Red7 Create

Módulo de generación de bases de datos

MATERIAL DE CLASES 1

INDICE

I. Introducción

Breve historia del procesamiento de datos

Un teléfono celular actual es mucho más rápido y potente que los computadores usados en la década de 1970. En aquella época, los computadores eran máquinas de gran volumen, pesadas y limitadas. Debido a su poca memoria interna, los procesos eran realizados a partir de instrucciones provenientes de tarjetas perforadas. Además, tenían que almacenar grandes volúmenes de datos en cintas magnéticas que, por su naturaleza, eran leídas en orden lineal.

Los censos de población y vivienda eran procesados en estas enormes máquinas. Los millones de casos y de variables –llamados microdatos- requerían de un programador para producir las tabulaciones que se analizarían posteriormente. Debido al costo y el tiempo que demandaba leer las cintas con los microdatos de un censo, era casi un imperativo crear, imprimir y publicar un conjunto de documentos con las tabulaciones oficiales a nivel de país y de las divisiones administrativas mayores.

Esto limitó los estudios sobre la base de datos censales a temas definidos y acotados, y redujo las posibilidades de explorar nuevas líneas de investigación. Dadas las dificultades para programar y producir tabulaciones especiales y el hecho que solamente las oficinas e institutos nacionales de estadística disponían de los microdatos de los censos, a los investigadores les resultaba muy difícil obtener tabulaciones adicionales específicas. Así, la retroalimentación entre datos y análisis era casi imposible.

Con el fin de solucionar este problema, Arthur Conning, entonces Jefe de Información y Procesamiento de Datos del CELADE[1], inició un nuevo proyecto dentro de la CEPAL. El objetivo era desarrollar un software capaz de identificar y localizar todas las tabulaciones y publicaciones censales, tanto generales como específicas.

Esta idea surgió después del éxito que tuvo DOCPAL[2] un sistema de documentación de datos demográficos en América Latina y el Caribe. Seguro de que estaba en el camino correcto, Conning visitó[3] las principales oficinas nacionales de estadística y de planificación de la región para buscar apoyo al proyecto.

Un descubrimiento fortuito

La conclusión después de este estudio de factibilidad realizado en varios países de América Latina y el Caribe fue concluyente: el proyecto propuesto no era capaz de

[1] CELADE – División de Población de la Comisión Económica para América Latina y el Caribe (CEPAL)

[2] DOCPAL tenía como objetivo crear una base de datos bibliográfica, conteniendo resúmenes y palabras clave, para auxiliar el análisis de encuestas de población hechas con SPSS. El nombre es un acrónimo para Documentación sobre Población en América Latina y El Caribe)

[3] Las misiones se concretaron con el apoyo económico del Centro Internacional de Investigación para el Desarrollo (CIID) de Canadá.

identificar o solucionar las necesidades reales de manejo de la información de los censos en la región. El interés de las oficinas nacionales de estadística apuntaba en otra dirección: tener la capacidad de generar tabulados censales específicos, que generalmente no eran considerados por los resultados oficiales publicados para cada censo, y disponer de información censal para áreas desagregadas que no correspondían a las divisiones administrativas mayores del país, e incluso para territorios que no necesariamente correspondieran a algún arreglo administrativo oficial.

La conclusión fue clara: el proyecto original no tenía futuro, y sin tener la intención, se había "encontrado en forma fortuita" (serendipity en inglés) un nuevo desafío mucho más interesante y de mayor envergadura para la región en esos momentos.

REDATAM: una solución rápida y amigable

A mediados de la década de 1980, CELADE, con el apoyo y financiamiento de muchos organismos internacionales, inicia el desarrollo de una herramienta más completa. En 1985 nace el proyecto REDATAM[4], responsable de crear una aplicación informática del mismo nombre dirigido a la generación y procesamiento de datos demográficos.

Basado en microcomputador (y no más en mainframes), esta solución permitió el acceso a combinaciones organizadas en forma jerárquica de grandes archivos de datos. REDATAM organiza y almacena toda la información de tal manera que el usuario pueda obtener de forma rápida y amigable cualquier tabulación u otro tipo de estadística básica hasta el área jerárquica más pequeña definida en los datos. En función de su principal característica, el lema del proyecto es "Fast & Friendly", como se muestra en la Figura 1.

Figura 1. Logotipo de la última versión de REDATAM

A lo largo de estos treinta años de vida del proyecto REDATAM, el software a pasado por diversas actualizaciones y mejoras. Actualmente la herramienta se encuentra en su quinta generación o séptima versión, razón por la cual se denomina *Redatam7*. En la Tabla 1, describimos brevemente las principales versiones de REDATAM, desde el inicio del proyecto hasta los días actuales.

Versión	Generación	Publicación	Sistema operativo, ambiente, plataforma	Ronda censal

[4] REDATAM es un acrónimo en inglés para Retrieval of Data for small Areas by Microcomputer, o recuperación de datos para áreas pequeñas por microcomputador.

Redatam DOS	Primera	1985	DOS	1980
Redatam 3.1	Primera	1986	DOS	
Redatam-Plus	Segunda	1991	DOS	1990
winR+	Tercera	1997	Windows	
Redatam+G4	Cuarta	2001	Windows, WebServer	2000
Redatam+SP	Cuarta	2004	Windows, WebServer	
Redatam7	Quinta	2014	Windows, WebServer, 64 bits, Unicode-multilingüe	2010

Tabla 1. Evolución de REDATAM (1985-2014).

Redatam7: rapidez, eficiencia y facilidad de uso

El desarrollo de *Redatam7* es guiado por tres objetivos principales:

i. Aumentar la velocidad de procesamiento de datos;
ii. Facilitar la programación de indicadores; y
iii. Mejorar la experiencia técnica del usuario.

Buscando facilitar el desarrollo de aplicaciones y funciones especificas requeridas por los usuarios, el desarrollo de esta última versión de REDATAM se basó en lenguajes de programación C++, Delphi y JavaScript. Actualmente *Redatam7* está disponible para Windows y estamos probando versiones en Linux, Macintosh y sistemas operativos móviles.

Al contrario de las versiones anteriores, *Redatam7* está dividido en tres módulos: *Create*, *Process* y *Admin*. El primer módulo genera una base de datos y un diccionario en formato REDATAM a partir de datos de entrada originales. El segundo procesa estos datos, generando tabulaciones definidas por el usuario. Y el tercero permite administrar las bases de datos en formato REDATAM, una tarea esporádica que, para mejorar su utilización, no está incluida en los otros dos módulos.

Redatam7 incorpora nuevos recursos en relación con la versión *Redatam+SP*, de 2004. Muchos de los cambios surgen a partir de la contribución de nuestra red internacional de usuarios, tanto de América Latina y de El Caribe, como de otras partes del mundo. Además de hacer más amigable la programación en REDATAM, la presentación del ambiente Web fue mejorada sustancialmente. A continuación se presentan las principales actualizaciones y sus beneficios.

Recurso	Beneficios
✓ **Lenguaje XML como estándar**	Ayuda a sincronizar las tareas de documentación, capacitación y programación, así como la interconexión con otros sistemas. Con esto se intenta disminuir la brecha entre el desarrollo y su conocimiento por parte de los usuarios.
✓ **Nuevo compilador**	Diseñado y programado para definir una gramática o sintaxis que adecua el lenguaje a nuevos requisitos, mejorando así la detección y visualización de errores.
✓ **Soporte al formato Unicode**	Fundamental para generar aplicaciones en diversos idiomas adicionales a los ya contemplados originalmente (inglés, español, portugués, francés e bahasa[5]).
✓ **Tabulaciones personalizadas**	Ofrece acceso a cada uno de los elementos involucrados en la presentación de tabulados, permitiendo al usuario definir cuáles serán mostrados y como serán posicionados. Las salidas se presentan en XML para facilitar la exportación a otros sistemas.
✓ **Tabulaciones con dimensiones ilimitadas**	No existen más límites al número de variables a ser incluidas en una tabulación. Hasta la versión anterior, solamente era posible cinco variables, más un quiebre de área.

[5] Lengua indonésia.

II. Conceptos básicos y definiciones

Para trabajar con una base de datos REDATAM, es necesario primero entender mejor su estructura. A continuación explicamos algunos conceptos y definiciones básicas que nos ayudaran en el camino.

Entidades

Entidades son conjuntos de objetos lógicos organizados en forma jerárquica en la base de datos. Una entidad, por ejemplo, puede ser el conjunto de provincias o ciudades en un determinado país.

En la Figura 2, se presenta una base de datos simple, que contiene cuatro entidades organizadas jerárquicamente: PAIS, PROVINCIA, DISTRITO y VIVIENDA.

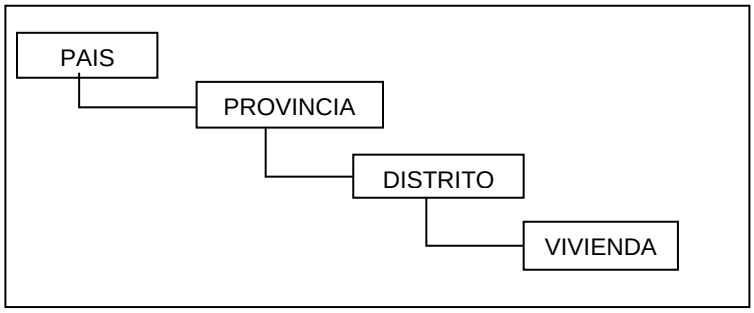

Figura 2. Estructura jerárquica simple con cuatro entidades

Entidades superiores

Las entidades pueden ser clasificadas de acuerdo con su posición en la jerarquía. Las **entidades superiores** son aquellas que están sobre una entidad determinada. Tomando nuevamente la Figura 2 como referencia, podemos decir que la entidad DISTRITO es superior a la entidad VIVIENDA. La entidad inmediatamente superior puede ser llamada también de **entidad madre**. Por definición, cada entidad tiene sólo una entidad madre, y en consecuencia, sólo hay un único camino de vuelta hasta la entidad raíz.

