El manejo de datos en la investigación educacional

[**Revista del Centro de Estudios Educativos** *(México), vol. VII, núm. 3, 1977, pp. 102-113*]

Richard Wolfe
Ontario Institute for Studies in Education (Canadá)

En la investigación educacional ocurre con frecuencia que datos recabados de acuerdo con diseños elaborados e implementados cuidadosamente sean objeto de análisis inadecuados. Lo común es que el análisis original de los datos se realice de manera apresurada en los últimos días de la investigación, cuando hay que enviar el informe final a quien contrató o patrocinó la investigación. Casi siempre la intención de los investigadores es volver a los datos para reanalizarlos ampliamente. En general, se reconoce que el análisis es un proceso difícil, que existen muchas alternativas para trabajar con los mismos datos, y que éstos pueden ser utilizados con propósitos muy diferentes, posiblemente no completados en las fases de diseño y recolección. Sin embargo, sorprende que existan tan pocos ejemplos de reanálisis de archivos de datos educacionales realizados con éxito.

Muchos de los problemas en el procesamiento y análisis original y de los obstáculos en el reanálisis, obedecen a las dificultades que presenta el manejo de los datos de la investigación. Los investigadores educacionales enfrentan grandes dificultades para limpiar los datos (eliminar las inexactitudes e inconsistencias de la codificación) y realizar las transformaciones de los mismos que requiere el análisis estadístico. Más aún, la información que define y describe la ubicación, formato y codificación de los datos almacenados en cintas de computadora es a menudo tan incompleta que dificulta enormemente el análisis o reanálisis de los datos. Muchas veces los análisis originales fracasan y los reanálisis se abandonan por errores obvios en la fuente de los datos y serias ambigüedades en la documentación de éstos. El propósito del presente trabajo es proponer un conjunto de procedimientos prácticos para el manejo de los datos en la investigación educacional.[1]

---

[1] El sistema para el manejo de datos presentado en este trabajo está basado en la organización del banco de datos del Departamento de Investigaciones Educacionales (DIE) de la Dirección de Planeación, del Ministerio de Educación de Venezuela. Dicho banco de datos fue organizado en 1971 principalmente por Carlos Rodríguez, del DIE, y el autor, quien fue asesor de la Fundación Ford.

En ese mismo año contamos con el excelente asesoramiento de David Nasatir y Camilo Femenias, de la Universidad de California, Berkeley, quienes habían desarrollado el esquema de procedimientos utilizados en su centro e investigación, muchos de los cuales pudimos adaptar directamente a las necesidades del DIE. El banco de datos venezolano evolucionó y en un año se convirtió en una pequeña unidad burocrática dentro del DIE y continúa en funcionamiento. Rodríguez y Wolfe (1971) describen la organización y el conjunto de normas que integran ese banco.

Fundamentación

La gravedad del problema del procesamiento de datos en la investigación educacional obedece, entre otras razones, a la particular complejidad de los archivos de datos en este campo. En otras áreas de investigación se trabaja generalmente con una matriz simple de observaciones por variables; en educación, en cambio, se dan varios niveles de observaciones tales como escuelas, maestros, alumnos; para cada nivel muchas veces existen diversas fuentes de información, tales como cuestionarios diferentes, tests y archivos escolares. Los archivos de datos requeridos para el análisis constituyen versiones integradas de diferentes archivos de datos inicialmente preparados en forma separada. El sistema apropiado para el manejo de datos de investigación debe permitir el establecimiento de un gran número de complejas interrelaciones de los archivos de datos.

Una segunda causa de dificultades en el procesamiento de datos educacionales radica en la carencia de una organización permanente en los equipos de investigación; en particular, en la movilidad de los programadores y analistas de computación. Los investigadores principales en un proyecto cualquiera rara vez son expertos en computación y están alejados de la rutina diaria de esta actividad. Los procedimientos propuestos aquí para el manejo de los datos pueden ser entendidos entonces como instrucciones que los investigadores deben impartir a los equipos encargados del procesamiento de datos.

