PRODUCE: LA TECNOLOGIA DE BASE DE DATOS PARA EL ACCESO EFICIENTE A DATOS DE POBLACION

Paul Cotton (*)

PRODUCE – USING DATABASE TECHNOLOGY TO PROVIDE EFFICIENT ACCESS TO POPULATION DATA

SUMMARY

This paper introduces CELADE's PRODUCE system which will use a database management system (DBMS) to provide efficient access to population census and survey micro-data. The RAPID DBMS developed by Statistics Canada is shown to be more appropriate for population data processing than traditional commercial DBMS products. Special emphasis is given to the usefulness of the transposed physical structure for the creation of population or statistical databases. An outline is given of how the RAPID DBMS will be integrated with existing statistical packages to make micro-data more cost-effectively available to the regional researchers and to investigators within CELADE.

(*) Consultor de *Statistics Canada* en CELADE, bajo el patrocinio de CIDA/ CANADA.

1. *Introducción*

El Programa Latinoamericano de Información sobre Población (INFOPAL) del CELADE ha adoptado un enfoque integrado de almacenamiento y recuperación de documentación y procesamiento de datos sobre población para atender a los requerimientos de los investigadores de la América Latina. En este artículo se describe en qué forma se están empleando técnicas modernas de manejo de datos, mediante el uso de computador, usando un sistema denominado PRODUCE, con el fin de mejorar el almacenamiento y la recuperación de datos de censos, encuestas y otras fuentes utilizadas en los estudios de población. Para esa tarea se ha elegido el Sistema de Manejo de Base de Datos (DBMS) RAPID (1), desarrollado por *Statistics Canada* para satisfacer sus propias necesidades de procesamiento de datos estadísticos.

En la sección 2 de este documento se comenta el proceso de cambio por que pasa la producción de datos de censos y encuestas, incluyendo el procesamiento electrónico, en los países de la América Latina, puesto que esa circunstancia ha influido en gran medida para la elección del RAPID por parte de CELADE sobre otros sistemas DBMS comerciales disponibles. La descripción de este sistema de procesamiento de datos se explica con mayores detalles en la sección 3, en la que se comenta la necesidad de contar con un diccionario integrado de datos en las bases de datos estadísticos.

El capítulo cuarto trata del concepto de archivos transpuestos y la forma en que esa estructura física de almacenamiento puede atender las necesidades que se indican en las secciones 2 y 3. En la sección 5 se describe cómo el RAPID se conecta con los paquetes de programas estadísticos existentes como base del sistema PRODUCE.

2. *El proceso de cambio de las necesidades de procesamiento de datos de población en América Latina.*

Cuando en los países latinoamericanos se trabaja con un archivo nuevo, como un censo, es usual que se empiece por diseñar un amplio conjunto de cuadros, con la expectativa de satisfacer las necesidades de los diferentes usuarios, de modo que prácticamente no sea necesario preparar tabulaciones adicionales. Durante la década de 1970, los programadores de la región pudieron trabajar con computadores relativamente pequeños para producir esas tabulaciones de grandes archivos censales mediante el uso de paquetes de programas como el CENTS (4) y el COCENTS (5) que, con elevado nivel de eficiencia, producen

cuadros listos para fotografiar y publicar en la forma en que salen del computador.

Los usuarios de datos de encuestas, las que tienen menor número de casos, también tienden a producir grandes cantidades de cuadros utilizando paquetes de programas menos eficaces, pero con mayor versatilidad para fines estadísticos y de uso más fácil, como el SPSS (3) (Statistical Package Programs for the Social Sciences).

La tendencia a elaborar todos los cuadros que se consideran necesarios en cada pasada de los datos por el computador es bastante racional, desde el punto de vista del costo y del acceso al computador, puesto que cuando se utilizan programas corrientes éstos deben leer todos los registros para realizar cualquier tabulación independiente.