Entidades inferiores

Las **entidades inferiores**, a su vez, son aquellas entidades localizadas jerárquicamente debajo de una entidad determinada. En la Figura 2, se observa que la entidad DISTRITO es inferior a la entidad PROVINCIA. Las entidades inmediatamente inferiores pueden ser llamadas **entidades hijas**. Por definición, una entidad puede tener un número variable (igual o mayor que cero) de entidades hijas.

Elementos

Los **elementos** son los miembros individuales de cada entidad. Siguiendo con el ejemplo de la Figura 2, para la entidad PROVINCIA, los elementos son cada una de las provincias que componen el país.

El número de elementos varía significativamente de acuerdo a la entidad. Puede ser que existan unos pocos elementos, tal como las provincias, o varios millones como los individuos en la entidad PERSONA.

En el entorno estadístico, los elementos pueden asumir otros nombres, entre los más comunes pueden ser: observaciones, casos, instancias o registros. Para efectos didácticos, cuando nos referimos a una estructura de bases de datos, como la de REDATAM, optaremos por la expresión elemento. Para el caso de señalar el contenido de un archivo físico, usaremos la expresión registro.

Los elementos, así como las entidades, siempre respetan una jerarquía de estructura de bases de datos. Esto implica que cada elemento de una entidad madre está subdividido en los respectivos elementos de sus entidades hijas. Volviendo a la Figura 2, podemos deducir que cada elemento (provincia) de la entidad PROVINCIA está subdividido en elementos (distritos) de la entidad DISTRITO.

Elementos identificables o seleccionables

Cuando los elementos de una entidad poseen nombres y/o códigos propios que los distingan individualmente, decimos que son **identificables**. Al generar una base de datos en formato REDATAM, podemos elegir las entidades que contienen sus elementos identificados, razón por la cual son clasificadas como entidades **seleccionables**,

Aunque, por definición, cualquier entidad puede ser seleccionable, en la práctica, se acostumbra asignar esta propiedad a las entidades geográficas debido a su relevancia estadística. Por tratarse de entidades mayores, desde el punto de vista de los formuladores de políticas públicas, existe interés en identificar sus elementos. En una base de datos cualquiera, por ejemplo, queremos saber cuál es el distrito más poblado de una provincia. Al marcar la entidad DISTRITO como seleccionable, REDATAM nos proporciona esta información relevante. Al contrario, difícilmente existe interés estadístico en identificar los elementos de las entidades menores, no geográficas. Siguiendo con el ejemplo anterior, cuál es la ventaja de conocer la vivienda más poblada de un distrito? Para los formuladores de políticas públicas, ninguna.[6]

En el entorno de REDATAM, los elementos de las entidades seleccionables también pueden ser mapeados gráficamente usando su **código geográfico**.

[6] A esto se suma el hecho de que entidades menores -como personas-, en muchos países están protegidas por el principio del secreto estadístico que asegura la inviolabilidad de la intimidad de las personas físicas y jurídicas (empresas), las cuales no pueden ser objeto de divulgación de datos que las identifiquen estadísticamente.

Código geográfico

El código geográfico permite identificar en una tabla o mapa la unidad geográfica que dicho código esta identificando y el cual puede llegar hasta los niveles más desagregados de una base de datos ya sean las manzanas o sectores de enumeración. Los elementos personas, hogares y viviendas no llevan código geográfico por lo tanto no pueden ser identificados en una tabla o mapa).

El código geográfico está compuesto por una secuencia alfanumérica específica para cada uno de los elementos de una entidad. Cada parte del código informa la pertenencia a una determinada entidad, razón por la cual también se conoce como código compuesto. Un elemento del distrito 07, de la provincia 02, del país 0 (por definición), por ejemplo, puede recibir el código geográfico 00207.

Elementos no identificables y/o no seleccionables

Cuando los elementos de una entidad no tienen nombres y/o códigos propios que los distingan individualmente, señalamos que son **no identificables**. Al generar una base de datos REDATAM, podemos elegir las entidades que no tendrán sus elementos identificables, razón por la cual son clasificadas como entidades **no seleccionables**. Este normalmente es el caso de entidades no geográficas, como vivienda y personas.

Aunque los elementos de entidades no geográficas no deben ser identificados individualmente, el procedimiento puede hacerse en conjunto. Esto se hace mediante la localización del grupo en elementos identificables de entidades superiores que son identificables. En el ejemplo anterior, un elemento específico probablemente una vivienda, que por definición, no es identificable, sin embargo, gracias a su código geográfico – que comparte con otras viviendas del distrito- podemos identificarlo en su conjunto.

La característica de seleccionable para una entidad es definida por el usuario, al crear la entidad en el diccionario. También se usa el término de entidad Identificada y No Identificada, con el mismo propósito de Seleccionable y No Seleccionable, por la existencia de un código de identificación para sus elementos.

Ramas y niveles

Una **rama** es el camino recorrido desde la entidad raíz a una determinada entidad. Por definición, una entidad (y por ende todos sus elementos) pertenece a una única rama.

Si se compara la estructura de una base multidisciplinaria (Figura 3) con un árbol, las "ramas" de la estructura son exactamente lo mismo que las ramas del árbol, y las entidades serían las hojas del árbol.

Una base de datos REDATAM puede tener más de una rama jerárquica y éstas a su vez pueden tener varias ramas, los conceptos y elementos de entidad son los mismos. Una entidad seleccionable no necesita ser necesariamente geográfica.

Por ejemplo, en la Figura 3 se presenta la estructura jerárquica de la base de datos de demostración correspondiente a Nueva Miranda[7] que se distribuye con *Redatam7*. En esta estructura la entidad *EDUCAC* es el establecimiento educacional, *TENSENAN* es el tipo de enseñanza, y *CURSOS* son los cursos existentes, vale decir que se puede seleccionar sólo los datos de educación en donde se agregan los datos de tipo de enseñanza y cursos.

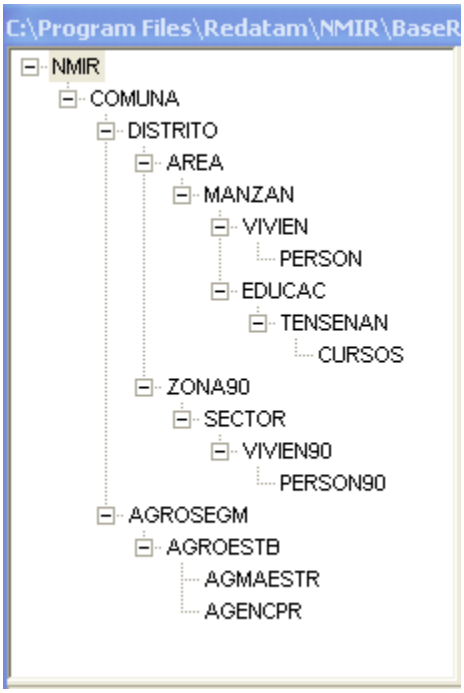

Figura 3. Estructura jerárquica multidisciplinaria

El **nivel** es el grado de profundidad de una entidad dentro de su rama. Por construcción, la entidad "raíz" tiene nivel cero, la(s) entidad(es) inmediatamente inferiores tienen nivel 1, las entidades bajo esta tienen el nivel 2 y así consecutivamente.

Variables

Las variables son informaciones referentes a los elementos que integran una entidad y por ende una base de datos. Una base de datos censal guarda información en **variables**, tales como el sexo, parentesco y edad de las personas, el tipo de techo y pared de las viviendas, cantidad de milímetros de agua caída en las comunas, etc. Cualquier entidad de una base de datos puede contener variables, por ejemplo, la PROVINCIA puede tener variables como su nombre, código geográfico, total de personas, total de hogares, índice de vulnerabilidad, etc. El DISTRITO puede almacenar el número de viviendas sin agua potable o el total de jefes de hogar femenino.

[7] Nueva Miranda es la base de datos de demostración que acompaña al programa REDATAM y corresponde a un país ficticio. Se guarda bajo el mismo directorio de instalación de Redatam7

La tabla 2, a continuación, muestra la relación entre los elementos y las variables de una entidad genérica.

	ENTIDAD		
	Variable 1	**Variable 2**	**Variable 3**
Elemento 1 →			
Elemento 2 →			
Elemento 3 →			
Elemento 4 →			

Tabla 2. Relación entre elementos y variables de una entidad.

III. Estructura de los archivos de una base de datos REDATAM

Una base de datos REDATAM contiene necesariamente cuatro tipos de de archivos, cada uno con características específicas: archivo del proyecto, archivo del diccionario de datos, archivos de datos, y archivos de punteros.

Los archivos de datos guardan los datos ya en formato REDATAM los cuales fueron cargados a través del modulo Red7 Create junto con los archivos de punteros y se pueden procesar utilizando el modulo Red7 Process a través del diccionario

Proyecto (redPrjX)

Se incorpora el proyecto como elemento administrador de la base de datos, diccionario, programas, selecciones geográficas, estilos de tablas, mapas y otros elementos.

Diccionario de la base de datos REDATAM (.dicX)

Este es un solo archivo que se genera para almacenar toda la información de la base de datos REDATAM. Almacena los metadatos, es decir la información de los datos. En el módulo Red7 Process se accede a la base de datos solamente a través del diccionario el cual incluye la estructura jerárquica, las entidades tanto seleccionables y no seleccionables, las claves de identificación, y las variables.

