Distr.
RESTRINGIDA

LC/DEM/R.64
Serie A, No. 197
Octubre 1989

ORIGINAL: ESPANOL

ARQUITECTURA Y FILOSOFIA DE BASES DE DATOS: EL MODELO REDATAM-PLUS

Documento a ser presentado en el
Seminario sobre Utilización de Bases de Datos
Cuernavaca, México, 27 de noviembre al 1o. de diciembre de 1989

NACIONES UNIDAS / CEPAL
Centro Latinoamericano de Demografía (CELADE)
Casilla 91, Santiago, Chile
Fax: 480252 Tel: 485051

TABLA DE CONTENIDO

ref:\red\texto\redplus.doc R2filos.pub CELamc61 (10/11/1989 <09:34>|)

I. INTRODUCCION

1. Este documento tiene dos propósitos principales, relacionados con el sistema REDATAM 1/. El primero, es presentar una descripción más técnica del sistema, su diseño interno, "modus operandi" y algunas características particulares. El segundo propósito, es el de intentar definir una filosofía y criterios de utilización del sistema en un ambiente de producción de datos estadísticos.

2. El sistema REDATAM, desarrollado por el CELADE 2/, tiene su versión 3.1 y su utilización ha sido establecida en unos 18 países de América Latina y el Caribe 3/, pero salvo algunas excepciones, todas las bases de datos son de censos de población y vivienda, y casi todas centralizadas en las oficinas nacionales de estadística. Con el desarrollo de su nueva versión (REDATAM-Plus o REDATAM+), prevista para estar disponible en el segundo semestre de 1990, el CELADE espera contribuir aún más a su difusión y aceptación, pero además espera también que esto establezca en verdad una herramienta de dominio general, tanto en el ámbito del tipo de datos (fuera de los censos demográficos) como en el del tipo de usuario (fuera de las oficinas nacionales de estadística).

3. Volver a escribir sobre su versión 3.1 sería repetitivo, tanto desde el punto de vista de los actuales usuarios, que ya conocen sus ventajas y debilidades, como de la existencia de varios documentos sobre el mismo tema, escritos por el CELADE y por estos mismos usuarios y presentados en varios seminarios pasados. Por otro lado, escribir solamente sobre la versión Plus sería presumir que la versión actual ya es suficientemente conocida y difundida en la región. La solución pareció ser la de hacer una mezcla de las dos estrategias, pero con una concentración mayor de detalles sobre los nuevos aspectos, siempre mirados desde el punto de vista técnico.

1/ REDATAM = REcuperación de DATos para Areas pequeñas por Microcomputador

2/ REDATAM se ha desarrollado gracias al generoso aporte del Centro Internacional de Investigaciones para el Desarrollo (CIID) y con el apoyo del Fondo de Población de las Naciones Unidas (FNUAP) y de la Agencia Canadiense para el Desarrollo Internacional (ACDI).

3/ Arthur Conning y Ari Silva. Microcomputer technology to extend the use of population data in developing countries: Multidisciplinary databases, geographic information systems and REDATAM, IUSSP GENERAL CONFERENCE, 20-27 September 1989, New Delhi, India. LC/DEM/R.62, Serie A-195 (August 1989).

4. Los tres primeros capítulos describen la arquitectura de una base de datos, su implementación y lenguaje de programación. Luego, en los capítulos IV a VIII, se describen los principales módulos del sistema con sus características de mayor relieve. En el capítulo IX, se discute la "filosofía" de utilización de bases REDATAM y en el capítulo X, se intentan algunas conclusiones sobre la materia.

II. ARQUITECTURA DE LA BASE REDATAM

5. El sistema REDATAM posee una estructura de archivos que combinan los conceptos de bases de datos "jerárquicas" y "relacionales", agregando la característica de ser "multidisciplinaria", produciendo un ambiente en el que se buscan las propiedades deseadas de cada modelo. Para que se entienda lo que significa exactamente cada parte de esta definición, es necesario primero que se comente sobre los componentes de una base REDATAM y se definan los conceptos de *variables* y *entidades*.

6. La información almacenada en cualquier base de datos REDATAM está separada por *variables*, que es el ítem elemental común a todas las observaciones de un archivo de datos. Por ejemplo, el sexo de las personas, el tipo de techo de las viviendas, etc. Las variables describen *entidades*, que son conjuntos de objetos lógicos que están organizados jerárquicamente en la base de datos. Una entidad puede ser un grupo de provincias, municipios, casas ó personas. Todas las variables pertenecientes a una entidad poseen ciertos atributos en común, tales como el número de elementos que la componen (para una entidad "provincia", existe un valor de la variable "cantidad de agua caída" para cada una de las provincias específicas) y el mismo nivel jerárquico.

7. Las variables pueden contener microdatos, como en el caso del sexo de una persona, ó datos ya agregados, como el ingreso per cápita de un municipio. Estas variables agregadas pueden provenir de variables microdatos de la misma base de datos como, por ejemplo, el número de mujeres en edad fértil para cada distrito de enumeración, ó pueden provenir de agregaciones de variables externas a la base como, por ejemplo, el precio promedio de un hectárea en la zona rural.

8. La creación de variables agregadas provenientes de variables microdatos de la misma base es una redundancia de información y, por consecuencia, una pérdida de espacio en su almacenamiento. Tiene ventajas cuando su cálculo es muy complejo, o se la necesita muy a menudo. Otro ejemplo de redundancia, es el almacenamiento de la variable edad de la persona y otra variable derivada de ésta, con la categorización de la edad por grupos quinquenales.