Ciertamente que es útil contar con una serie de tabulaciones de interés general sobre un archivo de datos. No obstante, la creciente importancia que se atribuye en la región a las políticas de población y a la incorporación de variables demográficas al proceso de planificación del desarrollo, hacen difícil establecer a priori, con relativa precisión, qué tabulaciones serán necesarias para un determinado estudio. Hay dos situaciones, muy corrientes, en las que vale esa afirmación. Una situación es la de los planificadores que necesitan información detallada sobre población para proyectos concretos de desarrollo que tienen lugar en ciertas regiones que a menudo no se ajustan a las fronteras político-administrativas normales. Lo mismo ocurre con las categorías sociales de la población usadas habitualmente. Por ejemplo, puede ser preciso contar con información sobre familias que viven en extrema pobreza en una región donde se va a ejecutar un proyecto de riego. En tal caso habría que producir tabulaciones especiales a partir de los datos censales, experimentar quizás con índices complejos para definir la extrema pobreza y estructurar la región a base de los segmentos censales de empadronamiento.

La importancia de los estudios en profundidad destinados a mejorar la comprensión de los fenómenos, con fines de política, ha creado una segunda situación en la que hace falta un enfoque interactivo de la tabulación de los datos. Un demógrafo que tenga acceso a los datos nacionales de la Encuesta Mundial de Fecundidad puede estar interesado en hacer un estudio en profundidad de las mujeres que conocen los anticonceptivos y que dicen que no desean tener más hijos, pero no utilizan un método para controlar la fecundidad. La mayoría de las encuestas

nacionales realizadas en el marco de la Encuesta Mundial de Fecundidad (WFS) abarcan unos 5.000 casos e investigan cerca de 350 variables. Después de definir las hipótesis de trabajo y examinar el conjunto de tabulaciones de interés general ya publicadas para tener una idea de los datos, el investigador pediría una serie de tabulaciones específicas, sucesivamente, a medida que fuese avanzando en su estudio. En tal caso, habría que hacer de cada vez sólo unas pocas tabulaciones y volver al computador cuando el mismo análisis sugiriese la necesidad de nuevos datos.

En ambas situaciones se dan ciertas características comunes. Los planificadores e investigadores *a)* necesitan tabulaciones especiales para realizar análisis detallados que son imposibles de prever antes de iniciar el estudio particular; *b)* deben volver repetidamente a la información básica, usando cada vez sólo una parte de las variables que contiene el archivo de datos; *c)* ocupan todos los casos o bien un subconjunto definido de ellos; y *d)* necesitan los resultados rápidamente y a un bajo costo, para que el estudio sea viable. Tales usuarios, que en la América Latina cuentan con fondos limitados para procesamiento y no siempre tienen fácil acceso al computador, necesitan de un sistema de procesamiento de datos que tome en cuenta estas condiciones y actúe en forma más adecuada que los procedimientos utilizados tradicionalmente en la región. Una solución a este problema es el empleo de un sistema de manejo de base de datos (DBMS). En las secciones siguientes se presentarán algunas ideas acerca del procesamiento de datos estadísticos mediante el uso de una base de datos.

3. *Requerimientos de la base de datos estadísticos.*

Cada DBMS ofrece al usuario una estructura lógica mediante la cual se puede imaginar la forma en que se presentarán los datos. Esta imagen puede ser muy distinta a la estructura física en que se almacenan los datos. El sistema RAPID emplea una estructura lógica que permite al usuario imaginar sus datos como un conjunto de cuadros sencillos (*). Este modelo de cuadro es especialmente aplicable en instituciones que, como el CELADE, cuentan con usuarios que son muchas veces estadísticos o demógrafos acostumbrados a ver los datos que analizan en forma tabular o de matriz.

(*) Aquí no se explicará el modelo tabular. El lector interesado puede consultar la referencia (2), donde encontrará una buena introducción a éste y a otros modelos de base de datos.

FIGURA 1

VARIABLES (ATRIBUTOS)

	Líneas	Columnas				
		1	2	. . .	. . .	N
REGISTROS	Línea 1					
	Línea 2					
	Línea 3					
	. . .					
	. . .					
	. . .					
	Línea M					

Cada cuadro en la base de datos se presenta como en la figura 1. Un cuadro describe una *entidad* de interés para el usuario y se compone de líneas y columnas. Cada línea representa un caso u ocurrencia de la entidad descrita en el cuadro y también es llamada *registro.* Las columnas representan diferentes variables (edad, estado civil, etc.) que describen algunos atributos (soltero, casado, etc.) de la entidad. Como ejemplo se muestra en la figura 2 cómo un usuario puede ver los datos de una encuesta con las entidades VIVIENDAS y PERSONAS.