Archivos de datos REDATAM (.rbf)

REDATAM posee funciones para leer y convertir datos guardados y escritos originalmente en formato CSPro[8], SPSS[9], xBase[10], CSV[11] y ASCII[12] al formato REDATAM. Este proceso lee los "meta datos" desde su diccionario original y en base a la información que se define en un esquema REDATAM, se estructuran los datos a la forma interna de REDATAM bajo una estructura binaria y comprimida con la extensión RBF (del inglés *Redatam Binary File*).

REDATAM guarda la información de cada variable en su propio archivo de datos, con un "registro" (línea) por cada elemento de la entidad. De ahí que cada variable es un vector de los datos. Existe un archivo para sexo, uno para estado civil, otro para tipo de vivienda, etc.

En la estructura de archivos de REDATAM el número total de registros (líneas) corresponde necesariamente al número de elementos (observaciones) que hay en la entidad a la que pertenece esa variable. En otras palabras, todos los elementos de una entidad deben tener un valor para esa variable en cuestión, ya sea un valor en blanco o ignorado. Para ilustrar esto se tiene la siguiente tabla con entidades y registros en cada una:

ENTIDAD	ELEMENTOS
PROVINCIA	5
DISTRITO	23
VIVIENDA	4.587

Tabla 3. Ejemplo de número de elementos por entidad.

Analisando a Tabla 3, vemos que cada uno de los archivos de datos (archivo de variables) para la entidada PROVINCIA tiene 5 registros (líneas), para la entidad DISTRITO tienen 23 registros (líneas) y para la entidad VIVIENDA tienen 4.587 registros (lineas).

Los nombres qe se asignan a cada uno de los archivos con variables esta compuesto por el nombre de la base (definido al momento de creación de la base, por ejemplo PAIS) y

[8] CSPro es un programa estadístico, de dominio público, desarrollado por la Oficina de Censos de Estados Unidos (Census Bureau), la consultora Internacional ICF y SERPRO SA.

[9] SPSS es un programa estadístico, de naturaleza comercial, desarrollado por la empresa homónima y comprador por la IBM en 2009.

[10] xBase es el nombre genérico que se la da a las estructuras y lenguajes desarrollados a partir de dBASE, lenguaje comercial muy popular en la década de los 80.

[11] CVS es el nombre que se le da a los archives de datos que utilizan texto simple (comma separated values).

[12] ASCII es un tipo de código binario usado mundialmente. Codifica un conjunto de 128 símbolos: 95 gráficos (letras del alfabeto, puntuación y símbolos matemáticos) y 33 símbolos de control.

una secuencia numérica ascendente. De esta forma, podemos tener los arquivos pais0032.rbf para la variable sexo y pais0033.rbf para a variable edad.

Archivos transpuestos: menos espacio y mayor velocidad

En la mayoría de los programas estadísticos las variables para cada registro son guardadas en un único archivo de datos rectangular. En estos casos cuando se desea obtener los valores de una determinada variable, el sistema debe recorrer todo el archivo desde el primer registro, elemento por elemento, hasta encontrar el buscado, lo que toma mucho tiempo y exige mucha CPU.

REDATAM por otro lado, almacena cada variable por separado en su propio archivo de datos, permitiendo al sistema leer solo el archivo que contiene la variable deseada, ignorando los demás datos haciendo el proceso mucho más veloz.

Estos archivos separados para cada variable pero lineales para cada registro se conocen como archivos transpuestos, para distinguirlos de la forma más común de organización de los datos estadísticos rectangular. Entre otras ventajas, ello hace que el procesamiento del sistema sea eficiente, ya que permite que el sistema lea solamente las variables específicamente involucradas en un determinado proceso, a su vez ocupa menos espacio en el disco ya que cada valor en un archivo de una variable tiene el mismo tamaño podemos comprimir los archivos usando una técnica simple. Esta compresión también garantiza mayor velocidad de lectura de los datos.

El proceso de compresión adoptado en REDATAM transforma los valores de las variables en números binarios. El número de bits necesarios para cada variable depende del valor máximo que pueda presentarse en una variable y es definido por el usuario al momento de crear una base de datos.

La variable sexo, por ejemplo, contiene los valores 1 y 2 (siendo 1 para hombres y 2 para mujeres). Esta variable es comprimida en 2 bits, espacio suficiente para almacenar un número menor o igual a 3. Otro ejemplo, seria la variable edad con valores de 0 a 125 (los años de una persona). Esta variable es comprimida en 7 bits, espacio suficiente para almacenar un número menor o igual a 127.

Archivos de punteros o índices (.ptr)

Estos archivos son los responsables por la conexión entre los elementos de las entidades y sus entidades inferiores. Cada entidad tiene un archivo de punteros o índices, con elementos que "apuntan" desde los elementos de la entidad superior hasta los elementos de la entidad misma.

Pertenencia

El concepto de pertenencia es fundamental para entender el funcionamiento de los archivos de datos. En una estructura jerárquica, la pertenencia indica la relación de subordinación entre las entidades y sus elementos. La figura 4, muestra una estructura jerárquica genérica con cuatro entidades: PAIS, PROVINCIA, DISTRITO, VIVIENDA

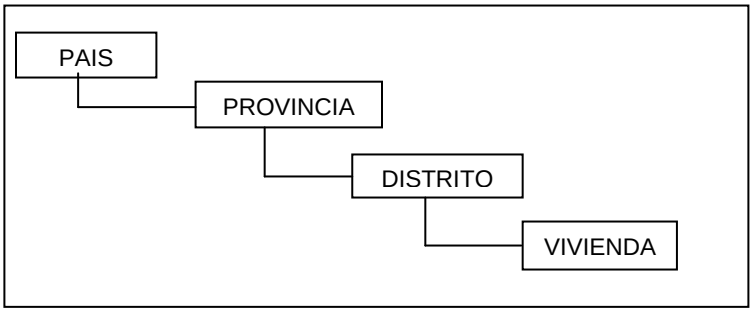

Figura 4. Estructura jerárquica

Analizando la figura 4 podemos decir que la entidad VIVIENDA “pertenece” a DISTRITO, que a su vez "pertenece" a *PROVINCIA*, que "pertenece" a *PAIS*, y esta, por definición, "pertenece" al sistema.

El concepto de "**pertenencia**", por ejemplo para una división político-administrativa, puede ser entendido como: la provincia *3* "del" país, el distrito *14* "de" la provincia *2*, la vivienda *358* "del" distrito *21*. Cada entidad puede tener un número distinto de elementos (casos), es decir, *x* distritos, *y* provincias, etc.

Para aplicar el concepto de pertenencia en una base de datos jerárquica, REDATAM asigna números internos secuenciales a cada una de las entidades a sus elementos y sus respetivas variables. Por definición, la numeración de las entidades comienza con cero siendo la raíz siempre la entidad 0 y la entidad "raíz" tiene **SIEMPRE** un único caso, o sea, es como si fuera el país.

Para efecto de ejemplo, en la figura 4 se tiene PAIS como la entidad 0 (entidad raiz), PROVINCIA entidad 1, DISTRITO la entidad 2 y VIVIENDA la entidad 3.

Es importante mencionar que cada entidad puede tener diferente cantidad de elementos. Si retomamos la tabla 3, se observan los siguientes números de casos para las entidades: PROVINCIA tiene 5, DISTRITO tiene 234 y VIVIENDA tiene 4,587 elementos (numerados del 1 al……..4,587).

Supongamos como muestra la tabla 4 y 5 las siguientes distribuciones, es decir, la primera provincia tiene tres distritos, la segunda cinco distritos, la tercera cuatro distritos, etc., y a su vez el primer distrito tiene 300 viviendas, el segundo 250 viviendas, y el tercer distrito tiene 130 viviendas, estos tres distritos pertenecientes a la provincia 1, luego sigue la lista de distrito de la provincia 2 y así sucesivamente.

PROVINCIA	DISTRITO
1	3
2	5
3	4
4	6
5	5
TOTAL	23

Tabla 4. Distribución de Distritos por Provincias

DISTRITO	VIVIENDA
1	300
2	240
3	130
…	…
23	90
TOTAL	4,587

Tabla 5. Distribución de Viviendas por Distritos

Almacenamiento de los datos de entrada

Para migar una base de datos hacia REDATAM estos datos deben estar almacenados en uno de estos formatos CSPro (cdf), SPSS (sav), dBASE (dbf), CVS y ASCII (txt) y deben estar relacionados a un diccionario de datos. Ahora bien, una vez realizada a migración a REDATAM estos datos son convertidos y almacenados como archivos de datos binarios propios de REDATAM (rbf) junto con un diccionario (dicX) y los punteros (ptr).

Por otra parte los registros en un archivo externo, se pueden almacenar de dos maneras: con un formato único (SPSS, dBASE, CVS) o con un formato múltiple (CSPro):

Formato único o simple

El archivo de entrada tiene **un sólo tipo de registro**, esto es, todos los registros tienen el mismo formato e identifican a una misma entidad. En este caso, REDATAM asume que ellos pertenecen solamente a una entidad, por ejemplo PERSONA. Es posible mezclar variables de diferentes entidades en un archivo con un formato único siempre y cuando la entidad inferior este presente. Por ejemplo, si el archivo con formato único representa a personas, pero además se incluyen variables de la vivienda, todas las variables de la vivienda (luz, agua, piso, techo) deberán duplicarse en cada registro correspondiente a las personas dentro de esa vivienda (ver Tabla 6).

