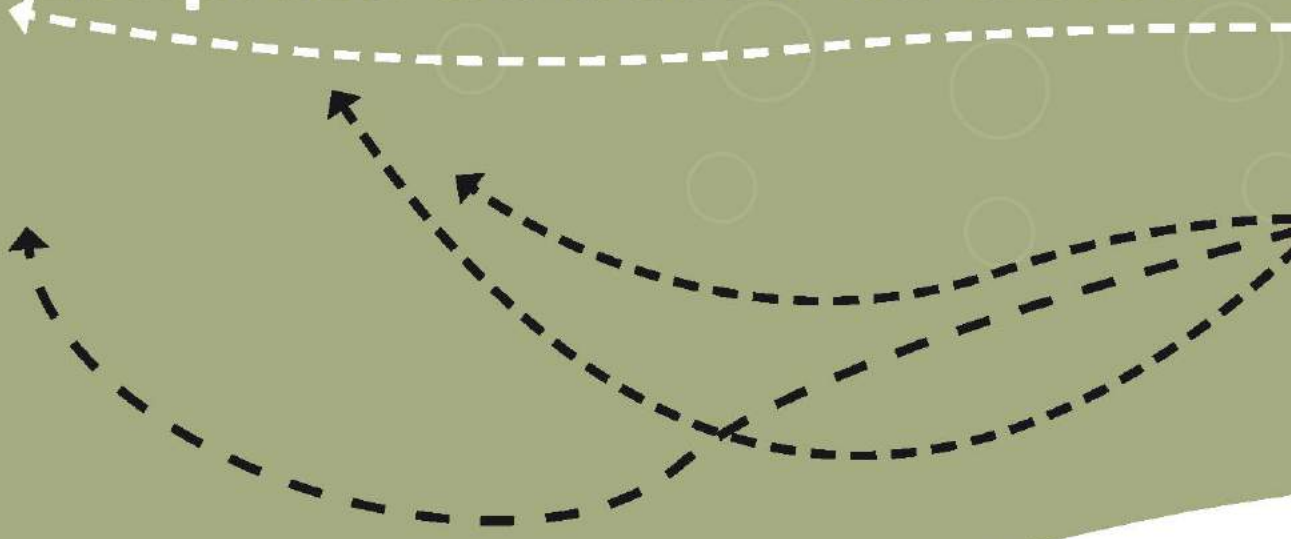
Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment. *Communications of the ACM* , 211-218.

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* , 1-11. Recuperado desde http://www.acm.org/sigs/sigmod/disc/disc01/out/websites/deb_december/rahm.pdf.

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting,diagnosing, and editing data abnormalities. *PLoS Medicine, 2* (10), e267.

Serie
de Cuadernillos Técnicos

Limpieza de bases de datos



Gobierno de Guatemala
Ministerio de Educación

Dirección General de Evaluación
e Investigación Educativa
-Digeduca-