La definición de algunos conceptos que se utilizan en el DBMS, como *clave, variable prima* y *dominio,* son fundamentales en el modelo tabular.

Se entiende por clave una variable o conjunto de variables que identifican únicamente una línea del cuadro. Por ejemplo, en una encuesta la clave puede ser definida como el número del cuestionario. Las variables que identifican una línea que permite la construcción de la clave se llaman variables primas. En cambio, las variables que le son

FIGURA 2

CUADRO DE VIVIENDAS

Casa	Casa No.	Número de ocupantes	Número de cuartos	Tipo de material empleado
Casa 1	1	3	3	Cemento
Casa 2	2	2	2	Ladrillo
Casa 3	3	1	7	Cemento
. . .				
. . .				

CUADRO DE PERSONAS

Persona	Casa No.	Número de personas	Edad	Sexo	Estado civil
Persona 1	1	1	27	M	Casado
Persona 2	1	2	26	F	Casado
Persona 3	1	3	7	M	Soltero
. . .					
. . .					

propias, como por ejemplo la edad, el sexo o el estado civil de las personas se llaman variables no primas.

La única variable prima en la figura 2, que corresponde a un cua-

dro de *viviendas,* es el *número de la casa,* desde que él sólo identificaría la línea (clave) que permitiría ubicar la vivienda en la base de datos. En el cuadro de *personas* la clave sería dada por el número de la vivienda y el número de la persona, dado que los dos juntos son necesarios para identificar una persona determinada, en el cuadro de personas.

Un dominio es el conjunto de todos los posibles valores que puede asumir una variable, aun cuando en la base de datos se represente sólo un subconjunto de ellos en un momento dado. Por ejemplo, el *material usado* en una vivienda puede tener como dominio: ladrillo, cemento, barro, acero, paja, madera. Es posible que en una encuesta particular de vivienda no se encuentren todas las seis posibilidades. Aun cuando la no inclusión de todas las características de un dominio no afecta materialmente la estructura lógica, desde el punto de vista del usuario de la base de datos, tiene importancia tanto para armar la estructura física de almacenamiento que emplea el DBMS como para el desempeño eficiente de la base de datos.

Para examinar los tipos de datos estadísticos se considerará un dominio caracterizado por el número de valores discretos que contiene, a los que se llamará su *rango,* y en seguida la capacidad que tiene un valor del dominio para identificar una entidad específica en la base de datos.

En el procesamiento de encuestas se pueden distinguir tipos de datos cuyos dominios difieren en cuanto a las características generales mencionadas. Los dominios cuantitativos tienen un amplio rango, pero escasa capacidad para identificar una unidad. Están en ese caso el ingreso y la fecha de nacimiento, por ejemplo. Los dominios que tienen un pequeño rango y poca capacidad para identificar son por lo general dominios cualitativos, que corresponden a atributos codificados, como son el sexo, el estado civil, etc.

Los atributos cualitativos y cuantitativos no constituyen típicamente dominios de atributos primos de un cuadro, debido a su baja capacidad como identificadores. En la práctica, los dominios de atributos primos consisten en identificadores creados para ese fin, los que deben caracterizarse por un rango mediano a grande y una elevada capacidad para identificar. Por ejemplo, la clave del cuadro de personas, en la figura 2, se compone de dos atributos creados: número de la casa y número de la persona. Cabría señalar que los atributos, como por ejemplo la dirección, fueron descartados en favor de un número arbitrario establecido con el fin de asegurar que sea el único.

Más adelante se analizarán los métodos particulares que emplea RAPID para almacenar físicamente estas distintas clases de dominios; primero se estudiarán las funciones usuales de procesamiento estadísticos en su relación con estos tipos de datos.

Requerimientos típicos de procesamiento de datos estadísticos

Una ventaja de los sistemas de base de datos está en la facilidad de acceso directo a los registros lógicos particulares (eso es, un caso o un cuestionario cuando se trata de datos de población) con fines de recuperación o actualización. Por ejemplo, para contestar a la consulta: ¿cuál es la edad de la persona número 613? se pediría al DBMS que recupere la variable edad del registro de persona cuya clave es 613.