Identificación					Variables			
PROV								
	COM							
		DIST						
			VIV					
				PER				
ID	ID	ID	ID	ID	LUZ	AGUA	SEXO	EDAD
01	01	001	01	1	1	1	2	23
01	01	001	01	2	1	1	1	33
01	01	001	01	3	1	1	2	12
01	01	001	02	1	2	3	1	56
01	01	001	02	2	2	3	2	52
01	01	001	02	P	2	3	2	16

Tabla 6. Archivo de entrada de registros únicos (rectangular)

Para efectos de simplificación, a las entidades de índole geográfica (PROV, COM, DIST) solo se les ha asignado el código identificador. Este código es utilizado para mantener la **pertenencia** de una persona a una vivienda a un distrito a una comuna y a una provincia.

En el ejemplo se muestran las primeras personas pertenecen a la vivienda 01 y las otras tres personas siguientes pertenecen a la vivienda 02. Ambas viviendas pertenecen al distrito 001 que a su vez pertenece a la comuna 01 de la provincia 01. Por eso hay tanta repetición de códigos.

Formato múltiple

Por otra parte, si se está tratando con un archivo de formato múltiple, tales como aquellos encontrados en encuestas de hogares, censos de población y agrícolas, **se tienen dos o más tipos de registros** en un mismo archivo. Por ejemplo, cuando se tiene un registro (una línea) para los datos de VIVIENDA y otro registro (una línea) para los datos de PERSONA, este archivo debe tener una columna específica (la primera) para identificar el tipo de registro.

La tabla 7, a continuación muestra un ejemplo de un archivo de registros múltiples con dos tipos de registros señalados en la primera columna, "1" para vivienda y "2" para persona.

Tipo-reg	Identificación			Variables
	PROV			
		COM		
			DIST	
				VIV-PER
	ID	ID	ID	ID…….
1	01	01	001	Vivienda 1
2	01	01	001	Persona 1
2	01	01	001	Persona 2
2	01	01	001	Persona 3
1	01	01	001	Vivienda 2
2	01	01	001	Persona 1

Tabla 7. Archivo de entrada de registros múltiples (tipo censos)

Nuevamente para efectos de simplificación, a las entidades de índole geográfica (PROV, COM, DIST) solo se les ha asignado el código identificador. Este código es utilizado para mantener la **pertenencia** de una persona a una vivienda a un distrito a una comuna y a una provincia.

En este caso, cada tipo de registro (el tipo 1 para vivienda, y el tipo 2 para personas) será referido a una entidad en la base de datos en formato REDATAM y los datos en cada tipo de registro serán cargados ordenadamente sólo a esa entidad. Cada tipo de registro y

por ende la línea entera debe tener variables pertenecientes sólo a la entidad específica a que se refiere, más las variables de identificación (pertenencia) de las entidades de niveles superiores en la jerarquía como DISTRITO, COMUNA y PROVINCIA.

Para las entidades que no tienen variable de identificación, como es VIVIENDA y PERSONA, el sistema REDATAM "sabrá" cuándo crear un elemento detectando su tipo de registro en el archivo de entrada. Esto es, usando el mismo ejemplo de un típico archivo de censos, cada tipo de registro "1" creará un elemento vivienda para la entidad VIVIENDA, y los registros tipo "2" que vienen después de este registro tipo "1" creará registros de personas para la entidad PERSONA manteniendo la pertenencia a la vivienda previa. Cuando encuentre otra vez un registro tipo "1"cerrará la vivienda previa y abrirá una nueva vivienda para contener a las personas que pertenecen a esta nueva vivienda. Las viviendas sin personas no tendrán registros tipo "2" a continuación de su registro tipo "1". Para detectar aquellas viviendas sin personas se aconseja trabajar con registros de formato múltiple, ya que con el formato único se crearía de todas maneras una entrada para una persona aunque las variables se consideraran con valores no aplica.

Es importante notar que el registro de vivienda tiene que venir ANTES de los registros de personas correspondientes. Es decir el archivo debe tener los registros ordenados en la forma vivienda-hogar- persona y no al revés. Se sugiere por ende marcar con "1" a las viviendas con un "2" a los hogares y con un "3" a las personas.

Solo el programa **CSPro trabaja con estructura de registros múltiples**.

El archivo de formato de registros múltiple es algunas veces llamado archivo "jerárquico", porque existe una cierta "jerarquía" entre los registros tipo "1" y "2", esto es, los registros tipo "2" "dependen" del registro previo tipo "1". Sin embargo, este no es realmente un archivo jerárquico, porque la parte de identificación (generalmente la información geográfica) es duplicada para todos los registros. Un verdadero archivo jerárquico tendría un tipo de registro por cada nivel de información (entidad), como en la tabla 8.

Tipo-reg	Entidad
1	Provincia
2	Comuna
3	Distrito
4	Vivienda 1
5	Persona 1
5	Persona 2
5	Persona 3
4	Vivienda 2
5	Persona 1

Tabla 8. Representación de un archivo múltiple jerárquico

Cada entidad tendría un tipo de registro específico para almacenar sus datos. La información "geográfica" no sería repetida para cada registro en el archivo, pero sería almacenada sólo en el tipo de registro específico. Por ejemplo, la información de

provincia (como su código, nombre u otra información de ese nivel) sería almacenada sólo en registros tipo "1".

Los archivos de registros múltiples no son manejados por muchos programas, (mencionamos aquí CSPro). Sin embargo, este tipo de archivo es muy útil para ahorrar espacio.

IV. Principios básicos de una base de datos REDATAM

Estructura lógica de REDATAM

Estos conceptos son importantes para entender exactamente la construcción de una base de datos REDATAM muy compleja. Para un usuario común, el contenido interno de la base debiera ser transparente.

Una base de datos REDATAM puede ser definida lógicamente en tres partes, cada una con su función específica: el diccionario, los punteros y los datos. Las dos primeras partes son consideradas como partes estructurales de una base de datos, mientras que la última es la parte que contiene los datos propiamente tales. Cada uno de estos grupos está compuesto por uno o varios archivos.

Número interno

Otra información almacenada en estos archivos, pertinente a las explicaciones que vienen a continuación, es el número interno de las entidades y variables de la base. Cada una de las entidades y variables posee, además de su nombre por el cual son conocidas en los procesos estadísticos, un número interno único, asignado por el sistema en la creación del Diccionario. Por definición, la primera entidad (entidad "raíz", con el nombre de la base de datos), es la **0**. Estos números pueden ser consultados (pero no cambiados) por el usuario en **el Diccionario de Datos**.

Punteros

Estos archivos, también llamados de "**punteros**", son los responsables por la conexión física entre los elementos de las entidades y sus entidades inferiores. Cada entidad tiene un archivo de índices, con elementos que "apuntan" desde la entidad superior hasta la entidad misma.

Ellos tienen la extensión "**.ptr**" (de "puntero"), y el nombre de cada uno de ellos es formado por el nombre de la base de datos, agregándose el número interno de la entidad, completado con ceros a la izquierda.

Todas las entidades en la base de datos (inclusive la entidad "raíz") tienen un archivo ".ptr" asociado. Para entender el funcionamiento de estos archivos, tomemos como ejemplo la base de datos mostrada en la Figura 5.

Esta base tiene cuatro entidades, y por lo tanto, deberá tener cuatro archivos "**.ptr**". Para simplificación, supongamos que las entidades tengan números internos correlativos empezando de cero (*PAIS* es la entidad *0*, *PROVINCIA* es la *1*, *DISTRITO* es la *2*, *VIVIENDA* es la *3*). La entidad *VIVIENDA* "pertenece a" (o depende de) la entidad *DISTRITO*, que a su vez "pertenece" a *PROVINCIA*, que "pertenece" a *PAIS*, y esta, por definición, "pertenece" al sistema.

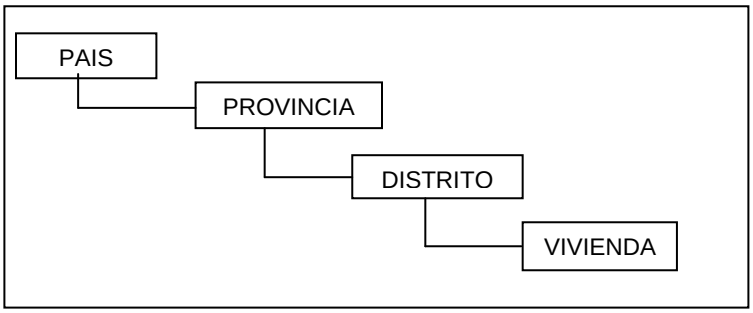

Figura 5. Estructura jerárquica

El concepto de "pertenencia", por ejemplo para una división político-administrativa, puede ser entendido como: la provincia *3* "del" país, el distrito *14* "de" la provincia *2*, la vivienda *358* "del" distrito *21*. Cada entidad puede tener un número distinto de elementos (casos), es decir, *x* distritos, *y* provincias, etc. Para efecto de ejemplo, supongamos que se tenga los siguientes números de casos para las entidades:

ENTIDAD	ELEMENTOS
PROVINCIA	5
DISTRITO	23
VIVIENDA	4.587

Por definición, la entidad "raíz" tiene **SIEMPRE** un único caso, o sea, es como si fuera el país. Supongamos también las siguientes distribuciones, es decir, la primera provincia tiene tres distritos, la segunda cinco distritos, etc., y el primer distrito tiene 300 viviendas, el segundo 250 viviendas, etc.

PROVINCIA	DISTRITO
1	3
2	5
3	4
4	6
5	5

Figura 6. Distribución de Distritos por Provincias (pais0002ptr)

DISTRITO	VIVIENDA
1	300
2	240
3	130
…	…
23	90

Figura 7. Distribución de Viviendas por Distritos (pais0003ptr)

Contenido lógico

El concepto lógico de los archivos de índices es un vector de contadores, en donde cada registro contiene el número de casos de la entidad inferior que "pertenecen" a cada elemento de la entidad superior. En otras palabras, cada archivo índice tiene tantos elementos cuantos sean los elementos de la entidad "que apunta", y cada elemento de estos contiene el número de casos de la entidad "apuntada".