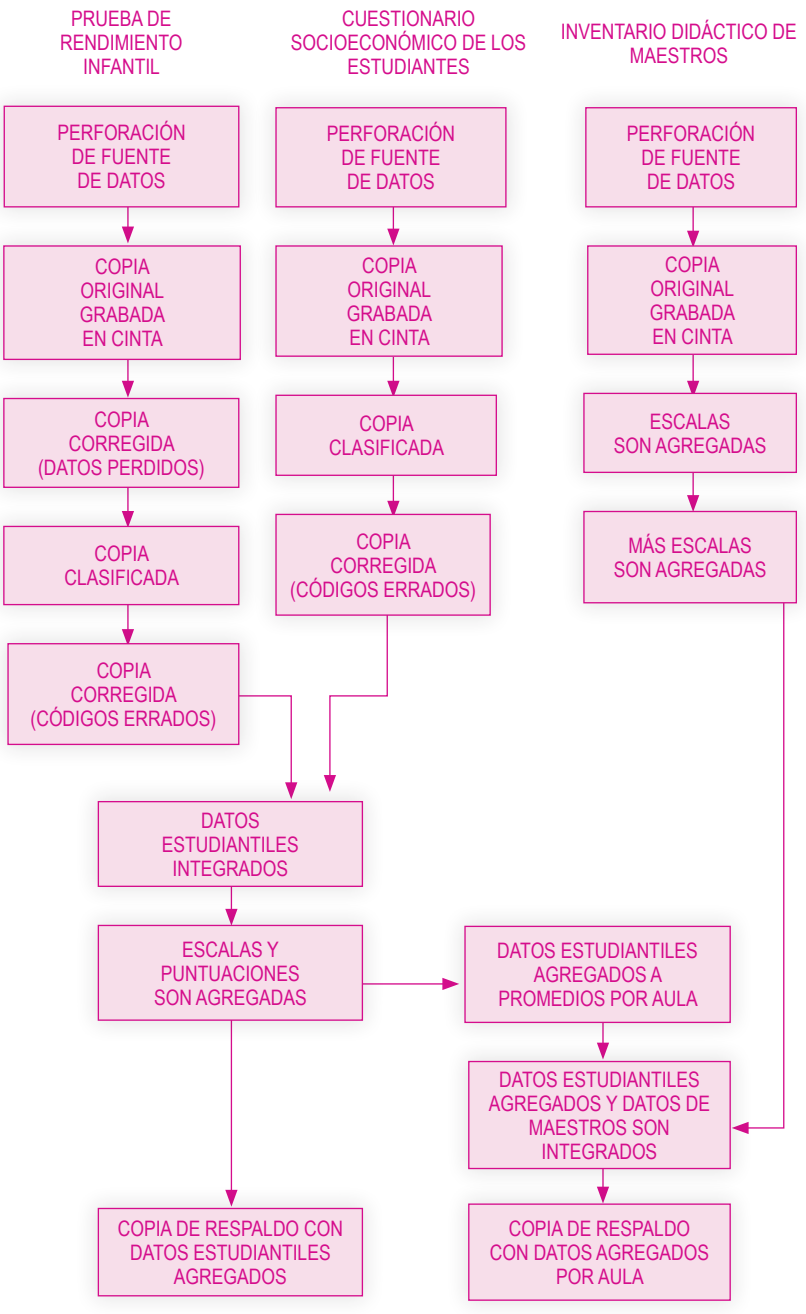
Dentro de este contexto, es útil adoptar la "reconstructibilidad" como el criterio para la documentación de los archivos de datos. Si el sistema de documentación asegura que el archivo de datos que se emplea para el análisis puede ser reconstruido a partir de la información almacenada y mantenida sistemáticamente, entonces y sólo entonces puede considerarse que tal archivo ha sido documentado correctamente. Esto significa que la codificación y definición de cada ítem pueden ser relacionadas sistemáticamente con su punto de partida en el archivo de datos original. Así, cuando se encuentran errores, es posible hacer correcciones y recrear los archivos correspondientes. Por supuesto, la reconstructibilidad debe estar basada en una buena documentación almacenada de manera segura y no debe depender de la memoria del equipo de investigación que trabajó en el procesamiento original. [2]

Sistema

La figura 1 presenta, a manera de ejemplo, los pasos del procesamiento de datos en un estudio que ha recolectado información utilizando dos fuentes diferentes para los estudiantes y una para los docentes. Cada uno de los archivos de datos de las fuentes originales se perfora en tarjetas y se copia en cintas. Para cada archivo se crean luego series de versiones que corresponden a los pasos de procesamientos necesarios para limpiar y clasificar los datos y para crear escalas o índices compuestos. Los dos archivos de estudiantes corregidos y clasificados se integran en uno solo y se desarrollan nuevas escalas. Entonces se crea un archivo con las medias

[2] Los procesamientos esforzados aquí parecerán demasiado rígidos, complejos y burocráticos a algunos investigadores. Quienes así opinen deben demostrar que sus datos son reconstruibles; deben probar, con evidencia clara, lo correcto de las variables contenidas en sus archivos.