Se puede clasificar las consultas a la base de datos según el número de registros comprendidos en la respuesta. Normalmente las preguntas como en el ejemplo anterior requieren un número muy pequeño de registros lógicos o un registro individual y son consideradas como *informativas.* Aquellas consultas que se refieren a un pequeño sub-conjunto de registros (digamos menos de un 10 por ciento) son consideradas como *consultas operativas.* Por ejemplo, si se desea conocer la relación entre la condición de migrante y el nivel de educación de todas las mujeres incluidas en una encuesta, hay que leer todos los registros que contiene la base de datos para la entidad mujer.

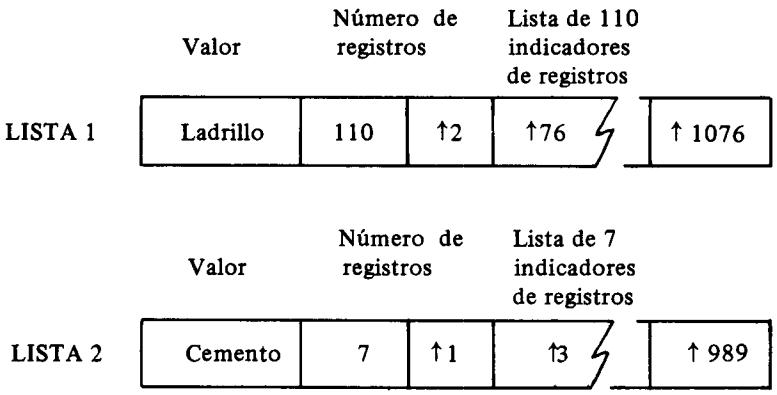
Los dos primeros tipos de consultas (informativa y operativa) se hacen normalmente a partir de las *variables primas.* Existen actualmente muchos sistemas comerciales de base de datos que emplean diversas técnicas, como por ejemplo archivos de acceso directo, listas invertidas, etc. para poder contestar en forma eficaz a las consultas informativas y operativas. La importancia que las bases de datos comerciales atribuyen a las claves para atender a las consultas basadas habitualmente en sistemas administrativos generalmente no son apropiadas para consultas estadísticas. Estas son normalmente definidas en términos de *variables no primas* y muchas veces en términos de dominios cualitativos. Por ejemplo, podría haber consultas estadísticas acerca de "madres que trabajan" "hombres mayores de 15 años" y "casas de ladrillo". Los conjuntos resultantes en las consultas estadísticas pueden abarcar todos los registros de un cuadro, como sería el caso en la creación de frecuencias multivariadas o estudios de correlación. Dichos procedimientos tienen que ver generalmente con un pequeño número de variables (menos de

FIGURA 3

INDICE INVERTIDO
CUADRO DE VIVIENDAS

Casa	Casa No.	Número de ocupantes	Número de cuartos	Tipo de material empleado
Casa 1	1	3	3	Cemento
Casa 2	2	2	2	Ladrillo
Casa 3	3	1	7	Cemento
. . .				

TIPO DE CONSTRUCCION
INDICE INVERTIDO



↑ 1 es un indicador de la vivienda No. 1

10, por ejemplo) sobre una gran cantidad de registros (las muestras censales del Banco de Datos de CELADE alcanzan cifras que van de 50.000 a más de 1 millón de casos).

Cuando los sistemas comerciales se ven enfrentados a consultas de tipo estadístico, que requieren una selección de registros a base de variables *no primas,* la técnica más común es la de crear un *índice invertido* sobre cada atributo. El índice invertido se compone de listas de identificadores de registros, una para cada uno de los valores del dominio correspondiente. Por ejemplo, en el caso de la figura 2, que corresponde a un censo de vivienda, un DBMS comercial podría atender a las consultas sobre el material empleado y producir un índice invertido sobre esa variable. Para cada valor en el dominio, el índice contendría una lista de identificadores de registros que abarcaría todos los registros con aquel valor; por ejemplo, una lista de identificadores para todas las casas construidas de ladrillo, otra para todas las casas construidas de cemento, etc., como se muestra en la figura 3.