9. Las entidades, a su vez, se interrelacionan con otras entidades de una manera jerárquica, en forma de árbol, es decir, una entidad puede tener una ó más entidades subordinadas. Las entidades sin entidades subordinadas se llaman *entidades terminales* (véase Gráfico 1 al final del documento). Como restricción, una entidad sólo puede ser subordinada a una única entidad superior.

10. Las entidades pueden o no ser identificadas en la jerarquía, como en el caso de las entidades llamadas "geográficas" (país, departamento, etc.). En este caso, las entidades tienen un código de identificación y pueden también, opcionalmente, tener un nombre (ej: 01 - Carchi). Como ejemplo de entidades no identificadas, se puede mencionar comúnmente la vivienda y la persona. En el caso de las entidades identificadas, su código (y su nombre) también son consideradas como *variables*.

11. ¿Para qué sirve la identificación de una entidad? Las entidades identificadas tienen la propiedad de ser distinguidas por su código, y por lo tanto, pueden participar del proceso de selección jerárquica. Eso significa que solamente las entidades identificadas pueden ser seleccionadas jerárquicamente, es decir, se pueden elegir para el proceso los departamentos A y B, pero no se podría seleccionar la vivienda X porque no está identificada.

12. Como restricción, en el árbol jerárquico no se puede tener una entidad identificada, subordinada a una no identificada. Por ejemplo, si la entidad persona está subordinada a la entidad vivienda, y vivienda no está identificada, persona tampoco podrá estarlo.

13. La arquitectura de una base REDATAM es considerada *jerárquica* porque las entidades están organizadas en una jerarquía, donde una entidad puede tener cero, una ó más entidades subordinadas.

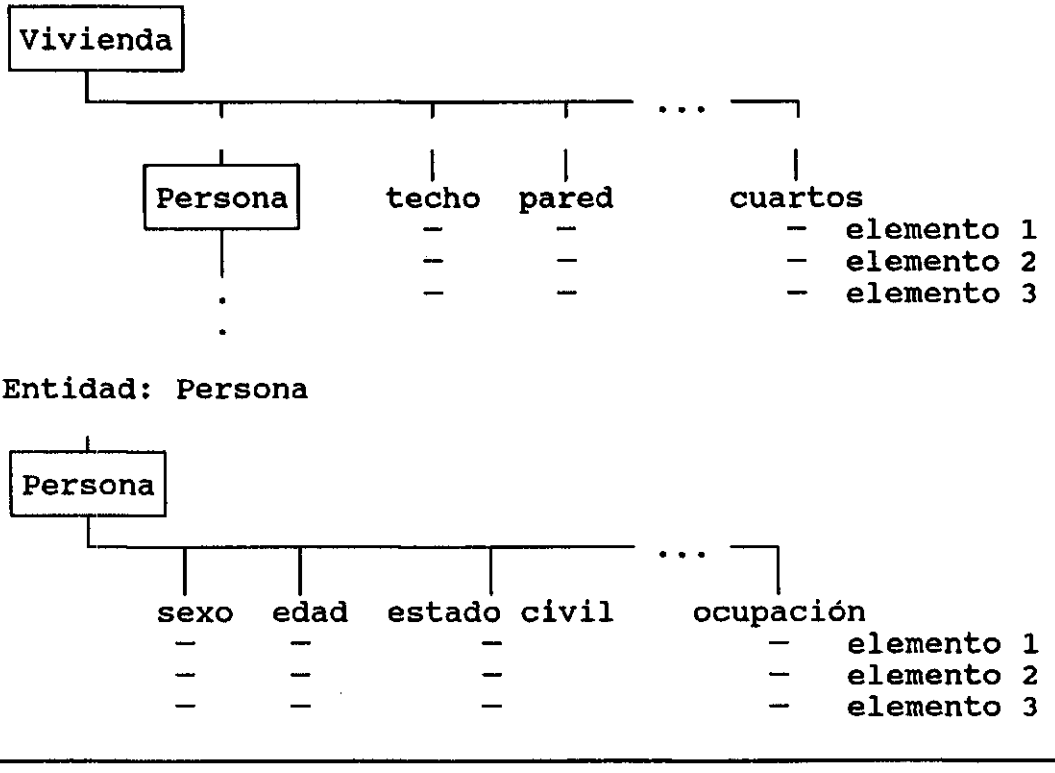
14. Por otro lado, una base de datos puede tener información de varias fuentes distintas, como por ejemplo, censos demográficos, agropecuarios, económicos, etc., siempre y cuando se los organice en *entidades* jerárquicamente relacionadas. Eso le da el criterio de *multidisciplinaria*.

15. Por último, la base de datos REDATAM puede ser considerada *relacional* porque la información de las variables es almacenada de una manera transpuesta, es decir, cada variable es un vector de información y la ubicación de un valor para una observación dada es por su posición relativa en el vector. Por ejemplo, la persona número 100 en la base tiene toda su información almacenada en la posición 100 de los vectores de variables que pertenecen a esta entidad, como sexo (100), edad (100), etc.

16. En ese sentido, cada entidad es una relación y cada variable de esta entidad es una columna de la relación. Por no ser un modelo estrictamente relacional, no se

pueden aplicar todos los operadores relacionales en su forma estricta generando nuevas relaciones.

Figura 2: La entidad VIVIENDA y sus variables.