FIGURA 1.
Pasos del procesamiento de datos en una investigación típica



por aula de las variables medidas a nivel de estudiante. Este archivo tiene tantas observaciones como el de los docentes; al fusionar los dos archivos se forma uno solo integrado a nivel de aulas. En el ejemplo de la figura 1, ésta es una de las cintas finales que se utilizan en el análisis. Obsérvese que si bien para el análisis sólo interesan dos cintas, las etapas intermedias de análisis y procesamiento requirieron generar quince cintas.

Un primer paso indispensable es el establecimiento de un sistema de codificación que permita referirse a cada archivo y versión que se producen. En la mayoría de los proyectos de investigación basta con adoptar un esquema de codificación de dos niveles en que el primer código indique el archivo y el segundo represente la versión dentro del archivo. Los códigos correspondientes al ejemplo de la figura 1 se presentan en la figura 2. Obsérvese que cuando el procesamiento de los datos conduce a integrar la información contenida en varios archivos de datos o reducir sustancialmente la información mediante agregación, muestreo o selección de variables, se ha creado un nuevo archivo. En otras palabras, un archivo corresponde a un instrumento utilizado como fuente o a un conjunto de información creado mediante el procedimiento de datos; una versión es un paso más o menos reversible en el trabajo con un archivo y corresponde generalmente al contenido de una cinta de computadora.

FIGURA 2.
Código de archivos y versiones

Archivo 01. Datos de Prueba de Rendimiento Estudiantil

Versión 01-01. Copia original de las tarjetas grabadas en cinta
Versión 01-02. Archivo completo-datos estudiantiles perdidos con añadidos
Versión 01-03. Clasificación por aula y estudiantes dentro del aula
Versión 01-04. Códigos, errados corregidos

Archivos 02. Cuestionario Socioeconómico de los Estudiantes

Versión 02-01. Copia original de las tarjetas grabadas en cinta
Versión 02-02. Clasificación por aula y estudiantes dentro del aula
Versión 02-03. Códigos errados corregidos

Archivos 03. Inventario Didáctico de Maestros

Versión 03-01. Copia original en cinta
Versión 03-02. Clasificación por código de aula
Versión 03-03. Estatus socioeconómico y otros índices son añadidos
Versión 03-04. Escalas actitudinales son añadidas

Archivos 04. Datos Estudiantiles Agregados

Versión 04-01. Archivos de rendimiento y estatus socioeconómico estudiantil agregados
Versión 04-02. Escalas de estatus socioeconómico y puntuaciones en las pruebas agregadas
Versión 04-03. Copia de respaldo

Archivos 05. Datos Agregados por Aula

Versión 05-01. Promedios de datos estudiantiles agregados por aula
Versión 05-02. Promedios agregados y datos de maestros integrados
Versión 05-03. Copia de respaldo

A partir de esta nomenclatura básica de archivos y versiones, se puede diseñar un sistema integrado para el almacenaje de información. La figura 3 presenta un esquema general del mismo. El centro del sistema es el índice de archivos. Como se describe más adelante, éste contiene una página para cada archivo y versión creados durante el procesamiento de los datos. Se refiere a la información almacenada en los archivos de tarjetas de procesamiento, de listados y de documentos. El índice de archivos tiene un índice cruzado con el de cintas. Cada vez que se carga una cinta con una nueva información, es decir cada vez que se crea una nueva versión de un archivo, se añade una página al índice de cintas. El índice permite referirse a éstas tal como están almacenadas en el centro de computación.