Se pueden atender a las consultas sobre casas de ladrillo buscando sólo en la lista de identificadores asociados con el valor ladrillo, en vez de leer el cuadro completo y verificar cada registro para determinar si describe *casa de ladrillo.* Por supuesto que las listas invertidas se deben almacenar en la base de datos en adición a los datos originales a partir de los cuales también se puede atender las consultas, sólo que a un costo más alto de almacenamiento y mantención. Lamentablemente, el empleo de los índices invertidos presupone que el responsable de la base de datos puede elegir algunos subconjuntos de variables no primas para preparar el índice, dado que pocas aplicaciones justificarían el costo de un índice sobre todas las variables.

De igual modo, en sistemas estadísticos, normalmente es muy difícil definir *a priori* qué variables serán objeto de consultas *ad hoc* por parte de un usuario. Además, en ciertos archivos de encuestas, como por ejemplo en la Encuesta Mundial de Fecundidad, que contiene 350 variables para cada registro, sería muy difícil seleccionar las variables que se deberían invertir.

El costo y la dificultad para emplear índices invertidos en aplicaciones estadísticas es otro factor que impide el uso de programas comerciales de base de datos en estos casos. En resumen, las consultas estadísticas que se refieren a unas pocas variables elegidas *ad hoc,* para un gran número de registros, no pueden ser atendidas eficientemente con los DBMS disponibles en el comercio, ya que en ellos se enfatiza la recuperación por registro único, tan común en aplicaciones no estadísticas. Incluso aquéllos que utilizan índices invertidos para variables no primas son inadecuados para el manejo eficiente de base de datos estadísticos.

Por todo ello el CELADE eligió el DBMS RAPID, que emplea la organización transpuesta que se describe en la sección 4 y resulta más apropiada para solucionar los problemas específicos de bases de datos estadísticos.

Requerimientos para la descripción de datos estadísticos

Si bien la descripción de los datos integrados en una base de datos es siempre importante, lo es aún más cuando se trata de variables cualitativas, tan comunes en las bases de datos estadísticos. Por ejemplo, la variable estado civil podría estar clasificada según las categorías *divorciado, casado, separado, soltero* y *viudo.* Muy pocos sistemas almacenan las categorías mencionadas en la forma escrita en la lista anterior, para economizar almacenamiento y aumentar la eficiencia de procesos. Los códigos efectivamente almacenados en la base de datos, para cada una de las categorías, son completamente arbitrarios, desde el punto de vista humano, y esto ha sido una fuente de errores de interpretación en sistemas estadísticos para computadores.

Lo que RAPID hace es separar lo que el usuario ve en los valores codificados de una determinada variable (para el estado civil esta visión externa corresponde a *divorciado, casado,* etc.) de la visión interna del computador, esto es, los códigos físicos realmente almacenados en la base de datos. Al establecer una correspondencia entre los valores que se presentan al usuario y los códigos particulares almacenados, el sistema RAPID logra optimizar el almacenamiento de cada categoría codificada, a fin de ocupar la cantidad mínima de espacio de la unidad de memoria y aumentar la eficacia del proceso.

Si bien esta correspondencia es particularmente útil en el caso de las variables codificadas, RAPID tiene un enfoque similar para todas las variables, de modo que cada cuadro dentro de la base de datos contiene su propia definición o descripción. Esto es importante por dos razones: *a)* las bases de datos estadísticos generalmente se archivan para uso futuro y el hecho de contar con un diccionario integrado que describa el cuadro permite disponer de una información estructurada; *b)* dado que cada cuadro en una base de datos RAPID es autodefinido, se simplifica enormemente la preparación de programas generalizados que permitan hacer consultas a la base de datos acerca de las variables que están disponibles y del formato en que se presentan.

4. *Archivos transpuestos y bases de datos estadísticos*

Como se menciona en la sección 3, el usuario imagina sus datos en una base de datos RAPID como un conjunto de cuadros compuestos de registros y variables (ver figura 1). Esta visión lógica de los datos no obliga al DBMS a almacenar registro por registro, como se hace en muchos sistemas convencionales. Una ventaja de la tecnología de base de datos es que la visión externa del usuario no tiene que ser idéntica a la estructura física del almacenamiento usada en el DBMS. De hecho el RAPID almacena cada cuadro mediante una estructura física transpuesta, lo que implica que en vez de hacerlo registro por registro (esto es, reuniendo todas las variables de cada registro) el sistema almacena cada cuadro por variable. La figura 4 muestra cómo las variables de un cuadro pueden ser transpuestas a subarchivos que contienen cada uno una variable. Se explicará cómo la estructura usada por el RAPID soluciona los problemas de las bases de datos estadísticos que se mencionaron en las secciones 2 y 3.