17. En la estructura jerárquica, las entidades subordinadas también son funcionalmente variables de la entidad superior. En el caso de la Figura 1, la entidad "vivienda" tiene variables propias (techo, pared, etc.) y tiene una entidad subordinada ("persona"), que es representada también como una columna en la relación "vivienda". A su vez, la entidad "persona" tiene también sus propias variables (sexo, edad, etc.), pero no tiene ninguna entidad subordinada (sería una entidad terminal).

18. Este tipo de estructura es similar al modelo RAPID de Statistics Canada, con la diferencia que en el RAPID, las variables son guardadas en un archivo único por entidad, y en el REDATAM, cada variable es almacenada en un archivo separado (en sus versiones iniciales, el RAPID también separaba las variables por archivos).

19. Al tener las variables separadas por archivos, se obtienen las siguientes ventajas:

A. Se puede añadir ó eliminar variables de una entidad fácilmente, sin necesidad de tocar las otras variables de la entidad;

B. Se puede aplicar un proceso de compresión sencillo y eficaz a cada archivo (variable), porque todos sus elementos tienen el mismo tipo de información (en REDATAM se usa la compresión llamada "binaria", que se basa en el mayor valor que la variable puede tener para calcular el número de "bits" necesarios para almacenarla). Por ejemplo, la variable sexo ocuparía 1 byte (8 bits) en el almacenamiento tradicional, pero como sus valores varían entre 1 (hombre) y 2 (mujer), se necesitaría sólo 1 bit para almacenarla en forma binaria (la base REDATAM ocupa alrededor de un 20 por ciento del espacio original del archivo de entrada);

C. Al procesar la base de datos, solamente se leerán los archivos de las variables involucradas, es decir, para satisfacer una tabulación de parentesco por sexo para las personas de más de 15 años, el REDATAM procesará solamente tres archivos, los de las variables parentesco, sexo y edad.

20. La estructura jerárquica de entidades tiene las siguientes ventajas:

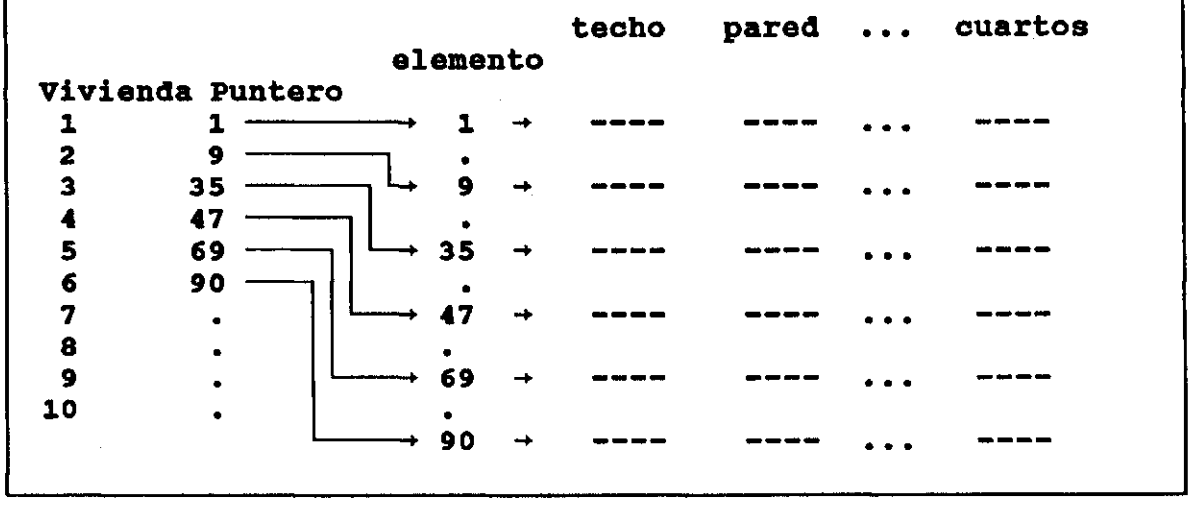
A. Permite la selección de algunas entidades, y/o algunos elementos (observaciones) de estas entidades. De esta manera, al procesarse una selección de entidades y/o elementos, el REDATAM solamente leerá las partes de los vectores de datos que correspondan (pertenezcan) a la selección;

La Figura 2 muestra que el distrito 2 "apunta" para la vivienda 9, y que el distrito 3 "apunta" para la vivienda 35. Eso significa que el distrito 2 tiene 26 viviendas, y al seleccionarse el distrito 2, solamente las viviendas de 9 a 34 serán leídas.

B. Permite el ahorro de espacio al no repetir información común a los elementos de una entidad, almacenándolos en el nivel superior correspondiente. Por ejemplo, el código de identificación geográfica de las viviendas y personas es almacenado separándolo por partes, y cada parte es guardada en su nivel jerárquico correspondiente, una sola vez para todos los elementos comunes.

C. Por supuesto que esta facilidad en el ahorro de espacio tiene que ser acompañada con la capacidad de combinación de variables de entidades distintas como, por ejemplo, saber cuántos niños en edad escolar viven bajo ciertas condiciones de vivienda. Como restricción, no se pueden combinar variables que pertenecen a entidades localizadas en ramas distintas, sin una subordinación común entre las mismas (véase Gráfico 1, ramas House90 y Agric90).

Figura 3: Ejemplo de una matriz con punteros para leer solamente una parte de los datos.



D. Permite la creación de variables agregadas a niveles superiores como, por ejemplo, el cálculo del ingreso familiar, ó el número de viviendas sin las condiciones mínimas de calidad por área geográfica (como en un estudio del mapa de pobreza de una ciudad).