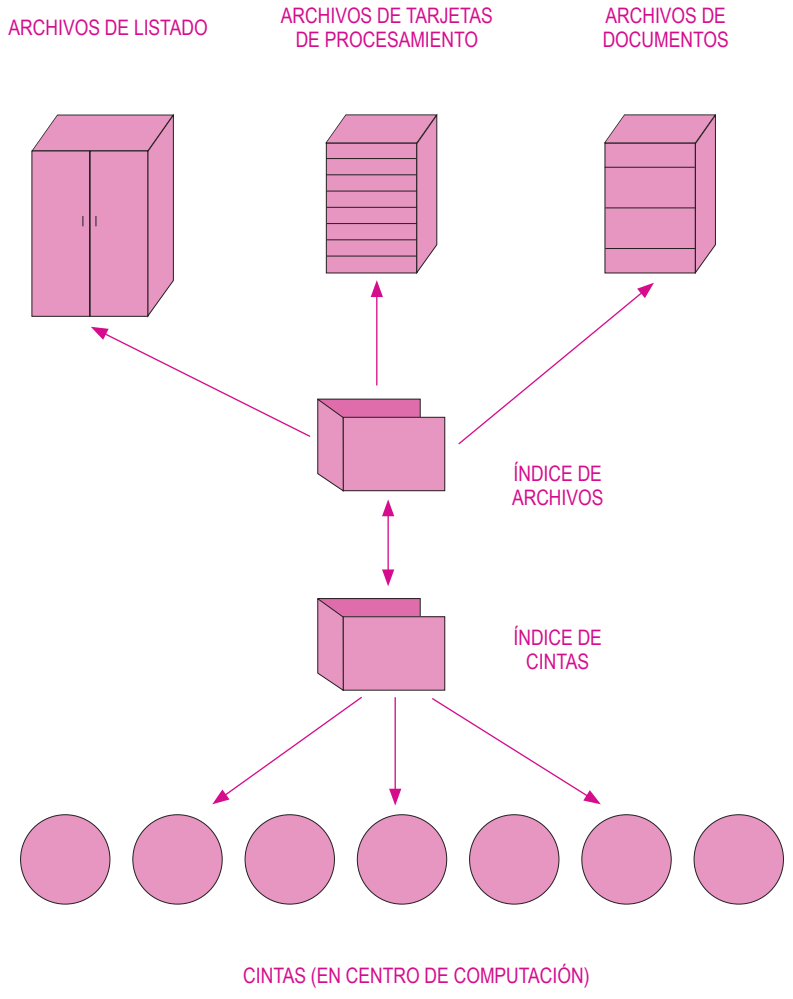
Archivo de tarjetas de procesamiento

Éste es un gabinete de tarjetas. Debe ser una regla inalterable en el procesamiento de datos que cada vez que se use un programa especial o un programa estándar con el propósito de generar una nueva versión de un archivo, las tarjetas utilizadas en la corrida se almacenen permanentemente en este archivo. Nunca habrán de usarse nuevamente, excepto para copiarlas. Se deberá establecer una norma rígida que no permita a los programadores alterar los juegos de tarjetas utilizadas en una corrida para preparar otra corrida similar. Las tarjetas empleadas para producir una nueva cinta son indispensables para reproducir la cinta y proveen información esencial a fin de documentar los contenidos de la versión del archivo. A menudo, la información contenida en las tarjetas de corridas no se encuentra impresa o descrita en ninguna otra parte. Por ejemplo, es posible que los números de identificación de escuelas eliminados de un archivo por un programa sólo puedan ser rescatados examinando los datos contenidos en las tarjetas de corridas. Para identificar los juegos de éstas habrá que escribir en ellos los códigos de las versiones correspondientes.

Archivo de listados

Es un estante que contiene las salidas de la computadora originadas por las corridas efectuadas para crear nuevas versiones. Es importante conservar todas las hojas de las salidas recibidas, incluyendo los mensajes del sistema que con frecuencia aparecen al principio y al final de la corrida. De nuevo, estos listados constituyen documentación esencial para los archivos creados y no se debe permitir que los programas se queden con dicho listado o los descarten. Cierta información no se encuentra duplicada en parte alguna: por ejemplo, la información técnica sobre la asignación y formato de las cintas de datos y las indicaciones sobre los errores de la computadora encontrados durante el procesamiento.

FIGURA 3.
Representación del sistema de almacenaje de información



Archivo de documentos

Consiste en un gabinete tipo archivo. Los archivos de tarjetas y listados se utilizan sólo en caso de error o cuando se va a realizar un trabajo de reconstrucción. Para fines de referencia y no de circulación, el archivo de documentos contiene los materiales convencionales que los usuarios necesitan para poder utilizar los datos. En este archivo deben almacenarse copias de los instrumentos originales, libros de códigos, esquemas, etc. Es útil incluir informes de muestras, listas de escuelas y tabulaciones básicas. Los documentos deben ser organizados en primer lugar por archivos y luego, dentro de cada archivo, tomando en cuenta si contienen información general sobre el archivo o información particular sobre una versión.