FIGURA 4

ARCHIVO TRANSPUESTO

Vivienda No.	Persona No.	Edad	Sexo	Estado civil	Año en que se casó
1	1	27	M	Casado	72
1	2	26	F	Casado	72
1	3	7	M	Soltero	

Una base de datos estadísticos debe ser capaz de atender consultas sobre cualquiera de las variables almacenadas y contestarlas de manera eficiente aun cuando el número de registros sea elevado. El RAPID, como un sistema completamente transpuesto, atiende a estos requerimientos en dos formas. Primero, cuando una consulta exige recuperar muchos registros para una o más variables, el trabajo realizado para encontrar la información rquerida y transferirla para ser procesada por el computador es independiente de la cantidad y el tamaño de

las variables no seleccionadas. Considérese, por ejemplo, la recuperación de tres variables de una encuesta nacional de fecundidad, para todos los registros del cuadro. Si el archivo estuviese almacenado en forma convencional, ordenado registro por registro, la lectura se haría como está indicado en la figura 5. Aunque se necesitaran sólo 3 de las 350 variables del cuadro, un sistema orientado al registro obligaría a leer todas las variables para cada registro del cuadro, lo que correspondería a leer todos los registros de una cinta magnética. Esto se indica con las flechas que pasan horizontalmente (por registro) a través de los datos (figura 5).

FIGURA 5

ORGANIZACION REGISTRO POR REGISTRO

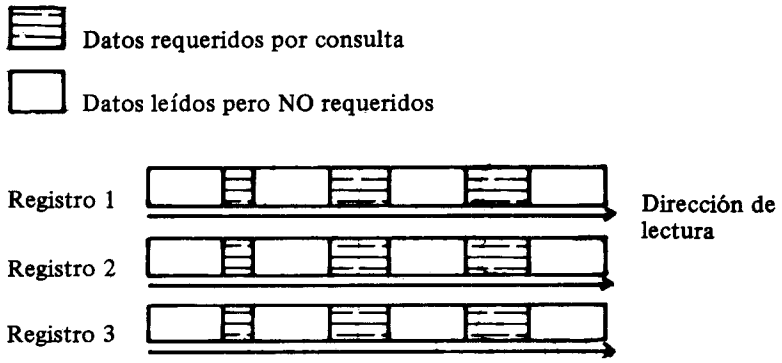
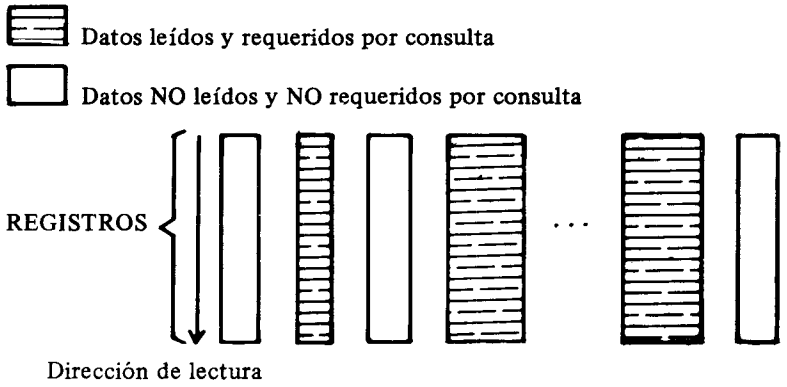


FIGURA 6

ORGANIZACION TRANSPUESTA



El sistema orientado al registro obliga a un trabajo extra, puesto que debe leer todas las variables del cuadro.

En la organización transpuesta, el resultado es muy distinto, como se ve en la figura 6. Como los datos están almacenados por variable, el sistema leerá el cuadro verticalmente y leerá sólo las variables que interesan. En los cuadros que tienen muchas variables, muchas líneas o ambas características, el ahorro puede ser muy significativo.