Por ejemplo, el contenido lógico del archivo de índices de *DISTRITO*, o sea, *pais0002ptr*, es exactamente la columna de la derecha en la Figura 6, es decir, el archivo tiene cinco registros lógicos (uno para cada provincia), con los valores, 3, 5, 4, 6 y 5. El archivo *pais0003.ptr* (para la entidad *VIVIENDA*) corresponde a la Figura 7, con 23 registros lógicos conteniendo respectivamente los valores 300, 240, 130, etc.

Se puede decir que un archivo de punteros, a pesar de llevar el número de la entidad hacia la cual él apunta, funciona como un archivo de datos (vea **Datos** más adelante), de la entidad superior, porque tiene el mismo número de elementos de esta entidad.

Contenido físico

En verdad, el contenido físico de los archivos índices es un poco distinto, asemejándose más a un concepto de punteros que de número de elementos. En vez del número de elementos, cada registro contiene el número correlativo (empezando de cero) del primer elemento de la entidad inferior. Además, existe siempre un elemento adicional, al final del archivo, con el número total de casos de la entidad inferior. Refiriéndose a la Figura 6 anterior, el contenido físico del archivo *pais0002ptr* para la entidad *DISTRITO* sería como en la Figura 8, y para la entidad *VIVIENDA* sería como en la Figura 9.

REGISTRO	CONTENIDO
1	0
2	3
3	8
4	12
5	18
6	23

Figura 8. Archivo PAIS0002.PTR

REGISTRO	CONTENIDO
1	0
2	300
3	550
4	680
…	…
23	4497
24	4587

Figura 9. Archivo PAIS0003.PTR

Para tener acceso a los elementos de la entidad inferior, el sistema usa entonces dos registros, el **registro$_n$** para el inicio de los elementos, y el **registro$_{n+1}$** para calcular, por substracción, el número de casos. Por ejemplo, los distritos de la provincia *3* empiezan en el distrito *7* (en el registro está *8*, pero hay que substraer *1* porque se empieza con el elemento cero), y son 4 (12 - 8).

La Figura 10 muestra un ejemplo de esta relación en forma gráfica, en donde se puede apreciar que cada entidad tiene variables (archivos .rbf, vea más adelante **Datos**), dentro de las cuales se dispone de una variable que contiene un código para los elementos de aquellas entidades seleccionables. Además se puede tener una variable que contiene un nombre asociado al código. Para recorrer la estructura jerárquica de las entidades se disponen de los índices (archivos .ptr) que establecen la conexión entre un elemento de una entidad superior con sus elementos de la entidad inferior, y se pueden visualizar como las flechas de la figura.

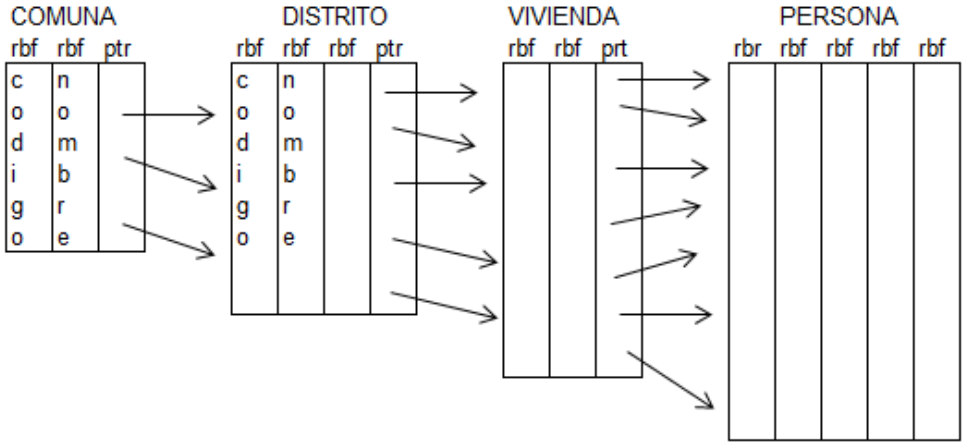

Figura 10. Relación entre punteros y variables

Los archivos de índices son creados en tiempo de generación de la base de datos. Cada registro (elemento del archivo) tiene un largo lógico de 4 bytes, y su contenido NO está en formato ASCII, el formato es propietario de REDATAM y por ende NO pueden ser leídos fuera de aplicaciones desarrolladas con REDATAM.

Datos

REDATAM guarda cada variable de una entidad en su propio archivo comprimido que tienen la extensión "**.rbf**" (de "binarios de REDATAM"), con un "registro" por cada elemento de la entidad, de ahí que cada variable puede ser vista como un vector de datos (tales archivos se conocen como archivos transpuestos). Existe entonces un archivo para la variable sexo, *pais0032.rbf*, otro para la variable tipo de vivienda *pais0005.rbf*, etc.

Cada archivo contendrá tantos registros (elementos) cuantos sean el número de casos de la entidad a la cual pertenezca la variable. Reportándose al ejemplo anterior descrito para los archivos de **Índices**, una variable a nivel de *PROVINCIA* tendrá *5* elementos, y una variable a nivel de *VIVIENDA* tendrá *4.587* elementos.

Una de las ventajas de este tipo de archivo es que, como cada registro (es decir, cada valor de la variable, o cada elemento del vector) tiene el mismo tamaño, se puede usar una técnica de compresión de datos muy sencilla para ahorrar espacio de almacenamiento y mejorar la eficiencia en la lectura de sus datos de disco a memoria, ya que hay que transferir menos volumen de información.

El proceso de compresión consiste en transformar en números binarios los valores de los elementos de cada archivo. El número de "**bits**" necesarios para cada variable depende del valor máximo que la variable pueda contener (también definido por el usuario). Una variable, por ejemplo sexo, que contiene los valores *1* y *2* (para hombres y mujeres respectivamente), es comprimida en *2* bits (con *2* bits se puede almacenar un número menor o igual a *3*). La variable edad, que puede variar, por ejemplo de **0** hasta *125*, necesita de *7* bits (almacena un número menor o igual a *127*), etc.

CREANDO UNA BASE REDATAM

V. Etapas preliminares

Una base de datos en formato REDATAM puede tener una estructura muy sencilla, con sólo una rama de información, o tan compleja como pueda imaginarse, con múltiples ramas en el árbol de la jerarquía. Puede ser muy pequeña en tamaño (sólo unos cuantos registros), o muy grande, incluyendo datos con cientos de registros. Obviando cuál es su caso, antes de comenzar el proceso de generación, hay que tomar en consideración dos cosas (1) garantizar que los datos de entrada están en un formato aceptado pro REDATAM y (2) verificar que los datos de entrada estén ordenados según la estructura deseada.

Formato de la base de datos de entrada

REDATAM puede leer archivos en los formatos de CSPro (dat y dcf), CSV (comma Separated Values), Dbase (dbf) o SPSS (sav). Si su archivo de entrada está almacenado en un tipo especial de formato, tales como aquellos usados en paquetes comerciales, por ejemplo, EXCEL, SAS, Oracle, SQL-Server, etc., se tienen que convertir todos los archivos de la base de datos (datos y diccionario) a uno de los formatos aceptados.

Verificación de los datos de entrada

El segundo paso preliminar para generar una base de datos en REDATAM es analizar el contenido de los archivos de datos originales, la comprobación de su consistencia (por ejemplo, que no existan casos duplicados con el mismo código de identificación, que no existan hogares sin un jefe de hogar, etc.). Este paso es fundamental para el éxito del proyecto, ya que si los datos originales no muestran la frecuencia requerida y / o son inconsistentes, la calidad de la base de datos se verá comprometida.

VI. Etapas básicas

Consideraciones generales

La generación de una base de datos en el módulo Red7 Create implica tres pasos básicos: (i) la definición de la estructura jerárquica; (ii) la definición de variables para el nuevo diccionario; y (iii) la carga de la base de datos misma. A continuación se explica en detalle cómo funciona cada paso. Estas definiciones se guardan en un esquema correspondiente a un archivo con la extensión .wipX. Este esquema es el paralelo del diccionario que se usara en la base final, es decir, se utiliza como referencia para la generación de la base de datos en formato REDATAM.

Definición de la estructura jerárquica

El proceso de generación de una base de datos en formato REDATAM comienza por establecer las entidades de la jerarquía y sus variables en la ventana del esquema.