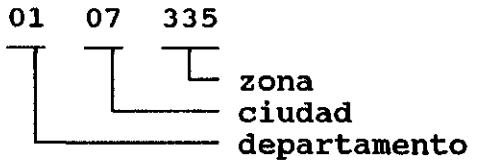
E. Como resultado de lo anterior, permite la combinación de variables agregadas de varias entidades distintas, en la entidad superior común a ellas. Como ejemplo, se puede calcular para cada ciudad (entidad), la relación del número de personal ocupado en la agricultura (censo demográfico, personas con la variable ocupación en cierto rango), con el número de hectáreas cultivadas (censo agrícola, suma de las áreas de cada finca). Favor referirse al Gráfico 1.

III. IMPLEMENTACION

21. Una base de datos REDATAM tiene tres tipos de archivos: a) diccionario; b) índices; y c) datos. El diccionario almacena toda la información de control de la base de datos, la estructura del árbol jerárquico, el ordenamiento entre las entidades, las características de cada variable, sus códigos y categorías, etc. En verdad, este tipo de información puede ser considerada la parte de los metadatos del sistema.

22. Existe un archivo índice para cada entidad subordinada. Este archivo es del tipo vectorial (igual que una variable común y corriente), y cada elemento de este vector contiene la dirección de la primera observación de la entidad subordinada. Por ejemplo, en el caso de las entidades de vivienda y población, la entidad vivienda tiene

Figura 4: Estructura de códigos geográficos.



un archivo índice cuyos elementos apuntan para la primera persona de cada vivienda.

23. Como ya se ha dicho anteriormente, cada variable de la base está almacenada en un archivo separado, del tipo vectorial, donde cada elemento es el valor de la variable para la observación relativa a la entidad. Por ejemplo, sexo (1000) contiene el valor de la variable sexo para el milésimo elemento de la entidad persona en la base de datos. Por supuesto, la utilización de estos vectores de variables es totalmente transparente al usuario, siendo manejado internamente por el sistema.

24. Los archivos índice son similares a los archivos de datos en cuanto al método de almacenamiento. La diferencia está en que en los archivos índice, el valor almacenado (dato) es un puntero (o dirección) para los archivos de las variables de la entidad subordinada.

25. La separación de la información en diccionario, índices y variables tiene las siguientes características convenientes:

A. La simplicidad de la estructura permite trabajar con los índices como si fueran variables de la entidad, es decir, una columna más en el modelo relacional.

B. Se pueden tener varios usuarios consultando una base central única y común, pero, al mismo tiempo, estos usuarios podrían tener variables auxiliares que fueran de conocimiento particular de cada uno. Bastaría para eso que una copia del diccionario estuviera almacenada en el directorio particular del usuario. Cualquier modificación ejecutada en este diccionario no afectaría el diccionario común y sería "visible" solamente a este usuario. En otras palabras, cada usuario puede crear sus variables adicionales, sin que eso afecte a los demás usuarios (siempre y cuando eso se haga en su directorio particular y que los archivos comunes de la base estén protegidos, por ejemplo, con los atributos de lectura del sistema operativo).

C. Por otro lado, también se puede tener una versión completa del diccionario de datos de modo centralizado, y "entregar" a cada usuario una copia modificada de este diccionario, solamente con las entidades y variables que este usuario

pueda "ver". Eso es nada más que la implementación del concepto de esquemas y sub-esquemas para archivos.

D. Aplicándose el mismo concepto, conjugado con la selección jerárquica, se puede generar fácilmente, con comandos del propio REDATAM, sub-bases de datos para la descentralización de la información. Estas sub-bases REDATAM pueden ser fracciones jerárquicas (todos los datos de la provincia de Córdoba, por ejemplo), fracciones lógicas (sólo algunas entidades y variables, por ejemplo, todas las variables de vivienda y algunas variables de población para una aplicación en la compañía de electricidad), ó una combinación de ambas.

IV. PROGRAMACION

26. El sistema REDATAM está totalmente programado en lenguaje C, compilador Microsoft versión 5.1 y usa también un conjunto de bibliotecas comerciales de función, de las cuales las más importantes son la "C-Worthy" para las funciones de despliegue y recuperación de datos en pantalla, y la "Greenleaf", para algunas funciones específicas.

27. Está dirigido a la familia de computadores IBM/PC ó compatibles, con el sistema operativo DOS 3.0 ó versiones posteriores. No requiere más que los 512 Kb de memoria comunes, tampoco exige un coprocesador matemático. El sistema está diseñado modularmente, de tal forma que los programas ocupen un mínimo necesario de espacio en memoria. Al mismo tiempo, para los usuarios que no tengan que crear y generar bases de datos, se puede tener una versión compacta del sistema, solamente con los módulos de selección jerárquica y procesador estadístico.

28. Toda la información de textos del sistema (menús, mensajes, ayudas, etc.) está almacenada fuera de los programas ejecutables, en archivos separados por tipo de texto, lo que hace más sencilla la tarea de mantención del sistema y, además, permite generar versiones del mismo en varios idiomas. Actualmente, existen las versiones de REDATAM 3.1 en español e inglés, la versión en francés va a ser producida (básicamente para algunos de los países africanos) y en portugués está en estudio (para Brasil y algunos países de Africa). Sin embargo, la sintaxis de los comandos del Procesador Estadístico es única e inmutable (CROSSTABS a BY b no puede ser cambiado para CRUCE a POR b).

V. DICCIONARIO DE DATOS

29. No hay limitaciones para el número de entidades y variables de un diccionario. En la práctica, por efectos de diseño de la pantalla, para desplegar el árbol jerárquico de entidades, el máximo de entidades en una misma rama es 18.