Índice de archivos

Es una carpeta de hojas sueltas y constituye el núcleo del sistema de manejo de datos. Contiene una Hoja Muestra de Archivo para cada archivo. La figura 4 presenta un ejemplo de dicho formulario. En la parte superior de identifica el archivo por número de código, fecha de incorporación al gabinete y un breve título descriptivo. Viene a continuación una serie de datos descriptivos, como son el resumen, la fuente (instrumentos u otros archivos), la unidad de observación (escuela, estudiante, etc.)... El formato general se presenta, por ejemplo, de la siguiente manera: 80 caracteres de registro con tres registros por estudiante. Se ha de marcar un número cada vez que se crea una nueva versión de archivo. Finalmente, en la parte inferior del formulario se presenta la biografía de la información sobre este archivo almacenada en el gabinete de documentos.

Para cada versión de archivo se inserta una Hoja de Versión. La figura 5 ofrece un ejemplo del formulario correspondiente. En la parte superior se identifica la versión por número de código, fecha de incorporación al archivo, fecha de expiración y un breve título. (Obsérvese que esta versión expira cuando se desarrollan otras más exactas o útiles. Sin embargo, la documentación de todas las versiones ha de mantenerse intacta en los gabinetes e índices). La descripción presentada en este formulario define el contenido particular de cada versión y lo diferencia de versiones previas del mismo archivo. La fuente de entrada de una versión se determina mediante los números de la cinta, del archivo y de la versión o identificando las tarjetas de datos que sirvieron como fuente. El programa de computación empleado y el proceso de computación y sus resultados definen el procesamiento que se aplicó a los datos. Se consigna luego el número de la cinta que contiene la salida, es decir, la nueva versión. Debe señalarse con marcas que los listados de las corridas de la computadora y los juegos de tarjetas utilizados han sido archivados. En la parte inferior del formulario se presenta la bibliografía de la información correspondiente a esta versión, bibliografía que está almacenada en el archivo de documentos.

Índice de cintas

Es una carpeta de hojas sueltas. En el sistema de manejo de datos descritos en este trabajo se supone que el investigador almacenará sus datos en cintas de computadoras. Aun cuando esto parezca innecesario en el caso de archivos pequeños, es recomendable hacerlo para asegurar que los archivos sean almacenados de manera

FIGURA 4.
Ejemplo de una hoja maestra de archivos

ÍNDICE DE ARCHIVOS: HOJA MAESTRA
DE ARCHIVOS PARA ARCHIVOS NÚMERO: _____

FECHA: _____

Título: _____
Descripción: _____
Resumen: _____
_____

Fuente: _____
Unidad de observación: _____
Muestra y tamaño de muestra: _____
Fechas de administración: _____

Formato general: _____
Largo del registro: _____
Registros por observación: _____

Versiones producidas (marque)

1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 20

Documentos archivados:

a. _____
b. _____
c. _____
d. _____
e. _____

FIGURA 5.
Ejemplo de una hoja de versiones

ÍNDICE DE ARCHIVO: HOJA DE DETALLES DE
VERSIONES PARA ARCHIVO NÚMERO: _____

EN VERSIÓN NÚMERO: _____
FECHA: _____
FECHA DE EXPIRACIÓN: _____

Título: _____

Descripción: _____
_____
_____

Entrada: _____
Número de cintas: _____ Número de archivo y versión: _____

Tarjetas: _____

Procesamiento: _____
Programa: _____
Procedimiento y resultados: _____

Salida: _____
Número de cinta: _____
Gabinete de listados: ☐
Gabinete de tarjetas: ☐

Documentos archivos:
a. _____
b. _____
c. _____
d. _____
e. _____

apropiada y no sean alterados inesperadamente. Generalmente, el gasto adicional por el uso de computadora es despreciable. Lo común es que el equipo a cargo del proyecto de investigación no maneje ni controle directamente las cintas de datos; casi siempre pertenecen a los centros de computación donde son numeradas y almacenadas. Dicho equipo debe informarse sobre cuáles cintas se reserva el centro de computación y sobre el manejo que se hace de sus archivos y versiones.

