Notas y comentarios

El formato Redatam*

Pablo De Grande**

El paquete estadístico Redatam es un software desarrollado por la CEPAL y utilizado ampliamente en los países de América para la difusión de estadísticas censales. Aunque es de uso gratuito, su código no es abierto y la estructura del formato utilizado para alojar la información no es pública. En este artículo se presentan resultados de un trabajo de investigación sobre la estructura de datos de esta herramienta. Entre ellos se destacan: a) *una especificación preliminar del formato Redatam,* b) *la publicación de una herramienta para la exportación de bases de datos Redatam, y* c) *la evidencia respecto de que, contrariando lo establecido en la documentación técnica, el software no implementa estrategias de compresión y de encriptación de los microdatos por él almacenados.*

Palabras clave: acceso abierto; CEPAL; Redatam; análisis estadístico; confidencialidad.

Fecha de recepción: 6 de julio de 2015.

Fecha de aceptación: 16 de febrero de 2016.

The Redatam format

The Redatam statistical package is a software package developed by ECLAC and widely used in countries of America for the dissemination of census statistics. Although it is free to use, it is licensed as proprietary software (not open source) and stores its data in a non-public format. This article introduces research results describing the data structure used by this software. They include: a) *a preliminary specification of the Redatam format,* b) *a tool for accessing and exporting its databases, and* c) *the evidence that –contrary to what the technical documentation states– Redatam does not implement strategies for compression and encryption of the microdata it stores.*

Key words: open access; CEPAL; Redatam; statistical analysis; confidentiality.

* Agradezco en la elaboración de esta publicación los sensibles y provechosos comentarios de Alejandra Russo, Mariana Bordoni y los pares evaluadores. Quiero también agradecer la confianza y el apoyo dados por Agustín Salvia al conducir con tanta paciencia mi interés hacia la indagación de datos censales y otros problemas estadísticos.

** Universidad del Salvador, Instituto de Investigación en Ciencias Sociales (Idicso). Dirección postal: Pte. Perón 1818, piso 3, Ciudad de Buenos Aires (C1089AAU), Argentina. Correo electrónico: <pablodg@gmail.com>.

Introducción

Con la creciente disponibilidad de medios informáticos para la circulación de grandes bases de datos, la tensión entre dos derechos fundamentales se instaló como desafío en la difusión de resultados censales. Por una parte, la protección de la privacidad de los datos proporcionados por las personas sugiere que sólo un pequeño número de cuadros técnicos debería poder consultar los datos obtenidos, en estrictos términos de confidencialidad. Por otra, la relevancia social de las dimensiones estudiadas impulsa a la defensa del derecho al pleno acceso a dicha información estadística para su análisis y discusión.

Este es el problema en torno al cual surgió Redatam, un paquete desarrollado por la CEPAL para hacer compatible la circulación de datos censales con la protección de la confidencialidad de los datos personales que potencialmente pudieran estar allí contenidos.

En la actualidad Redatam es un software para la distribución y explotación de datos demográficos altamente difundido en países de América. Fue creado por Serge Poulard en el Centro Latinoamericano y Caribeño de Demografía (Celade), el cual es parte de la Comisión Económica para América Latina (CEPAL) de las Naciones Unidas. El Celade es, a su vez, el encargado del mantenimiento y de la distribución del mismo, organiza cursos y difunde material y nuevas versiones periódicamente.

Dando cuenta de su origen en el año 1986, en las últimas dos décadas Redatam se ha convertido en un "estándar de hecho" para la publicación de resultados censales. Así, Argentina, Colombia, Chile, México y Perú, entre otros, han adoptado esta herramienta para difundir sus bases censales, tanto vía web como en la modalidad de aplicación de escritorio para Windows.