30. Además de las variables numéricas enteras, el sistema REDATAM acepta variables numéricas decimales, valores negativos y variables alfanuméricas. Las variables alfanuméricas no pueden ser comprimidas.

31. Además de las categorías propias, cada variable posee dos códigos adicionales, uno para los valores fuera de rango y otro para el no aplicable (blancos). De esta manera, al cargar una base de datos, todos los valores que no estén definidos en rango, son transformados para el código informado como el "fuera de rango" y todos los blancos son transformados para el código informado como "no aplicable". Por ejemplo, supongamos que la pregunta de alfabetismo sea contestada solamente para las personas de más de 5 años, que sus categorías sean 1 (sabe leer) y 2 (no sabe), y que el archivo no esté totalmente depurado, con la posibilidad de la existencia de otros valores distintos de 1 y 2. Se puede definir entonces el valor cero para los no aplicables, y el valor 9, por ejemplo, para los fuera de rango. Al cargar la base de datos, los blancos son transformados en cero y los valores distintos de 1 y 2 son transformados en 9.

32. Los valores de los códigos de "fuera de rango" y "no aplicable" no necesariamente tienen que ser distintos entre sí, y tampoco tienen que ser distintos de las categorías informadas para la variable. Es decir, queda a criterio del administrador definir estos valores para cada variable, que pueden ser iguales entre sí y/o iguales a una categoría existente.

33. Adicionalmente, estos valores no están limitados al número de caracteres de la variable original. Por ejemplo, supongamos que la variable "qué hizo en la semana anterior" tenga originalmente una posición, y que sus códigos varíen de cero a nueve. Se necesita un código adicional para los no aplicables (personas que no contestan la pregunta) y, en este caso, se puede informarlo como 10. Al cargar la base de datos todos los blancos serán transformados para el valor 10.

34. Para aumentar el factor de compresión, el REDATAM usa una tabla interna de conversión para almacenar las variables de 1 y 2 bytes. Esta tabla es totalmente transparente al usuario y funciona de la siguiente manera: los códigos de la variable son convertidos a valores secuenciales empezando de cero, almacenados como tal, y al ser leídos son convertidos nuevamente al valor original. La ventaja es que, en general, los censos usan siempre los valores 9 y 99 para ignorado, y a veces las variables tienen los valores 1, 2 ó 9, ó 1 al 12 y 99, por ejemplo. De ser almacenadas así, las variables gastarían más espacio (4 bits para el 9 y 7 bits para el 99), mientras que convirtiendo de 1, 2 y 9 para 0 a 2 se gasta solamente 2 bits. Para el segundo ejemplo, al convertir de 1 a 12 y 99 para 0 a 12, se usan solamente 4 bits.

35. Las bases de datos REDATAM ya existentes para la versión 3.1 pueden ser fácilmente traducidas para la versión REDATAM+, a través de dos módulos, que

convierten automáticamente el diccionario y los índices, respectivamente. Los archivos de datos propiamente tales (que es el más voluminoso de una base) no necesitan ser convertidos.

VI. SELECCION JERARQUICA

36. El proceso de Selección Jerárquica permite restringir el número de casos procesados específicamente al área de interés del investigador (entidad identificada, normalmente geográfica), seleccionando las subdivisiones más pequeñas (ó agrupaciones de éstas) del archivo de entrada, limitado apenas por el nivel mínimo de códigos en la identificación de los cuestionarios.

37. Esta agrupación de partes seleccionadas no está restringida a los límites político-administrativos de un país, es decir, se puede crear un área de interés compuesta de dos municipios limítrofes que pertenecen a departamentos distintos.

38. La Selección puede ser ejecutada convencionalmente, por elección directa de los ítems deseados, con el distrito 54, ó el municipio de Limeira, ó a través de un proceso de selección "cuantitativa", donde el usuario elige un criterio de comparación para las entidades, como, por ejemplo, todas las viviendas con una sola persona, ó todos los municipios con menos de un cierto número de viviendas. Por "cuantitativo" se quiere decir que se establecen criterios basados en la cantidad de casos de entidades inferiores. No existe un proceso de selección "cualitativa" basado en el contenido de las variables, como por ejemplo, seleccionar las viviendas que no tengan agua por cañería. Este tipo de selección sólo puede ser ejecutado en el Procesador Estadístico.

39. Existe una protección a la confidencialidad de los datos, a través de claves ("passwords"), que tienen dos conceptos: a) nivel jerárquico mínimo de selección (un determinado usuario sólo puede seleccionar hasta municipios, sin poder hacerlo para distritos y sectores, por ejemplo); y b) número mínimo de casos (selección con más de 100 viviendas ó 500 personas, por ejemplo).

VII. PROCESADOR ESTADISTICO

40. Existen básicamente tres procesos para la producción de resultados estadísticos: frecuencias, cruces y promedios. La frecuencia produce la distribución absoluta y relativa del número de casos en cada categoría de la variable. El cruce (de hasta cuatro variables) produce la frecuencia de cada celda resultante de la combinación de los valores de las categorías de las variables involucradas. El promedio es un cruce que además de los resultados absolutos produce también el promedio de los valores de una variable adicional.

41. Existe otro proceso, muy poderoso, llamado tablas, que funciona como un conjunto de cruces y que tiene la característica de poder ser usado para la impresión directa de resultados.