Este índice contiene una Hoja Maestra para cada cinta que conserva el centro de computación. La figura 6 presenta un ejemplo del formulario correspondiente. El usuario del proyecto debe conservar los números de las cintas del centro de computación para relacionarlos con el sistema de manejo de datos. En la parte superior del formulario debe asentarse el número de la cinta y la fecha de incorporación al sistema. Dado que una cinta puede ser mudada de un centro a otro, se anotarán las ubicaciones sucesivas de la misma. Al pie del formulario se marcará un número por cada nueva generación que se grabe en la cinta. Obsérvese que se supone que cada vez se almacena sólo una versión en una cinta determinada, cosa que los programadores pueden interpretar como un uso ineficiente de las cintas. Sin embargo, esta norma, cuyo costo es muy bajo, hace más seguro el almacenaje de los datos y facilita organizar mejor la documentación de los datos de un proyecto. Las cintas para archivo múltiples sólo deberán utilizarse cuando sea necesario desde el punto de vista económico o cuando los archivos de datos hayan alcanzado su forma definitiva y estén listos para ser archivados permanentemente.

Cada vez que se almacene una nueva generación en una cinta,[3] borrándose su contenido previo, se añadirá al índice de cintas una nueva Hoja de Detalles de Generación. La figura 7 presenta un ejemplo del formulario que puede utilizarse. En la parte superior se identifica la generación con el número de cinta, número de generación, fecha de incorporación en el índice de cintas y fecha de expiración (la fecha de expiración es la misma de la versión; en este caso indica cuándo la cinta podrá ser usada nuevamente). Luego, se anotan los códigos y títulos informativos para relacionar la generación grabada en la cinta con el archivo y versión apropiada, tal como se presentan en el índice de archivos. En el centro del formulario se consigna la información técnica sobre el número del trabajo, fecha y hora de generación. La parte inferir del formulario provee más información técnica sobre las características de la grabación de los datos en la cinta.

Preparación y uso

Tres gabinetes y dos carpetas constituyen físicamente el sistema de manejo de datos antes descrito. No son necesarios equipos especiales ni gastos extraordinarios.

[3] Sorprendentemente, es común encontrar que el investigador carece de una clara conceptualización de lo que es una cinta magnética de computación estándar. Ordinariamente, los datos grabados en una cinta son leídos y escritos serialmente; los datos no pueden ser insertados o reemplazados. En consecuencia, cuando se va a escribir una nueva versión se selecciona una nueva cita para la salida y se escriben de nuevo los contenidos previos. Adoptamos aquí el término “generación” para referirnos a la grabación o regrabación particular de un conjunto de datos en una cinta de computación.

FIGURA 6.
Ejemplo de una hoja maestra de cintas

ÍNDICE DE CINTAS: HOJAS MAESTRA DE CINTAS PARA CINTA NÚMERO: _____

FECHA: _____

Tipo:
- Marca: _____
- Largo: _____
- Clasificación: _____
- Dueño: _____
- Fecha de reservación: _____

Ubicación:
- a. _____
- b. _____
- c. _____
- d. _____
- e. _____
- f. _____
- g. _____
- h. _____
- i. _____
- j. _____
- k. _____
- l. _____

Generaciones Producidas (marque)

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

FIGURA 7.
Ejemplo de una hoja de detalles de generación

ÍNDICE DE CINTAS: HOJA DE DETALLES DE GENERACIÓN PARA CINTA NÚMERO: _____

EN GENERACIÓN NÚMERO: _____
FECHA: _____
FECHA DE EXPIRACIÓN: _____

Contenidos
- Número de archivo: Título: _____
- Número de versión: Título: _____

Creación:
- Computador: _____
- Programa: _____
- Número de trabajo: _____
- Fecha y hora: _____

Características técnicas:

Tracks:	7	9	
Densidad:	800	1600	6250
Coding:	EBCDIC	ASCII	
Labelling:	Ninguno	ASCII	IBM

- Data set name: _____
- Número de archivo: _____
- Formato de registro: _____
- Largo de registro: _____
- Blocking: _____
- Otro: _____

Los componentes del sistema deben estar ubicados en el mismo lugar y, si es posible, cerrados bajo llave cuando no estén siendo utilizados. El otro aspecto del sistema es el conjunto de normas que han de regular las operaciones de programadores y usuarios. Debe destacarse que cada vez que se grabe una cinta habrá de registrarse la correspondiente información en los archivos e índices. Los pasos del establecimiento del sistema y las normas para su funcionamiento se presentan de manera resumida en la figura 8.