Al comenzar la creación de un diccionario se selecciona la opción "Nuevo" desde el menú principal, para abrir un esquema o ventana nueva, en donde se definirá la estructura jerárquica. Luego bajo el menú datos de origen se selecciona bajo “conectar” los datos de origen y así importar el diccionario de los datos originales (según el formato, ya sea CSPro, SPSS, DBF, o CSV**)**

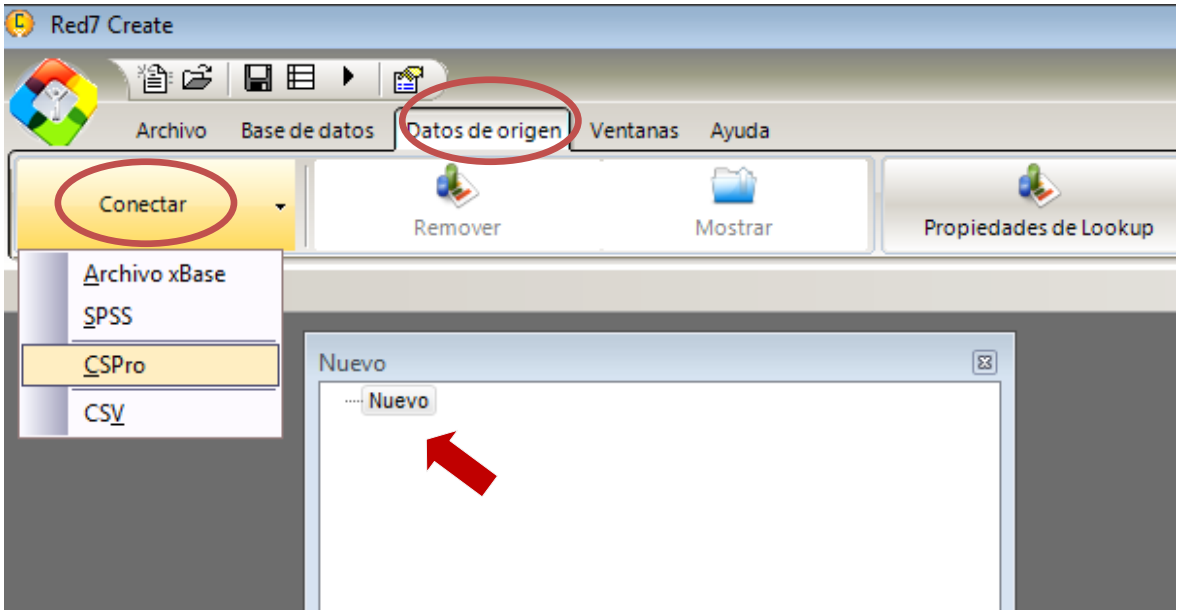

Figura 11. Menú de opciones de formatos para importar una definición de datos externa

En la Figura 11 se muestra el menú con las opciones para abrir una definición o diccionario de datos de origen dependiendo del formato en que ésta se encuentre (DBF, SPSS, CsPro o CSV) como se mencionó anteriormente.

En el ejemplo que se muestra en la Figura 12, se ha seleccionado un archivo de datos en formato xBase, el cual se encuentra en la carpeta de instalación de *Redatam7*, en la carpeta Samples (*Single_Input_Esp.dbf*).

En la ventana de la izquierda se definen las entidades que forman la estructura jerárquica de la base de datos en formato REDATAM (por ejemplo *DEPTO*, *PROVIN*, *MUNIC*, *DISTRITO*, etc.).

Nomes das entidades

En RED7 Create existe libertad en asignar los nombres a entidades y variables, además se puede asignar alias o segundos nombres a cada entidad o variable para mantener el nombre original y poder utilizar un segundo nombre. Sin embargo hay ciertas reglas a cumplir:

1. Los nombres pueden tener cualquier largo, se sugiere no sobrepasar 10 caracteres.
2. Pueden contener mayúsculas y minúsculas
3. Se sugiere no usar caracteres especiales (como @, #, $, _), aunque éstos están permitidos.
4. No pueden usarse caracteres en blanco (espacios), bajo ninguna circunstancia.

Las ramas o niveles de la estructura jerárquica también se pueden borrar o cortar y pegar en otra localización, por debajo de otro nivel o como hermano de otra entidad. Para ello, basta con hacer clic con el botón derecho en la rama deseada, seleccione "Cortar rama" (CTRL X), posicionarse en el nuevo nivel y seleccionar "Pegar rama debajo de" o "Pegar rama arriba de."

En la ventana de la derecha se despliegan la información en la parte General del formato del archivo y la ruta en donde se encuentran los datos, y en la parte de campos se despliegan las variables existentes en el archivo dBase, que caracterizan a las entidades *HOGAR* y *PERSONA*, y forman finalmente la base de datos, Figura 12. Recuerde guardar el esquema correspondiente a un archivo con la extensión .wipX asi pasa del nombre “nuevo” al nombre final del esquema.

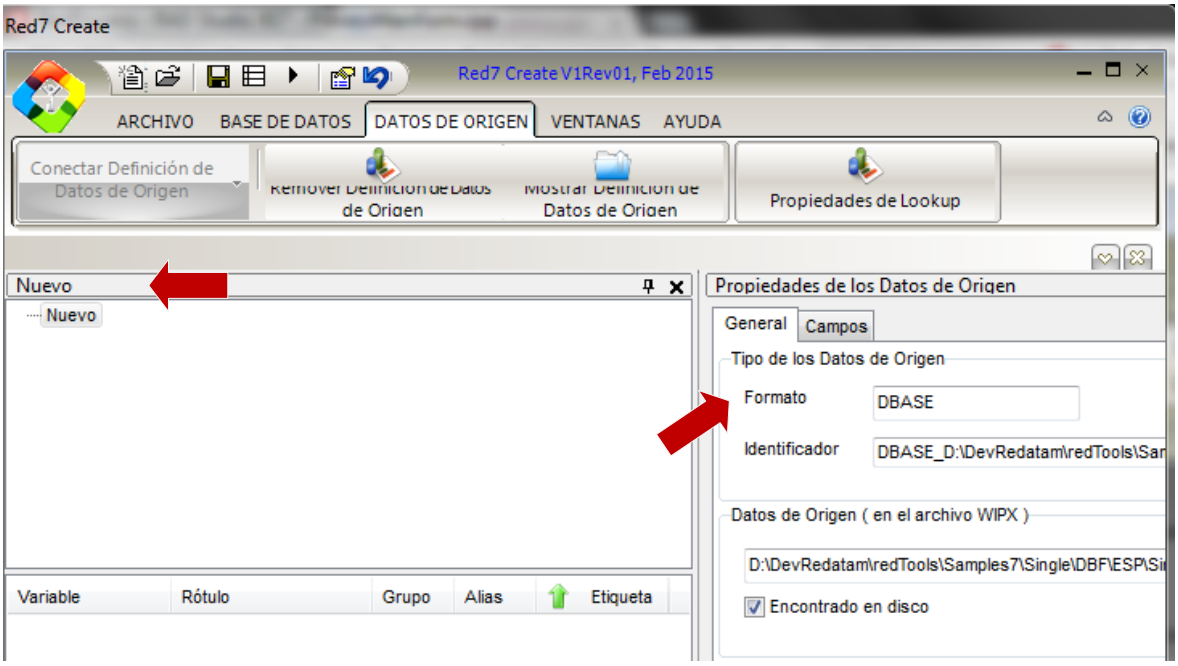

Figura 12. Definición proveniente de un archivo de datos xBase (.dbf).

Primero se selecciona la entidad a la cual se le va a agregar uno o varios campos y luego se arrastran desde la ventana de la derecha hacia el sector inferior de la ventana izquierda como muestran la figura 13.

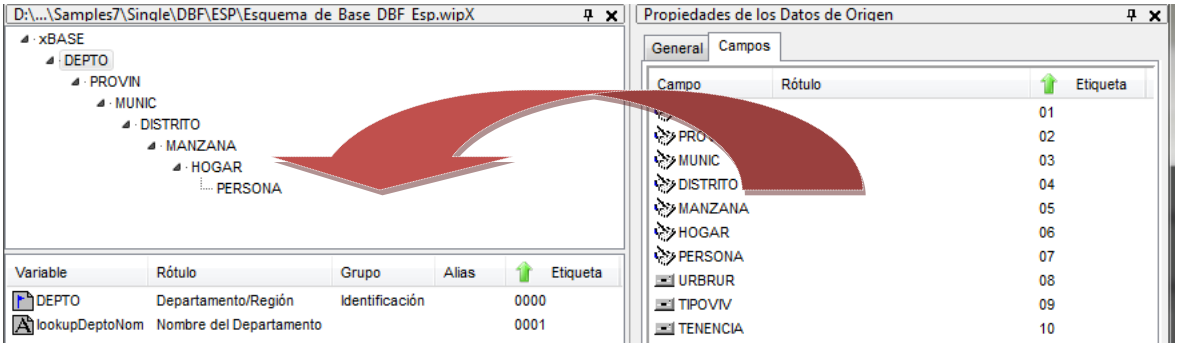

Figura 13. Estructura jerárquica - entidades y variables.

Para cada entidad se debe identificar cual de las variables contiene el código identificador (ordenador) de la entidad, en la Figura 13, para la entidad *PERSONA* sería PERSONA, esta variable debe ser de tipo carácter como se puede apreciar al abrir las propiedades de la variable PERSONA en la Figura 14. Si esta variable no fuera de tipo caracter se podría modificar posteriormente en el esquema REDATAM.

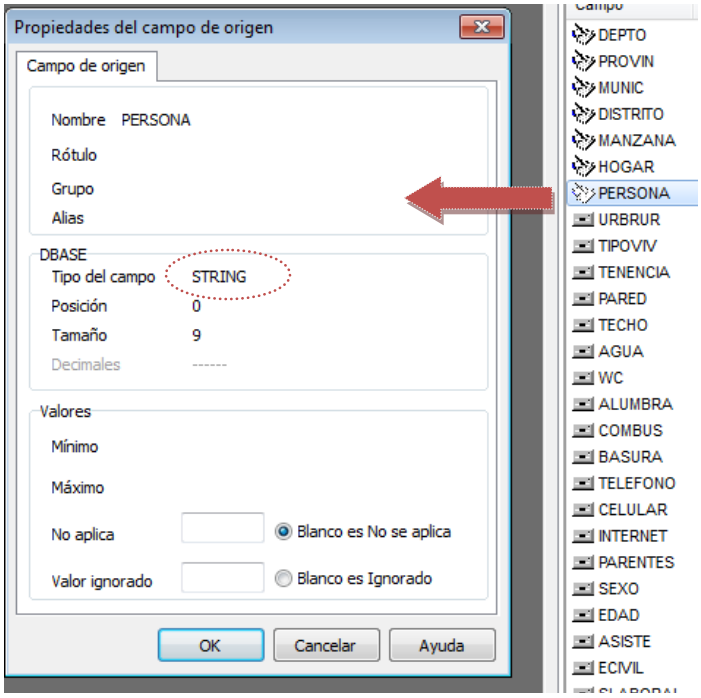

Figura 14. Propiedades de la variable original.