La modalidad de uso de Redatam permite a sus usuarios calcular totales y porcentajes en función de las variables disponibles. De esta forma, por ejemplo, es posible consultar la cantidad de hogares en una localidad o provincia, o la cantidad de personas desocupadas por sexo y edad.[1]

Esta herramienta ha representado un avance muy significativo en las capacidades de los usuarios en una diversidad de centros de inves-

[1] A través de una sintaxis específica, el software permite construir variables *ad hoc* en cualquiera de los niveles (por ejemplo, localidad, hogar, persona), dando flexibilidad de consulta con la condición irremovible de que las salidas sean conteos simples (o sus porcentajes directos).

tigación y dependencias estatales para realizar tabulados dinámicamente con datos censales. Ante las restricciones de las oficinas de estadística para facilitar datos primarios de sus censos, Redatam constituyó una propuesta superadora para la distribución de material estadístico. En este sentido, la misma resultó suficientemente conservadora como para ser aceptada por los productores de la información, y suficientemente potente como para ser adoptada (con las necesidades de capacitación que ello supuso) por investigadores y cuadros técnicos en la medida en que las bases de datos estuvieron disponibles.

Si bien Redatam es un software de uso público, el cual es posible descargar e instalar gratuitamente, cabe señalar que su código fuente no ha sido puesto a disposición de la comunidad académica (es decir, no es de código abierto). Asimismo, el formato utilizado para alojar los datos no ha sido documentado por la CEPAL ni por otros organismos. Este artículo presenta resultados de investigación vinculados al último de los dos aspectos mencionados, es decir, al carácter *cerrado* del formato de los datos.

Para asegurar una mayor transparencia de los procesos de investigación y una mejor capacidad de análisis de la comunidad científica sobre los datos demográficos disponibles, se planteó como meta analizar el formato en que Redatam almacena la información. Esto debía lograr dos objetivos: por una parte, evaluar el nivel de confiabilidad en la protección de los datos que ofrece el software, a la vez que –en caso de poderse decodificar el modo en que era guardada la información– permitir formas de análisis más complejas y dinámicas para los datos disponibles.

Como resultado de este trabajo, se ha arribado a una especificación parcial de la estructura de datos que se utiliza en la distribución de bases de datos Redatam. Dicha especificación permitió también elaborar una herramienta de código abierto para la lectura integral de bases de datos en formato Redatam (disponible en http://www.aacademica.org/conversor.redatam).

En la siguiente sección se discuten algunas limitaciones del paquete Redatam, destacándose sus barreras para la explotación estadística de la información y la ausencia de especificaciones respecto de la protección que realiza sobre los datos distribuidos. En segundo lugar, en la sección Metodología se indican los insumos y procedimientos con los que se realizó el análisis. En tercer lugar, en la sección Resultados, se detalla la estructura de datos inferida de las bases de datos Redatam. Finalmente, en las secciones Discusión y Conclusiones, se

resumen y ponen en contexto los principales hallazgos realizados, reconsiderando en qué posición se encuentra el equilibro entre confidencialidad y acceso luego del uso de Redatam en buena parte de las series censales 2000-2010 de América.

Dificultades para el análisis estadístico y la evaluación del resguardo de la información

Esta investigación se inició con la pregunta sobre cómo era posible que hubiera tanta información valiosa en formato Redatam y no se pudiera procesarla estadísticamente de maneras complejas. Redatam ofrece desde hace casi veinte años una aplicación y una sintaxis de consulta para extraer totales y porcentajes, sin opciones para extender esa capacidad de cálculo.

En la medida en que los datos sólo pueden consultarse como tabulados simples, el uso de ellos para operaciones estadísticas más avanzadas se ve gravemente obstaculizado. Para elaborar un análisis de varianzas, calcular la confianza de una diferencia de medias o realizar modelos inferenciales se requiere de extenso trabajo *artesanal* de extracción de datos calculando los totales para todas las combinaciones de categorías de las variables involucradas, reconstruyendo luego parcialmente con ellas una base de trabajo.[2] Este tipo de uso, aun en los casos en que es posible, supone habilidades avanzadas en Redatam, dificulta el análisis exploratorio, y requiere de grandes cantidades de tiempo para resolver operaciones que son básicas cuando se cuenta con los datos en forma de filas de una tabla.