42. Todos estos procesos tienen una serie de opciones que pueden ser elegidas, a criterio del usuario, con el fin de adecuarlos específicamente a sus necesidades, como porcentajes por fila ó columna, clasificación en orden ascendente o descendente de valores ó códigos, etc. El proceso de tablas tiene un conjunto de opciones propias para ajustar su salida, tales como tamaño de columna, títulos, notas de pie de página, elección de marcos, etc.

43. Si el conjunto de estos tres procesos no es suficiente, el REDATAM tiene la facilidad de exportar archivos tipo ASCII, sólo con los casos y variables seleccionadas, para su posterior utilización en paquetes estadísticos más poderosos, como el SPSS/PC. Esta generación va acompañada de otro archivo con la descripción, en sintaxis SPSS, del archivo generado.

44. Existe también la facilidad de exportación de datos de entidades geográficas seleccionadas para su utilización en sistemas de información geográficas (GIS), tales como el pcARC/INFO y el OSUMAP.

45. Selección de casos: Además de la selección del área de interés, se pueden también seleccionar cualitativamente los casos a procesar, como por ejemplo, las mujeres de 15 años y más, ó las personas de menos de 12 años que no asisten a la escuela, ó las viviendas que tengan ciertas características de material de construcción. Existen también otros dos tipos de selección de casos, que son la selección por muestra (procesar una de cada 5 viviendas de cada municipio, por ejemplo), y la selección de los primeros x casos (procesar las 20 primeras viviendas de cada municipio).

46. Entre otras facilidades adicionales, se puede mencionar que los casos pueden ser ponderados, se pueden crear nuevas variables, definir y nombrar sus categorías y, quizás la más importante, la capacidad de un procesamiento jerárquico para generación de variables a niveles superiores, como por ejemplo la suma de los ingresos de las personas de una misma vivienda, ó el ingreso promedio de los jefes de hogar de un conjunto de municipios (que podría ser después exportado a un GIS para producción de mapas temáticos).

VIII. CREACION DE BASES DE DATOS

47. Se deben tomar algunos cuidados antes de proceder a la carga de una base. Estos son:

A. Definir las entidades jerárquicas de la futura base;

B. Asegurar que el archivo de entrada esté clasificado según las entidades jerárquicas;

C. Verificar las frecuencias de códigos para cada variable, con el fin de facilitar la creación del diccionario;

D. Crear el diccionario propiamente tal, con la definición de entidades y variables, sus códigos, categorías, posición (desplazamiento y largo) en el archivo original, etc.

48. Para la generación propiamente tal, hay dos caminos a seguir, dependiendo del tamaño del archivo original. Si hay espacio suficiente en el disco duro para almacenar todo el archivo de entrada y más un 20 por ciento para la futura base, el proceso se llama *carga automática*, y en este caso lo único que hay que hacer es elegir esta opción en los menús de generación y procesarla. No se necesita en absoluto ningún conocimiento especial de programación.

49. Sin embargo, si el archivo original no cabe totalmente en el disco duro, existen dos alternativas. La primera, más fácil, sería partir el archivo de entrada en n partes, y generar estos sub-archivos por separado, usando el método descrito anteriormente. Luego, usar una opción existente de juntar todas las n partes, formando una base única. Este método sería el más adecuado, siempre y cuando el número de partes no fuera muy grande.

50. La otra alternativa, usada para archivos muy grandes (por ejemplo, censos), es partir también el archivo original, pero al revés de hacerlo "horizontalmente", se hace lógicamente por variables, es decir, leer el archivo de entrada y generar tantos archivos de salida cuantas sean sus variables. Por ejemplo, si la entrada es un registro de 80 posiciones agrupados en 45 variables y existen 200 mil registros, la salida será compuesta de 45 archivos con largo de 1, 2, etc. posiciones, de acuerdo al largo de cada variable, cada archivo con los mismos 200 mil registros. En el caso de un censo de población, se grabarían archivos de variables de viviendas (a veces también de hogares) y de personas.

51. También es necesario ejecutar un programa para la creación de los archivos que van a generar los índices de las entidades. Esto se hace a través de un conteo del número de casos de cada entidad para cada instancia de la entidad superior. Por ejemplo, en el caso de una encuesta que tenga cuatro niveles jerárquicos (Departamento, Distrito, Vivienda y Persona), a cada corte del código de Vivienda se graba el número de personas de la vivienda anterior, a cada corte de Distrito se graba el número de viviendas del distrito anterior, y a cada corte de Departamento se graba el número de distritos del departamento anterior, en tres archivos separados. El archivo correspondiente a departamentos va a tener tantos registros cuantos sean los

departamentos, cada registro con el número de distritos existentes en este departamento, lo mismo para el archivo de distritos (cuenta viviendas) y para el de viviendas (cuenta personas). Es evidente y fundamental que el archivo original debe estar clasificado por los campos que son los códigos de las entidades.

52. Tanto los archivos de variables como los de índices son copiados para el disco duro del microcomputador (tantos archivos cuantos quepan cada vez), y se procede a otra modalidad de generación, llamada de *carga manual*, que exige que se informe al sistema cuales variables se está cargando cada vez.

53. Estos programas a ser ejecutados (en el computador central) sobre el archivo original, son programas "ad hoc", muy dependientes del formato y organización del archivo de entrada y, por lo tanto, lo más que se puede entregar como ejemplo a los usuarios es un conjunto de programas "tipo", en COBOL por ser de fácil entendimiento, que tendrían que ser adaptados y/o convertidos a otro lenguaje, por programadores que conocieran un poco más profundamente los conceptos de programación.