FIGURA 8.
Pasos para el establecimiento y reglas para el funcionamiento del sistema de manejo de datos

1. Establecimiento del Sistema

a) Preparar la carpeta de hojas sueltas para el Índice de Archivos y el Índice de Cintas
b) Imprimir las copias de los formularios
c) Arreglo de los gabinetes para tarjetas, listados y documentos

2. Organización Inicial (también adición de nuevos archivos)

a) Asignación de números a los archivos y preparación de las hojas maestras de información para el Índice de Archivos
b) Almacenar copias de los instrumentos, libros de código, etc., en el Gabinete de Documentos
c) Preparar hojas maestras de información en el Índice de Cintas para todas las cintas reservadas en el centro de computación

3. Diseño del Procesamiento y Análisis de Datos

a) Identificar los archivos requeridos y sus versiones más recientes en el Índice de Archivos
b) Consultar los materiales citados en el archivo de Documentos y, posiblemente, las tarjetas de procesamientos y el Archivo de Listado
c) Obtener información técnica sobre las cintas del Índice de Cintas (el número de la cinta está en la hoja de versión en el índice de Archivos)
d) Si es necesario, seleccionar en el Índice de Cintas las cintas que han expirado para usarlas en las salidas
e) Preparar y hacer las corridas en el computador; si se crean nuevas cintas, siga los pasos que se indican a continuación

4. Documentación de Archivos de Datos Creados

a) Determinar si se ha creado un nuevo archivo o una nueva versión, complétense los formularios apropiados del índice
b) Completar una nueva hoja de generación para el Índice de Cintas
c) Almacenar las tarjetas de procesamiento y los listados en sus archivos correspondientes
d) Colocar cualquier material relevante en el Archivo de Documentos

5. Operaciones Periódicas

a) Colocar las fechas de expiración en versiones que ya no se necesiten; hacer lo propio con las hojas de generación correspondientes
b) Como respaldo, copiar nuevas versiones de los archivos críticos; aplíquese el procesamiento usual de documentación.

Es conveniente que dentro de una organización de investigación se establezca una pequeña unidad administrativa para asegurar que se cumplan las normas del sistema y que los índices y archivos se desarrollen de la manera apropiada. Podrá designarse una persona como bibliotecario de datos, responsable de asignar códigos y cintas, archivar documentos y ayudar a los usuarios en la utilización de los datos. Un punto clave del control lo constituye el hecho de que los programadores y otros usuarios tengan que pedir al bibliotecario las cintas para las salidas; esta medida le permite saber la fecha en que habrá de devolverse cierta documentación y, en consecuencia, le es fácil reclamarla.[4]

Complicaciones

El sistema aquí propuesto se basa en los principios de la documentación histórica y la reconstructibilidad de los archivo de datos. Dicho sistema constituye un banco de datos práctico que es de suma importancia para el desarrollo de un proyecto de investigación; sin embargo, no genera archivos de documentación tan fáciles de usar como uno desearía para el caso de análisis secundarios. Una organización complementaria estaría constituida por un archivo de datos en el cual se almacenaría, como en una biblioteca, sólo los archivos definitivos y su documentación. La referencia básica para establecer archivos de datos en gran escala es Nasatir (1973).

En algunas partes se comienza a sustituir las cintas de computación por discos en línea, lo cual crea nuevos problemas para el mantenimiento de la documentación de los archivos de datos. Ocurre que los usuarios pueden cambiar los registros individuales en un archivo o disco sin rehacer el juego de datos. De igual manera, puede ocurrir que no exista una copia adecuada del programa, procedimiento o instrucción utilizados para alterar o producir un archivo. El autor está probando actualmente algunos sistemas semiautomáticos para resolver estos problemas. La clave parece estar en el procesamiento de datos comerciales: a medida que se construye un archivo, se debe registrar un audit trail de manera de poder seguir y construir todos os puntos de los datos.

REFERENCIAS

Nasatir, David
1973 *Data archives for the social Sciences: Purposes, Operations and Problems,* París, UNESCO.

Rodríguez, Carlos y Richard Wolf
1971 "Banco de datos: manual de operaciones (versión preliminar)", Caracas, Ministerio de Educación, Dirección de Planeación, Departamento de Investigaciones Educacionales.

[4] Existe un control más rígido en el banco de datos del Ministerio de Educación de Venezuela, que se menciona en la nota 1, porque el bibliotecario controla también el procesamiento de datos y las corridas de análisis y no devuelve los resultados al usuario en tanto no se haya complementado toda la documentación apropiada.