En este punto resulta problemático no tanto el hecho de que Redatam no realice operaciones más complejas, sino que éste no pueda ser ampliado por terceros ni interactuar con otros paquetes estadísticos. Las vías para este tipo de interacciones podrían ser diversas, pero cabe destacar al menos tres de amplia utilización en la integración de sistemas:

1) Interfaces de extensibilidad: son frecuentes los casos en que los paquetes de software ofrecen canales para agregar módulos programados en forma externa que interactúen con la aplicación principal. Así, por ejemplo, el programa de información geográfica ArcGis permite a través de scripts en varios lengua-

[2] Esta estrategia fue utilizada por ejemplo en De Grande y Salvia, 2008.

jes acceder y modificar las capas de datos de sus mapas; el caso de las *macros* en Microsoft Office, los *plugins* en los navegadores web o las aplicaciones en los sistemas operativos (tradicionales y móviles) son casos de extensibilidad exitosa por esta vía.

2) Apertura del formato de datos: a través de una documentación controlada de las versiones del formato en que se guardan los datos es posible dar la posibilidad a otros proveedores de software o a investigadores y equipos de investigación de hacer compatibles sus aplicaciones con el propio formato. En el caso de Redatam, sólo utiliza formatos conocidos en la exportación de los resultados de los tabulados. El formato Acrobat PDF es un caso de éxito de crecimiento por medio de un formato especificado en forma pública.

3) Apertura del código fuente: la disponibilidad abierta del código de un paquete de software permite a otros programadores examinar las instrucciones que forman parte de un programa, realizando aportes o mejoras al mismo. Conocer los mecanismos internos de una aplicación permite también con frecuencia, para quien pueda leer el lenguaje en que fue realizado, despejar dudas y aprender sobre el comportamiento detallado del programa en cuestión. El sistema operativo Linux y el paquete estadístico R+ son dos casos exitosos de extensibilidad por apertura del código fuente.

Cualquiera de estos tres caminos podría permitir a la comunidad de usuarios de Redatam un crecimiento hacia formas de análisis más avanzadas sobre los datos actualmente disponibles.

Un segundo punto de partida para esta investigación estuvo representado por la pregunta sobre qué tan protegidos estaban los datos en una base de datos Redatam. Las principales oficinas de estadística han distribuido sus datos en Redatam, en lugar de hacerlo en formatos más conocidos (tales como tablas en DBF o bases de datos de SPSS), confiando posiblemente en que era un modo efectivo de resguardar la confidencialidad de datos primarios.

En este sentido, la CEPAL presenta a Redatam como un paquete que protege los microdatos encriptándolos (CEPAL, 2015), no permitiendo así que personas ajenas a los productores de la información puedan acceder a ella. En la introducción de la documentación de Redatam se indica:

> Los censos de población y vivienda, los censos agropecuarios, las encuestas de hogares, los registros vitales, etc., son bases de datos que contienen millones de registros sobre las viviendas, hogares y personas. Estos datos, organizados jerárquicamente en un formato Redatam son almacenados en forma encriptada y altamente comprimida, protegiéndose así el secreto estadístico de la persona misma (CEPAL, 2015).

De la misma manera, se resaltaba este aspecto en el lanzamiento de su versión del año 2002, al afirmarse que "las bases de datos externas se convierten al formato propio de Redatam, que comprime, encripta e invierte los datos originales con el fin de combinar la eficiencia con la confidencialidad de la información" (Faijer y Poulard, 2002: 326).

Pero, ¿cómo encripta Redatam la información? La criptografía es una disciplina específica, la cual ha ganado enorme masividad en los últimos treinta años (Katz y Lindell, 2007) con el desarrollo de protocolos para proteger conexiones de internet, operaciones bancarias, documentos, firmas personales y correos electrónicos, entre otros. Sin embargo, la documentación técnica de Redatam no da pistas sobre la clase de encriptación que realiza sobre la información. Del mismo modo, a la fecha no se han hallado registros de verificaciones sobre la fortaleza de este último aspecto por parte de la comunidad académica o de los institutos estadísticos que lo utilizan.