Una vez arrastrada la variable a la ventana de PERSONA se procede a definirla como código de la entidad, esto se hace abriendo la ventana de propiedades de la entidad PERSONA desde el menú dinámico (haga clic en el botón derecho del mouse sobre el nombre de la entidad PERSONA) en la ventana que se abre elija la variable PERSONA desde el casillero “Variable de códigos” como se muestra en la Figura 15.

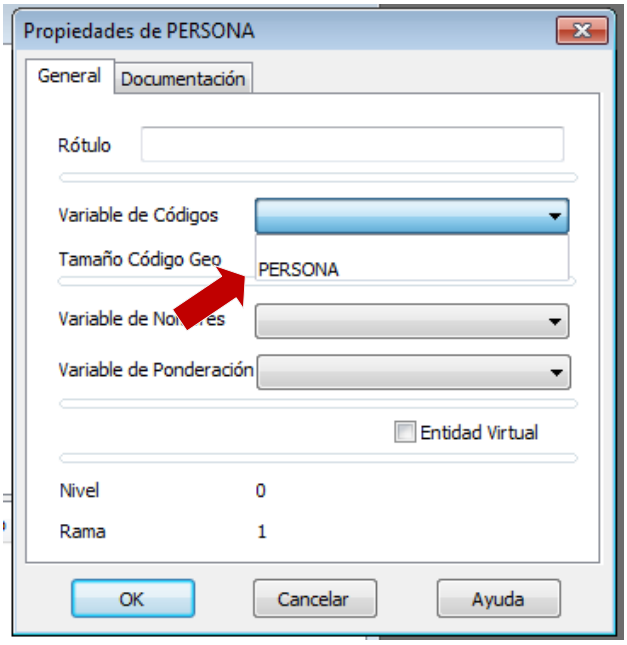

Figura 15. Menú de entidad – código de la entidad

Luego marque el casillero “Seleccionable” para distinguir a la entidad como entidad seleccionable como muestra la figura 16

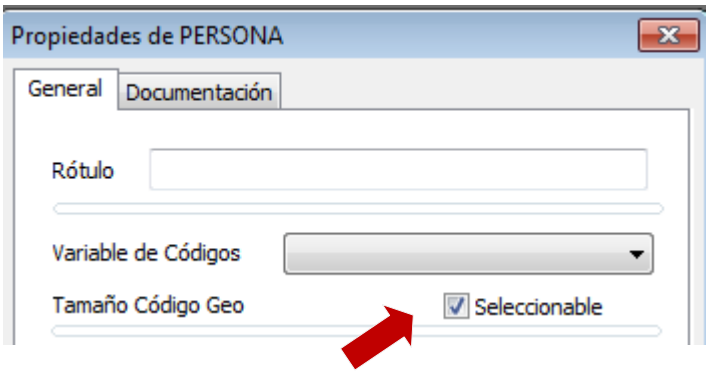

Figura 16. Propiedades de la entidad – entidad seleccionable

Repita este proceso para cada entidad de la estructura jerárquica y guarde el esquema.

Luego se procese a arrastrar el resto de variables desde la venta de origen hacia la ventana del esquema asignando la o las variables a cada grupo o entidad (las variables de vivienda de hogar o de personas según corresponda).

En la Figura 17 se muestran las variables de la entidad *PERSONA*. Luego se abrió la ventana de propiedades para la variable SEXO, en donde se define su rango, tipo, categorías, etc. La variable sexo contiene las categorías 1 para Hombre y 2 para Mujer.

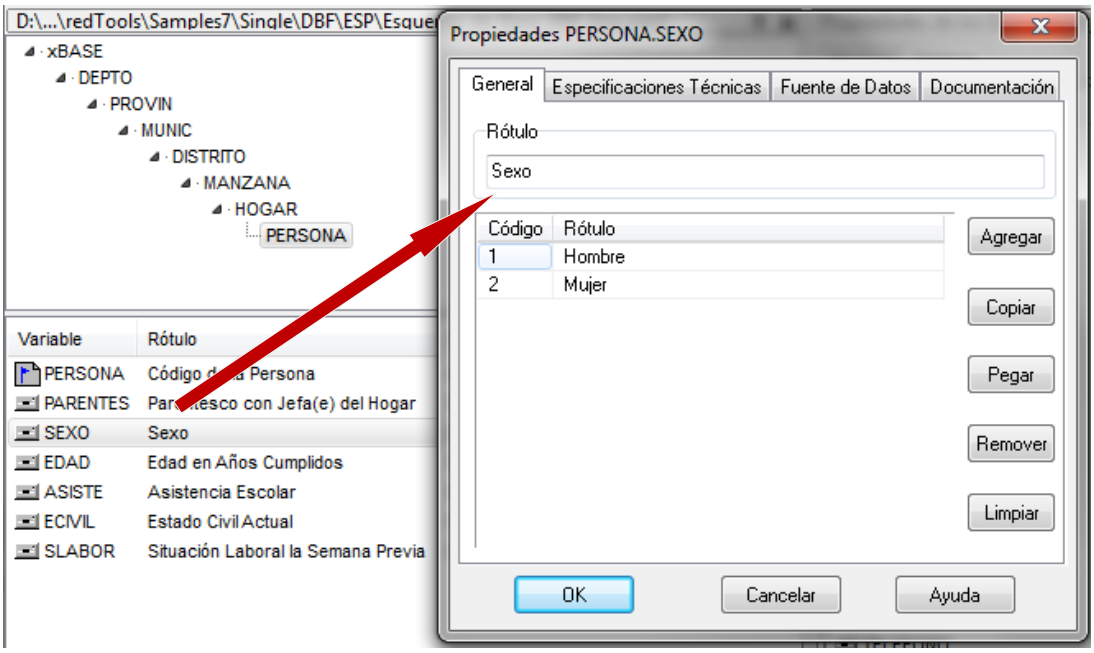

Figura 17. Propiedades de una variable.

Generación de la base de datos

A continuación se procede a generar la base de datos REDATAM. Para esto, seleccione la opción “Ejecutar Creación” desde el menú “Base de Datos” como muestra la Figura 18, también puede utilizarse la tecla de función F9.

Figura 18. Generación de una base de datos REDATAM.

La Figura 19 muestra la ventana para iniciar el proceso de generación de la base de datos, es necesario definir todos los parámetros que se incluyen en cada una de las páginas de esta ventana.

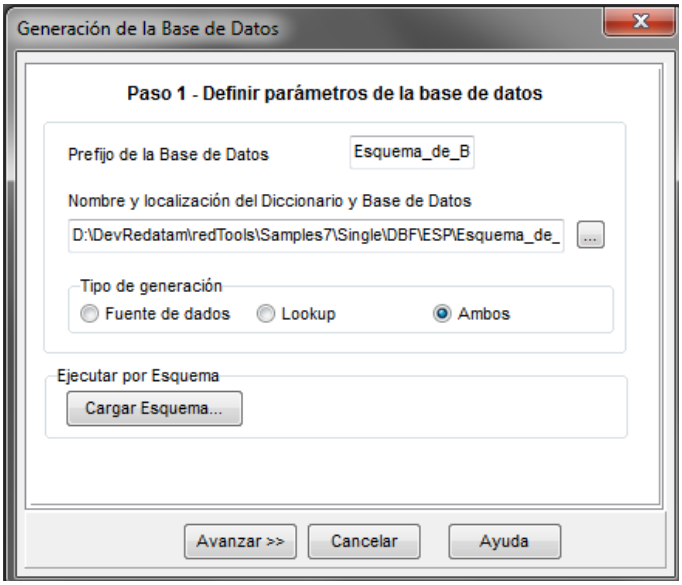

Figura 19. Generación: Paso 1 – Definir Parámetros de la base de datos.

El primer paso consiste en definir un nombre de la base de datos a generar, el directorio en el cual se van a localizar los archivos de variables (.rbf), entidades (.ptr) y diccionario (.dicX).

El segundo paso, Figura 20, consiste en definir el formato y archivo de entrada de los datos.

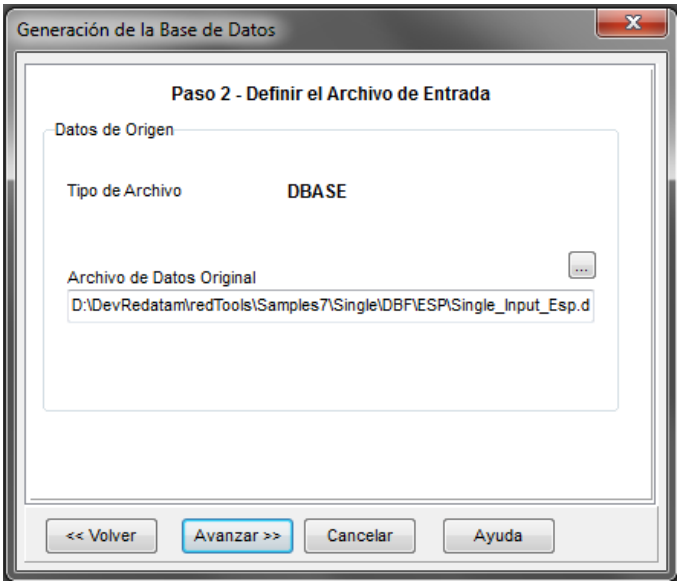

Figura 20. Generación: Paso 2 – Definir el Archivo de Entrada.