La segunda razón por la que el RAPID puede aumentar la eficacia de las consultas estadísticas se basa en el hecho de que el sistema almacena una sola variable en cada subarchivo, de tal manera que se minimizan los requerimientos de espacio de almacenamiento en disco magnético. Por ejemplo, el atributo *sexo* puede ser almacenado como un solo BIT (0 o 1), mientras los números decimales cero y uno necesitarían 8 BITS (un BYTE, en un computador IBM). Ello permite una comprensión que hace posible al RAPID almacenar un archivo en el 40 a 70 por ciento del espacio requerido por un sistema convencional orientado a los registros. Si bien la compresión en sí misma es importante, porque permite disminuir el costo de almacenamiento de la base física de datos, su efecto es aún mayor cuando se considera el trabajo de recuperar una variable. La compresión disminuye el número de bloques de datos del disco magnético que hay que leer para obtener los valores de una variable cuando se trata de un gran número de registros y esto a su vez disminuye el costo de la mayoría de las consultas estadísticas.

En RAPID también se transpone el contenido del diccionario de datos para cada cuadro. Cada variable tiene un "descriptor" básico y una descripción opcional que define la correspondencia entre los códigos internamente y los nombres de las categorías de las variables cualitativas. La transposición de la información descriptiva permite al sistema ubicar precisamente la información que necesita para atender un pedido del usuario. Así también se reduce el costo al usar una base de datos estadísticos que puede contener cientos de variables.

5. *El sistema PRODUCE*

El objetivo principal del sistema PRODUCE es proporcionar a los investigadores y planificadores un acceso más fácil y menos costoso a los datos de censos y encuestas de población.

El uso que hace el PRODUCE del DBMS RAPID será al comienzo exclusivamente en el campo de la recuperación. Las facilidades que

proporcionan el diccionario estándar de los datos y la creación del archivo de base de datos del DBMS RAPID serán empleadas para colocar los archivos más usados del Banco de Datos de CELADE en formato de base de datos. Entre ellos se encuentran los datos de las encuestas nacionales de fecundidad de varios países, realizadas en el marco de la Encuesta Mundial de Fecundidad (WFS), ya completamente revisadas y documentadas. Cada uno de estos archivos se compone de cerca de 350 atributos y 3.000 a 5.000 unidades de investigación. Estos datos se prestan especialmente para la organización de archivos transpuestos, dado que los futuros estudios en profundidad que se realizarán con la WFS se referirán en general a un pequeño número de variables con repetidas consultas a los datos por parte del investigador, a medida que avance su estudio.

Para que el sistema PRODUCE pudiese ser usado directamente por los investigadores se pensó en la necesidad de disponer de un conjunto completo de procedimientos estadísticos para manipulación de datos, a base de un lenguaje de alto nivel. Podría tomarse como ejemplo el SPSS, que es un lenguaje ampliamente utilizado (3). Para reducir el esfuerzo que se requeriría para crear y mantener la programación del PRODUCE, el SPSS se conectará con el RAPID, con el propósito de leer directamente la base de datos generada por RAPID. Así, los investigadores dispondrán inmediatamente de un lenguaje que muchos de ellos ya manejan y con la confiabilidad estadística de las funciones que se encuentran en la versión corriente del SPSS.

No obstante las facilidades del SPSS para producir tabulaciones cruzadas, ha quedado demostrado que otros sistemas orientados a tabulaciones censales, como por ejemplo el CENTS (4) y el COCENTS (5) pueden realizar ese tipo de tabulaciones en forma más rápida, cuando se trabaja con grandes archivos de entrada. Lamentablemente estos sistemas requieren el empleo de lenguajes que sólo pueden ser usados por programadores. Existe un sistema más reciente, el CENTS–AID (6), que incorpora el COCENTS y está destinado a solucionar este problema mediante el uso de un lenguaje muy parecido al que utiliza el SPSS. Con el fin de proporcionar a los usuarios del PRODUCE más facilidades de tabulación en forma eficiente, se elaborará la programación necesaria para permitir que el programa CENTS–AID lea la base de datos RAPID. Como el lenguaje del CENTS–AID se parece mucho al SPSS, les resultará relativamente fácil a los usuarios del PRODUCE utilizar uno de los sistemas que sea más apropiado para las investigaciones que realicen: SPSS, cuando se trate de archivos con un número relativamente pequeño de casos, pero requiera operaciones estadísticas complejas, o

CENTS–AID, cuando se trate de tabulaciones para grandes archivos.