Al tratarse de un paquete orientado a datos potencialmente confidenciales, esta investigación buscó aportar claridad sobre este aspecto, para permitir así a las oficinas estadísticas nacionales y a la comunidad de usuarios de Redatam en general, poder decidir de manera informada sobre qué columnas incorporar o no en las bases de datos, conocida la confiabilidad de los resguardos ofrecidos.

Metodología

Para analizar el esquema de almacenamiento de Redatam se utilizó un conjunto de bases de datos públicos en dicho formato, así como el paquete Redatam en su versión de escritorio para Windows R+SP V5. Esta versión tiene la capacidad de acceder y crear bases de datos, permitiendo desempeñar tanto los roles de consumidor de estadísticas como de productor de bases de datos. La misma puede descargarse en forma pública desde la página de la CEPAL.

El análisis se llevó adelante por medio de tres estrategias desplegadas en paralelo: por un lado, se siguieron criterios típicos de ingeniería inversa para investigar formatos desconocidos, observando variaciones en archivos simples (Eilam, 2005: 200); por otro, se analizaron muestras de bases de datos existentes en circulación; por último, se generó una herramienta que validara la hipótesis en construcción orientada a reconstruir los sets de microdatos originales.

En la primera estrategia se produjeron grupos de archivos elementales y se examinaron sus variaciones. Esto significó tomar como punto de partida la creación de una base de datos con solamente una tabla de una fila y una columna de tipo entero. Luego se agregó una variable adicional de igual tipo. Después se modificó el tipo de dato, y así sucesivamente, observando en las bases de datos los cambios producidos por la herramienta.

Para llevar adelante la segunda estrategia se definió un corpus de bases de datos preexistentes a utilizar como referencia. El mismo se constituyó principalmente a partir de las bases de datos disponibles en formato Redatam y SPSS en la página web del Instituto de Estadísticas y Censos de Argentina.[3] La selección de estas bases de datos de control tuvo como objetivo validar lo observado en bases pequeñas a partir de bases de datos *reales*, generadas en diferentes momentos y bajo diferentes necesidades. Asimismo, fueron utilizadas para observar cualitativamente los rasgos sobresalientes de la estructura de datos investigada, como la cantidad de archivos típica, las extensiones utilizadas o los tamaños generales de los archivos.

Para poder verificar en el curso de la investigación de manera veloz y masiva los hallazgos producidos para la descripción del formato, como tercera estrategia metodológica se desarrolló una herramienta que implementara estas definiciones y las aplicara en la realización de una reconstrucción de los microdatos contenidos en las bases de datos de Redatam. Esta herramienta tomó el nombre de Conversor Redatam, y se encuentra disponible en código abierto para su evaluación y uso experimental en el repositorio GitHub.[4] La misma cuenta en la actualidad con la capacidad de exportar a archivos SPSS (.sav) o a archivos de texto plano (.csv), la estructura y los microdatos desde bases de datos Redatam. Usuarios externos que descargaron la aplicación reportaron haber convertido con éxito bases de datos censales de Argentina, Bolivia, Chile y Uruguay.

[3] <http://www.indec.gov.ar/bases-de-datos.asp>.
[4] <https://github.com/discontinuos/redatam-converter>.

Resultados

Como se indicó anteriormente, el análisis realizado ha avanzado hasta el punto de tener una especificación parcial pero suficiente para la lectura total de los microdatos de una base de datos Redatam. En esta sección se presenta la estructura de archivos y datos reconocida, especificando la función de cada tipo de archivo y su estructura interna.

En primer lugar, pudo reconocerse que las bases de datos de Redatam se organizaban a partir de un archivo de "diccionario", que poseía la lista de entidades y variables y su definición. Adicionalmente al diccionario, existían también archivos de datos (donde estaban los valores para cada fila de cada variable) y archivos de correspondencias (donde se indica la relación entre las entidades de diferentes niveles, tales como a qué países corresponde cada provincia, o a qué hogar corresponde cada persona en una base de datos).

A continuación, se describen los tipos de datos identificados, para luego especificar los tipos de archivo en que estaban contenidos en las bases de datos analizadas.

Tipos de datos

En el marco del reconocimiento del formato de almacenamiento de Redatam se examinaron las variantes de datos que el software utiliza.