El tercer paso, Figura 21, consiste en definir para cada entidad si se crea o no un archivo de punteros y generar el **código compuesto** asociado a la entidad; también permite verificar la secuencia de los registros según el código geográfico identificador, para lo cual es necesario que esté activado el recuadro “Verificar Secuencia de Datos”.

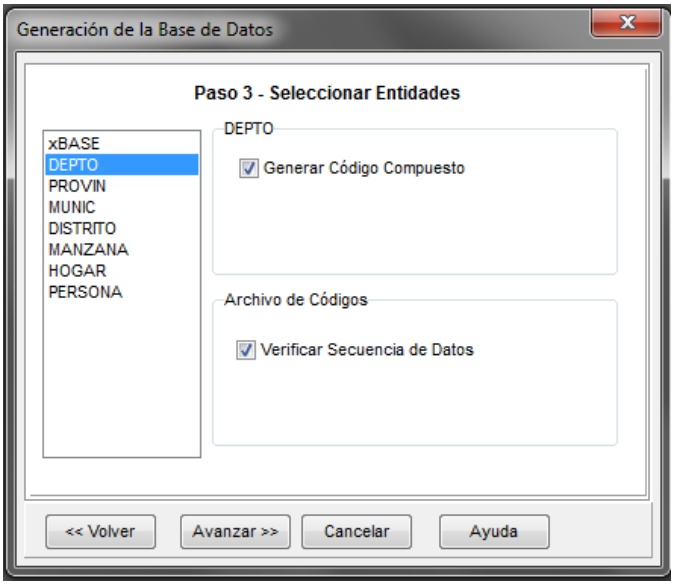

Figura 21. Generación: Paso 3 – Seleccionar Entidades.

En el cuarto paso, Figura 22, se debe definir las variables a incluir en la generación de la base de datos en formato REDATAM para cada entidad. Utilizando el mouse seleccione la entidad que contiene las variables y marcando en cada casillero para elegir las variables o en el botón “Seleccionar Todo” en el caso de todas las variables de la entidad seleccionada. Adicionalmente tiene la opción de “Seleccionar Todo de Todo” para marcar todas las variables existentes en el esquema de creación.

Figura 22. Generación: Paso 4 – Seleccionar Variables.

En el quinto paso y final, Figura 23, se puede almacenar los parámetros definidos en los pasos anteriores en un archivo especial (*.lsdX) con la opción "Guardar Esquema ", para no tener que repetirlos en una generación posterior de la misma base de datos.

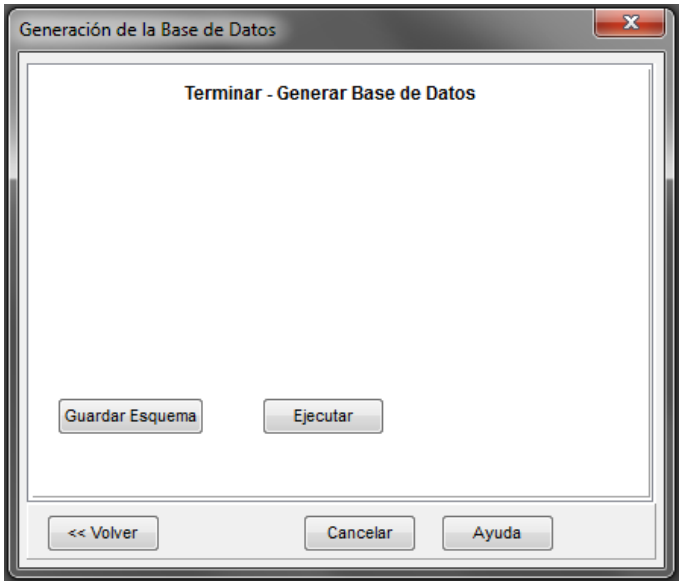

Figura 23. Generación: Terminar – Generar Base de Datos.

Finalmente, concluir el proceso de generación utilizando el botón “Ejecutar”.

Una vez terminado el proceso de generación de la base de datos, la ventana de ejecución se cierra y se despliega el Informe de Generación, el cual debe ser revisado minuciosamente para detectar posibles inconsistencias como: valores fuera de rango, valores no aplica, total de registros para cada entidad, etc. Figura 24.

Figura 24. Informe de generación – Cada archivo generado es indicado.

Se recomienda guardar el informe (archivo Excel) junto con la definición utilizada en la creación de la base de datos (.wipX).

Si el programa termina satisfactoriamente, significa que la base de datos ha sido creada así como el archivo de diccionario *Redatam7* (.wipX).

Para revisar la base de datos generada se debe utilizar el **módulo Red7 Process.** Si se desea utilizar la base de datos con la versión previa de REDATAM (*Redatam+SP*), se debe importar el diccionario desde su versión ASCII (archivo wxp).

Archivos creados por REDATAM

Debe mencionarse que todas las variables están comprimidas en formato binario. Es decir, la información correspondiente a las variables se han almacenado en archivos codificados (*.rbf), uno por variable. La información correspondiente a la pertenencia de una entidad a otra entidad superior se ha almacenado en los archivos de punteros (*.ptr), uno por entidad.

La definición de las variables y las entidades se ha almacenado en un archivo de definición (*.wipX), así como también se ha generado un archivo con el esquema de generación (*.lsdX) y no olvidemos el Informe de Generación (.xls).

Resumen de pasos a seguir

La creación de una base de datos en formato REDATAM involucra tener en cuenta los siguientes aspectos de manera secuencial:

1. Seleccionar el archivo original, verificando que su formato sea uno de los aceptados por *Redatam7*.
2. Análisis visual del archivo original para determinar las entidades y variables.
3. Analizar todos los campos del archivo original para determinar la estructura.
4. Abrir el programa Red7 Create.
5. Crear la jerarquía.
    a. Nombrar la entidad raíz teniendo en cuenta el objetivo de los datos, con un tamaño recomendado de 8 caracteres.
    b. Utilizar nombres de entidades que sean fácil de identificar por parte del usuario.
6. Guardar el Esquema de Creación (.wipX).
7. Abrir la definición del archivo de datos que corresponda.
8. Traspasar cada uno de los campos a su entidad respectiva.
9. Controlar que todos los campos se hayan pasado a la jerarquía (mostrar campos referidos).
10. Definir las entidades seleccionables (seleccionar la variables STRING que contiene los códigos de la entidad), es recomendable establecer como seleccionables todas las entidades -si se desea generar una base de uso interno y/o verificar la consistencia de los códigos de identificación-.
11. Utilizar la opción de “Verificar Valores” y/o “Verificar Máximo” para cada una de las entidades, guardar los archivos resultantes.
12. Establecer los valores mínimo y máximo (rangos) para las variables de tipo INTEGER y REAL, con ayuda de la información producida en el Paso 11.
    a. Utilizar como valor 0 para NA en el caso de que no se utilice en el rango de valores válidos.
    b. Para el valor MV (omitido) utilizar el valor máximo más uno.
    c. En caso de existir valores para la categoría 0, verificar existencia de valores NA.
13. Identificar las variables numéricas que están definidas como STRING y convertirlas a INTEGER, asignando el rango con ayuda de la información del Paso 11.
14. Verificar el tamaño de las variables de códigos de las entidades (posición relativa y tamaño).
15. Documentar (rótulos para entidades, rótulos y categorías para variables).
    a. En la entidad raíz, el rótulo es el nombre de la estadística que estamos creando.
    b. En la entidad raíz incluir información sobre fecha, autor y lugar de Creación de la base de datos.
    c. Grupo y Alias de las Variables.
16. Guardar el Esquema de Creación para actualizar los cambios.
17. Ejecutar la primera generación de la base de datos.
    Paso 1 – Definir parámetros de la base de datos:
    a. Nombre de la base de datos (nombre físico de los archivos .ptr y .rbf).
    b. Nombre del diccionario *Redatam7* a utilizar con los otros módulos

(Process, xPlan, Web, etc).

c. Localización de la base de datos REDATAM en un directorio diferente al de los archivos originales o de creación (.wipX, .lsdX).

Paso 2 – Definir el archivo de entrada.

Paso 3 – Seleccionar entidades.

Paso 4 – Seleccionar variables.

Seleccionar todas las variables para cada entidad (seccionar todo de todo)

Terminar – Generar la base de datos.

a. Guardar el esquema de generación (archivo .lsdX) en el directorio donde se encuentra el esquema de creación (archivo .wipX).
b. Ejecutar.

18. Examinar el informe de generación.
    a. Comparar con el paso 2 (número total de entidades y variables).
    b. Valores válidos, NA y MV (omitidos).
    c. Verificar que el total de registros de la última entidad corresponda al total de registros del archivo original.
19. En el caso de existir inconsistencias o errores en el informe, hacer los cambios en el esquema de creación (.wipX) y/o en el archivo original.
20. Antes de volver a ejecutar, borrar el contenido del directorio donde se genera la base de datos REDATAM.
21. Ejecutar la generación abriendo el esquema de generación previamente creado (Paso 17).
22. Verificar el informe de generación (Paso 18).
23. Ejecutar los pasos 19 al 22 hasta llegar a obtener la base de datos REDATAM correcta, optimizada y documentada.
24. Guardar el último informe de generación en el directorio donde se encuentre el esquema de creación (archivo .wipX).
25. Ejecutar una frecuencia de todas las variables de la base con Red7 Process.
26. Verificar las frecuencias de las variables y en el caso de inconsistencias volver al paso 19.
27. Respaldar la base de datos REDATAM.
28. Respaldar los archivos utilizados en la creación de la base de datos:
    a. Archivo de datos.
    b. Esquema de creación (.wipX).
    c. Esquema de generación (.lsdX).
    d. Informe de generación (.xls).