54. Además de la desventaja de la necesidad de programación en el equipo "mainframe", este método para bases de datos muy grandes también necesita más cuidado en la administración de las tareas y mantener un control rígido de todas las etapas (ejecución de los programas en el "mainframe", transmisión de los archivos, generación de cada uno, respaldo, etc.). La única ventaja es que es un proceso más rápido que la llamada carga automática, porque todas las tareas de identificación de las variables, conteo de los índices, cuidados en los cortes de cada código, etc., son ejecutados en un equipo muchísimo más veloz (el "mainframe").

55. El tamaño de una base de datos puede ser estimado inicialmente como un 20 a 25 por ciento del archivo original, es decir, si este archivo ocupa 20 Mb, la base ocupará alrededor de 4 Mb. Un cálculo más efectivo puede ser efectuado luego de la creación del diccionario, cuando entonces se sabe el número de bits ocupados por cada caso de la variable. Basta multiplicar la suma de estos bits por el número de casos de cada entidad, y agregar uno adicional para los índices (4 bytes para cada elemento de cada entidad), y el diccionario (1 o 2 Kb). Por ejemplo, supongamos las siguientes características para un archivo ficticio de entrada:

- registro de 80 posiciones, dos tipos de registro (uno para vivienda y otro para personas), ambos con la identificación geográfica;

- 6 niveles jerárquicos (departamento, municipio, distrito, sector, vivienda y persona);

- 3 departamentos, 10 municipios, 50 distritos, 600 sectores, 9.000 viviendas y 45.000 personas;

- 15 variables de vivienda que ocupan 22 bytes, y si están comprimidas por el REDATAM, según el diccionario, ocuparían, al sumarlas, 65 bits;

- 34 variables de población que ocupan 55 bytes, y si están comprimidas por el REDATAM, según el diccionario, ocuparían, al sumarlas, 143 bits.

Los cálculos serían los siguientes:

- Espacio para las variables de vivienda:

9.000 * 65 = 585.000 bits = 73.1 Kb

- Espacio para las variables de población:

45.000 * 143 = 6.435.000 bits = 804.4 Kb

- Espacio para los índices:

(3 + 10 + 50 + 600 + 9.000) * 4 = 38.7 Kb

El espacio total sería entonces:

73.1 + 804.4 + 38.7 = 916.2 Kb, ó aproximadamente, 920 Kb.

El archivo original ocupa:

(9.000 + 45.000) * 80 = 4.32 Mb

56. Cabe notar que es igualmente factible usar otros medios de almacenaje, como disco láser WORM (Write Once, Read Many times), CD-ROM (compact disc), etc., para bases de datos muy grandes.

IX. FILOSOFIA DEL MODELO REDATAM

57. El planteamiento de una "filosofía" REDATAM se basa en la utilización directa de los microdatos por los usuarios finales, sin la intervención de analistas y programadores, a través del almacenamiento de la información en microcomputadores.

58. También se considera, dentro de la misma línea de utilización directa, que estas bases pudieran estar disponibles para todos los usuarios de un sistema estadístico, *dentro y fuera de las oficinas de estadística*, resguardados los conceptos de la confidencialidad de la información.

59. A pesar de que el sistema REDATAM fue inicialmente diseñado para procesar pequeños conjuntos de grandes volúmenes de datos, su utilización no está limitada a éstos y las experiencias en varios países han mostrado sula eficiencia también con archivos pequeños, inclusive encuestas por muestra. Vale aquí resaltar que en este caso (muestras), el tamaño y diseño de la muestra a veces no permite la selección jerárquica a niveles geográficos menores.

60. El modelo no se limita a datos geográficos de censos de población y vivienda. Cualquier archivo que sea organizado jerárquicamente puede ser almacenado en una base REDATAM. A continuación, se mencionan algunos casos conocidos de implementación:

A. En Statistics Canada se hizo una prueba de crear una base de productos de exportación e importación, donde la jerarquía era una combinación del código del producto separado en grupos, sub-grupos e ítem, y el año y mes de la información;

B. En Costa Rica se creó una base con las estadísticas vitales, organizadas geográficamente;

C. En el CELADE, junto con la OPS (Organización Panamericana de Salud), se creó una base de datos para los registros de muerte de Brasil, organizada de dos maneras distintas: una con las causas de muerte separadas en grupos y sub-grupos, y otra por grupos de edad y edades simples de los fallecidos.

61. Si un archivo cualquiera va a ser procesado solamente un par de veces, quizás no se justifique la creación de una base REDATAM. Sin embargo, hay que mencionar también la facilidad, ó estandarización de una herramienta a ser usada por todo el personal de una oficina de estadística y que puede constituirse en un "lenguaje de trabajo" para sus técnicos.

X. CONCLUSION

62. Las ventajas de la utilización del REDATAM pueden ser resumidas en las siguientes:

A. El volumen de datos a procesar es mucho menor, porque, por un lado, se selecciona el área geográfica específica y, por otro, solamente se leen las variables involucradas en el proceso;
B. Los datos almacenados en forma comprimida ocupan un 20 por ciento del espacio original. Por eso, todos los datos de una ciudad, región ó país, pueden estar siempre disponibles;

C. La existencia de un diccionario de variables trae como consecuencia todas las ventajas de una base de metadatos (uniformidad de documentación, disminución en la programación, etc.);

D. El caracter multidisciplinario del modelo permite combinar datos de distintas fuentes, organizando un verdadero sistema de información para la planificación y como generadora de atributos para un GIS;

E. Lo amigable del sistema, dirigido por menús, unido a un lenguaje de programación muy similar a otros paquetes estadísticos, permite al usuario final (investigador, planificador, demógrafo, etc.) trabajar directamente con la base de datos, sin necesidad de la intervención de analistas y programadores, a no ser en el momento inicial para la creación de la misma;

F. La conexión con sistemas GIS permite la producción de mapas temáticos y el estudio espacial de la información;

G. Aun tratándose de archivos no muy voluminosos, cuando quizás no se justificara el esfuerzo de la creación de una base de datos, el hecho de que se use la misma metodología y filosofía de trabajo, lenguaje común, facilidad de acceso, etc., parece indicar que el REDATAM pudiera ser una herramienta en cierto modo "estándar" en las oficinas que trabajan con datos estadísticos en microcomputadores.