En el caso de los valores de texto pudo observarse que Redatam almacena cadenas de tamaño variable en la descripción del diccionario (que aquí llamaremos el tipo STRING)[5] y de tamaño fijo (que aquí llamaremos el tipo CHAR) en los archivos de datos. En ambos casos, los caracteres se almacenan utilizando la tabla de códigos Windows-1252 (8 bits).

En el caso de los valores numéricos con decimales, Redatam almacena para su persistencia valores con coma flotante de ocho bytes (tipo que llamaremos DOUBLE). Para los valores enteros, utiliza un conjunto de tipos de dato variable en función del rango de los valores a almacenar (los que llamaremos tipos INT16, INT32, y BITS(n)).

En el cuadro 1 se especifican estos tipos de datos, los cuales se utilizan en las descripciones posteriores para indicar las formas de almacenamiento de cada valor.

[5] En algunos casos los nombres de los tipos se desprenden de la denominación utilizada en Redatam; en otros se asignó un nombre *ad hoc*, buscando utilizar términos usuales en la especificación de estructuras de datos de paquetes o lenguajes informáticos.

CUADRO 1
Tipos de datos utilizados en la descripción

Tipo de dato	*Descripción*	*Ejemplo*
BITS(n)	Almacena secuencias de bits de tamaño arbitrario para alojar números enteros. Los valores de los campos BITS se recuperan leyendo enteros INT32, por lo que una serie de valores BITS siempre tendrá un tamaño múltiplo de 4 bytes.	0xA0860100 => 11000011010100000 => BITS(4) => 12; 3; 5; 0
BYTE	Número entero sin signo de 1 byte.	0x02
BYTE[]	Secuencia de bytes de tamaño variable.	0x02050202040405
CHAR(n)	Secuencia de caracteres de tamaño fijo. Al igual que el tipo STRING, los caracteres especiales se codifican siguiendo la tabla de caracteres predeterminados de Windows, o Windows-1252.	0x504552524F => PERRO
DOUBLE	Número de coma flotante, almacenado siguiendo el estándar IEEE 754 que utilizan la mayor parte de los lenguajes de programación.	0x547424971F88B340 => 5000,1234
INT16	Número entero sin signo de 2 bytes.	0x0401 => 260
INT32	Número entero sin signo de 4 bytes.	0xA0860100 => 100.000
STRING	Almacena cadenas de texto de tamaño variable. Presenta 2 bytes al inicio describiendo el tamaño del texto contenido, luego de lo cual se encuentra el texto propiamente dicho. En caso de requerir almacenar cadenas iguales o más largas que 65 535 caracteres (el tamaño máximo especificable en 2 bytes) indica el valor 65 535 en los primeros 2 bytes y reserva posteriormente un entero de 4 bytes para describir la duración del texto extenso.	0x43415341 => CASA
Nota: En todos los casos en que se almacenan valores mayores a 1 byte, la modalidad de almacenamiento es *little-endian*, es decir, el byte más pequeño se almacena primero.		0x0204 => 0x04; 0x02

Fuente: Elaboración propia con base en análisis de archivos.

Archivo de diccionario

En cuanto al archivo de diccionario, se constató que éste almacena la lista de entidades que componen la base de datos, incluyendo el detalle de variables y etiquetas para cada una de ellas. El esquema de datos de Redatam supone la existencia de datos jerárquicos, es decir, de un universo de datos en el cual las entidades se relacionan en la modalidad *padre-hija*. Típicamente en las estructuras censales esta relación toma la forma de una secuencia cuyo nivel superior es el país, el siguiente nivel es la provincia o estado, el siguiente son los departamentos, partidos o localidades, siguiendo niveles intermedios hasta llegar a los de vivienda, hogar y persona.

La estructura del archivo cuenta con un encabezado que posee atributos generales de la base de datos, el cual no ha sido descrito en esta etapa de la investigación por no ser vinculante para la descripción de los datos. A continuación del encabezado se encuentra una lista de bloques que describen a cada uno de los tipos de entidades contenidos en la base de datos (por ejemplo, provincias, departamentos, hogares, personas).