Las conexiones antes mencionadas con el RAPID se utilizarán estrictamente en los casos de procesamiento en serie (BATCH). Además, al comienzo los sistemas sólo proporcionarán acceso a un archivo simple de RAPID. Una segunda versión de estas conexiones permitirá investigaciones más complejas, con el uso de más de un cuadro en una base de datos, como por ejemplo los cuadros de *viviendas* y de *personas* que se presentan en la figura 2. Esta segunda versión también permitiría al usuario hacer consultas más complejas, como por ejemplo calcular el ingreso total de los miembros de un hogar, obtener el ingreso per cápita y asignarlo a cada miembro del hogar. De este modo, informaciones que por lo general no están hoy al alcance del investigador estarán disponibles fácilmente y a un costo relativamente bajo.

Una vez que se haya ganado experiencia en este trabajo, se considerará el uso del RAPID para cumplir otras funciones en el procesamiento de encuestas. Esa tarea abarcaría una conexión con el CONCOR (7), que es un sistema que se aplica con mucho éxito en los procesos de limpieza de datos y que fue desarrollado y es mantenido por CELADE.

La instalación del PRODUCE se está llevando a cabo en estrecha cooperación con *Statistics Canada,* con el apoyo de la Agencia Canadiense de Desarrollo Internacional (CIDA). CELADE se beneficiará por ese medio de los avances introducidos en el RAPID por *Statistics Canada* y por otros centros que utilicen el sistema.

Por ejemplo, se espera que la Oficina de Estadísticas Laborales de los Estados Unidos elabore una conexión entre RAPID y sus sistema de tabulación TPL (Table Producing Language) (8). En el caso de que CELADE lo adquiriera, sus investigadores dispondrían de un paquete de tabulación de alto nivel, capaz de producir cuadros con la forma requerida por el usuario y listos para ser fotografiados e impresos.

A partir de comienzos de 1980, se espera que el CELADE cuente con esta nueva combinación de tecnología de base de datos y paquetes estadísticos, la que estará también accesible a las oficinas de estadística, de planificación, etc., de los países de América Latina que cuenten con adecuados recursos humanos y de computación.

El sistema PRODUCE funcionará sólo con computadores IBM S/370, compatibles, equivalentes o superiores al modelo 135, trabajan-

do bajo el sistema operativo OS y espacio en disco magnético suficiente para almacenar la base de datos del usuario.

En el futuro, CELADE proporcionará capacitación y asistencia técnica a los países de la región que lo deseen para implementar y utilizar el PRODUCE y ofrecerá facilidades de uso para aquellos países que no dispongan de recursos propios adecuados.

REFERENCIAS

(1) RAPID DATABASE MANAGEMENT SYSTEM. Se puede obtener información sobre este sistema en Special Resources Sub-Division, Systems Development Division, Statistics Canada, Ottawa, Ontario, Canadá, K1A OT6.

(2) Date, C.J. *An Introduction to Database Systems,* Edition 2, Addison Wesley Reading, Mass., 1977.

(3) Nie, N. *et al., SPSS – Statistical Package for the Social Sciences,* McGraw-Hill, 1975.

(4) *CENTS II, Census Tabulation System – User's Manual,* International Statistical Program Center. Bureau of the Census, US Department of Commerce, Washington D.C. July, 1975.

(5) *COCENTS – Systems Reference Manual,* International Statistical Program Centre, Bureau of the Census, US Department of Commerce, Washington D.C., May 1977.

(6) *CENTS–AID II, User's Manual,* DUALABS, 1601 North Kent Street, Arlington, Virginia, March, 1976.

(7) Ortúzar, Julio. Revisión automática de datos de censos y encuestas mediante el uso de computadores medianos y pequeños, *Notas de Población,* agosto de 1978, número 17.

(8) *Table Producing Language, Version 3.5, User's Guide,* Bureau of Labour Statistics, US Department of Labour, Washington D.C., July, 1975.