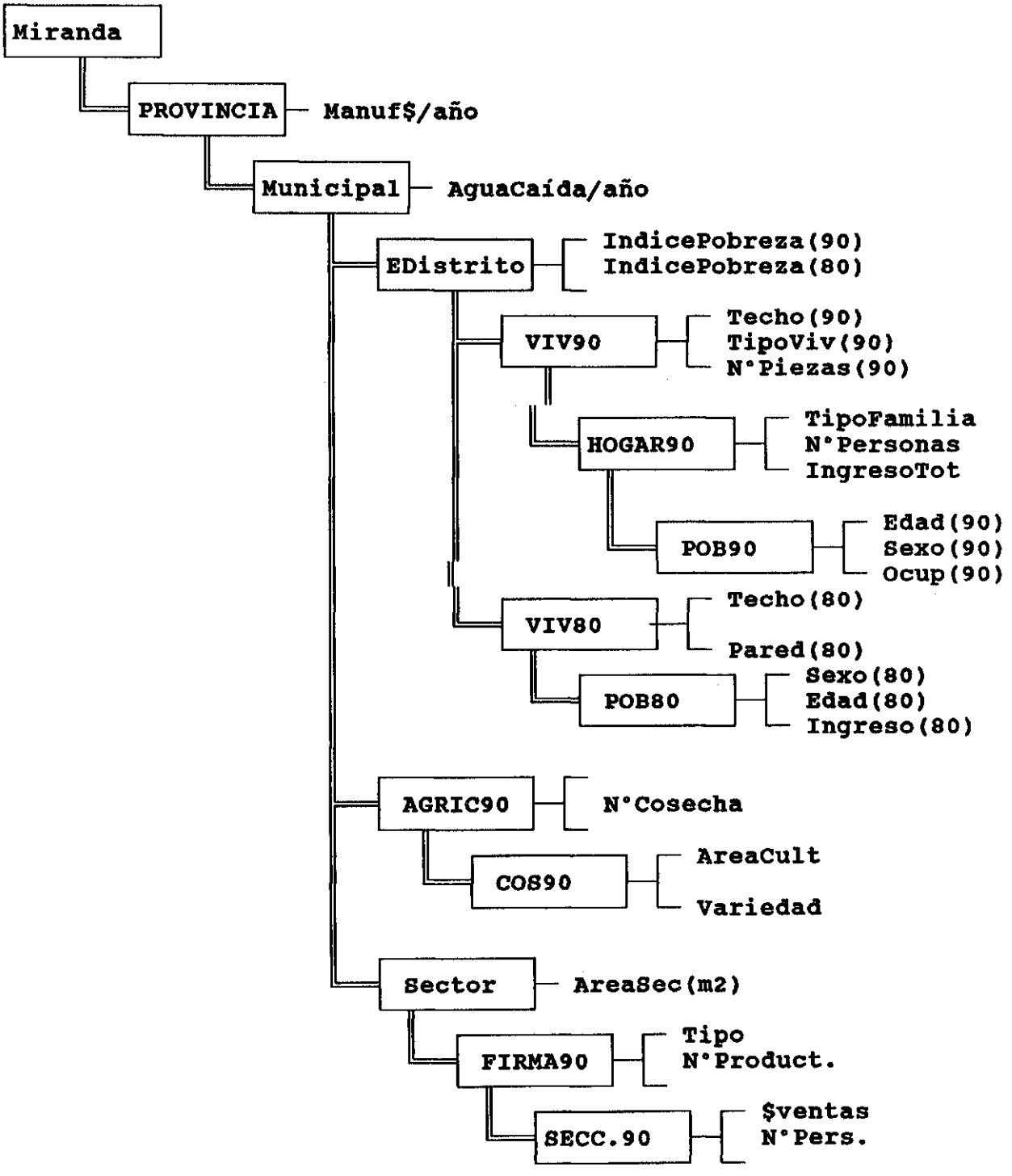
63. Las principales desventajas de un sistema REDATAM serían:

A. La creación de una base de datos para grandes volúmenes de información no es una tarea sencilla;

B. La limitación de sus procesos estadísticos (frecuencias, cruces, promedios y tablas) puede hacer que el usuario más sofisticado se vea obligado a trabajar con otros paquetes (usando facilidades de "exportar" de REDATAM).

64. Como conclusión final, el modelo REDATAM, al integrar las técnicas de almacenamiento jerárquico y relacional, permite entre otras, la facilidad de manipulación de informaciones multidisciplinarias, lo que aumenta en mucho la aplicabilidad del sistema, más allá de un archivo sólo de vivienda y población. Resta al usuario imaginar el potencial que puede ser alcanzado y las posibles implementaciones al integrar distintas fuentes de información en una misma base de datos.

Gráfico 1: Estructura de una base de datos REDATAM-Plus



R2DICese.pub CELamc60 (07/11/1989 <17:52>|) :\red\plus\dicciona.esq