Cada bloque de entidad, a su vez, se descompone de una lista inicial de atributos de la entidad (como su nombre, su entidad padre, el nombre de su variable de identificadores), seguida de una lista de bloques descriptores de cada variable que la entidad posee. El bloque de cada variable incluye a su vez atributos de la misma, que indican el tipo de datos, el nombre, la descripción extendida (su etiqueta) y las etiquetas de los valores posibles de la variable, entre otros elementos. En el cuadro 2 se encuentra una descripción detallada de estas estructuras.

Archivo de correspondencias

La observación arrojó también como resultado que los archivos .PTR (que hemos llamado aquí "de correspondencias") funcionan como índices o tablas de referencias para determinar a qué entidad de un nivel superior corresponde una entidad de un nivel inferior. Existe un archivo de correspondencia por cada tipo de entidad contenida en la base de datos. Los mismos permiten resolver, por ejemplo, a la hora de calcular un resultado, en qué provincia se encuentra cierto departamento, o en qué hogar se encuentra cierta persona.

CUADRO 2
Ficha descriptiva del tipo de archivo "diccionario"

Descripción			
Tipo de archivo	Diccionario		
Extensión	DIC		
Nivel de especificación	Parcial		
Objeto	Contiene el listado de entidades y sus variables (columnas).		
Estructura			
Campo	*Contenido*	*Descripción*	*Ejemplo*
Encabezado	BYTE[]	Desconocido. Reúne un grupo de datos que preceden a las entidades y que no fue analizado debido a que no aparecía como necesario para la lectura de los datos.	
Entidades	Secuencia de entidades	A continuación del encabezado se suceden entradas que describen a las entidades que forman parte de la base de datos.	
Nombre1	STRING	Nombre de la entidad.	DPTO
Nombre2	STRING	Repite el valor anterior. Se omite si la entidad no tiene padre (nivel superior).	DPTO
Padre	STRING	Nombre de la entidad superior respecto de la actual. STRING vacío en caso de ser la entidad del nivel superior.	PROV
Descripción	STRING	Descripción extendida de la entidad.	Departamento
Archivo de correspondencias	STRING	Detalla qué archivo describe las correspondencias de la entidad con su entidad padre.	CV100000.ptr
<desconocido>	INT16	2 bytes de uso no identificado.	

(continúa)

CUADRO 2
(continúa)

Campo	*Contenido*	*Descripción*	*Ejemplo*
Variable de identificadores	STRING	Especifica el nombre de la variable dentro de la entidad; mantiene códigos descriptivos de cada fila.	PROVID
Variable de descriptores	STRING	Especifica el nombre de la variable dentro de la entidad; mantiene descripciones textuales de cada fila.	PROVINCIA
<desconocido>	INT32	4 bytes de uso no identificado.	
<desconocido>	BYTE	1 byte de uso no identificado.	
Cantidad de variables (?)	INT32	Cantidad de variables. No resultó consistente en la totalidad de las bases, por lo que el conversor no utiliza este valor.	12
<pie>	BYTE[]	Desconocido. Final de la descripción de la entidad. No se decodificaron los valores correspondientes, no resultando necesarios para extraer la información.	
Variables	Secuencia de variables	Luego se suceden entradas describiendo cada variable de la entidad. El inicio de las mismas se reconoce por la existencia de entradas en la forma "<nombre de variable> DATASET"	12
Nombre	STRING	*Nombre de la variable*	PROV
Declaración	STRING	La declaración se especifica luego del prefijo DATASET. La misma consiste en tres elementos, separados los espacios. Los mismos son: el tipo de dato de la variable, el archivo donde se encuentran almacenados los datos correspondientes a la variable y el tamaño. Para la indicación del tipo de dato, los valores posibles son: BIN: valores enteros con tamaño fijo especificable almacenados en bloques de 4 bytes en modo *big-endian*. CHR: valores de texto con tamaño fijo especificable.	DATASET CHR 'CP200000.rbf' SIZE 2 DATASET BIN 'CP4541.bin' SIZE 7

		DBL: valores con decimales (con coma flotante) especificados en 8 bytes.[a] INT: valores enteros de 0 a 65 535. LNG: valores enteros de 0 a 4 294 967 296. PCK: valores enteros con tamaño fijo especificable almacenados en bloques de 4 bytes en modo *little-endian*. El tamaño es indicado en bytes en el caso de las variables CHR y en bits en el caso de las variables de tipo BIN y PCK. Las variables de tipo INT, LNG y DBL son de tamaño fijo, siendo 2, 4 y 8 bytes respectivamente.	
Filtro	STRING	Indica si la variable debe utilizarse solamente en ciertas condiciones.	VIVIENDA. V02 = 1 AND HOGAR. NHOG = 1
Rango	STRING	Valor mínimo y máximo posibles para variables numéricas, separados por el término 'TO'.	1 TO 10
Tipo	STRING	Tipo de dato almacenado, indicando si se trata de valores numéricos o de texto. Los valores posibles son INTEGER para enteros, REAL para números con decimales y STRING para texto.	INTEGER
Etiquetas	STRING	La lista de etiquetas a utilizar para la variable. Las entradas se encuentran separadas por Tabs (carácter 9), y los valores se encuentran separados de las etiquetas por espacios.	1 Varon{TAB} 2 Mujer

(continúa)

CUADRO 2
(concluye)

Descripción	STRING	Descripción extendida de la variable (etiqueta de la variable).	País de nacimiento
Descriptores	STRING	Se almacena una lista de elementos que permiten describir aspectos adicionales de la variable o sus valores. Los atributos son opcionales y se guardan separados por espacios. Los mismos son: ALIAS: permite definir un nombre alternativo a la variable. DECIMALS: cantidad de decimales a mostrar para los tipos de datos REAL. GROUP: permite indicar el nombre del grupo en que debe visualizarse la variable. MISSING: indica el valor que señala datos no registrados. NOTAPPLICABLE: indica el valor que señala datos no pertinentes.	MISSING 4 NOTAPPLICABLE 0 GROUP EDUCACION ALIAS ALFAB

[a] El rango de un tipo de dato de doble precisión (8 bytes) es -1.79769313486231570E+308 a -4.94065645841246544E-324 para valores negativos y 4.94065645841246544E-324 a 1.79769313486231570E+308.

Fuente: Elaboración propia con base en análisis de archivos.

La manera en que esto se resuelve es manteniendo en estos archivos una lista ordenada con tantos elementos como tenga la entidad de nivel superior. Cada uno de estos elementos contiene la cantidad de entidades de nivel inferior que se corresponden con la entidad de nivel superior, las cuales se encuentran ordenadas respetando dicho criterio (cuál es su entidad superior).

Tomemos por caso un ejemplo en el cual existe una tabla con 24 provincias, de la que depende otra tabla con 240 departamentos. El archivo de correspondencias indicado para la entidad "Departamentos" contendrá 24 elementos (luego de un valor de inicio en cero que posee el archivo), explicitando en cada uno de ellos la cantidad de departamentos que corresponden a cada provincia. Si los departamentos fueran homogéneos en su distribución –es decir, si cada provincia tuviera 10 departamentos en su jurisdicción– la lista estaría compuesta de una serie de 24 valores 10 (la cantidad de departamentos en cada provincia). Si en cambio la primera provincia tuviera 15 departamentos y la segunda tuviera 5, el contenido del archivo de correspondencias se iniciaría con el número 0, tal como siempre comienza, luego habría un 15 y luego un 5. El detalle de esta estructura puede observarse en el cuadro 3.

Archivo de datos

Los archivos de datos del paquete analizado, indicados en el diccionario para cada variable, contienen la información de los valores que cada variable posee en cada entidad. Esto implica que existe un archivo de datos para cada variable (por ejemplo, Persona. Edad, Persona. Sexo, Persona.Ocupación). Por esta razón no hay un archivo único de datos para cada tipo de entidad (como Personas), por lo que la consulta de una lista de entidades requiere la lectura de varios archivos en forma simultánea.

Esta estrategia posiblemente haya sido adoptada para acelerar la lectura de datos, ya que de este modo Redatam sólo accede a los bloques de datos correspondientes a las variables seleccionadas en cada consulta, evitando leer el registro completo de la entidad. El detalle de la estructura de almacenamiento se encuentra especificado en el cuadro 4.

CUADRO 3
Ficha descriptiva del tipo de archivo "correspondencias"

Descripción			
Tipo de archivo	Archivo de correspondencias		
Extensión	PTR		
Nivel de especificación	Completo		
Objeto	Contiene el modo en que se relacionan las entidades de diferentes niveles.		
Estructura			
Campo	*Contenido*	*Descripción*	*Ejemplo*
Fila inicial	INT32	Valor constante en cero.	0x00000000
Listado de filas por entidad	Secuencia de INT32	Presenta una secuencia de valores que indican la cantidad de filas de la entidad hija que corresponden con la entidad padre.	
Filas por entidad	INT32	Valor para la fila correspondiente a la posición en la lista.	512

Fuente: Elaboración propia con base en análisis de archivos.

CUADRO 4
Ficha descriptiva del tipo de archivo "datos"

Descripción			
Tipo de archivo	Archivo de datos		
Extensión	RBF. En bases de datos más antiguas puede encontrarse la extensión .BIN.		
Nivel de especificación	Completo		
Objeto	Contiene los valores correspondientes a una variable de una entidad. La estructura depende del tipo de dato almacenado.		
Estructura			
Campo	*Tamaño*	*Descripción*	*Ejemplo*
Listado de valores	Secuencia de valores	Presenta una secuencia de valores que permiten reconstruir el contenido de la variable para cada fila de la entidad. La lista tendrá tantas filas como elementos existan para la entidad.	
Estructura para tipo de dato BIN			
Valor	BITS(n)	Valor entero de tamaño arbitrario correspondiente a la posición en la lista. Las series de valores de BITS persisten en bloques de 4 bytes, los cuales poseen en primer lugar el byte de mayor valor (orden *little-endian*). Las bases más antiguas utilizan el formato de datasets BIN, mientras que las más modernas utilizan el formato PCK.	12; 3; 5; 0.
Estructura para tipo de dato CHR			
Valor	CHAR(n)	Valor de texto de longitud fija para la fila correspondiente a la posición en la lista.	PERRO
Estructura para tipo de dato DBL			
Valor	DOUBLE	Valor con coma flotante para la fila correspondiente a la posición en la lista.	5000,1234

(continúa)

CUADRO 4
(concluye)

Estructura			
Campo	*Tamaño*	*Descripción*	*Ejemplo*
Estructura para tipo de dato INT			
Valor	INT16	Valor entero corto para la fila correspondiente a la posición en la lista.	512
Estructura para tipo de dato LNG			
Valor	INT32	Valor entero largo para la fila correspondiente a la posición en la lista.	19772501
Estructura para tipo de dato PCK			
Valor	BITS(n)	Valor entero de tamaño arbitrario correspondiente a la posición en la lista. Las series de valores de BITS persisten en bloques de 4 bytes, los cuales forman un entero en formato *big-endian* (es decir, los bytes de mayor peso se encuentran al final). Al igual que en el tipo BIN, una vez leído el bloque de 4 bytes se toma la cantidad de bits correspondientes a cada elemento sucesivo.	17; 1; 8; 2.

Fuente: Elaboración propia con base en análisis de archivos.

Discusión

En criptografía y seguridad informática el término *seguridad por oscuridad* refiere a la estrategia por la cual se busca que una protección sea efectiva gracias a mantener en secreto los procedimientos de su aseguramiento. En oposición a ello existe en la criptografía contemporánea un consenso respecto a la validez del principio de Kerckhoffs, el cual sostiene que en un sistema criptográfico "nada debe ser secreto salvo su clave": es decir, que para maximizar la seguridad de una protección, el funcionamiento de sus mecanismos debe ser conocido (Ferguson, Schneier y Kohno, 2010: 74). De este modo, las formas de encriptación utilizadas para intercambios de datos cifrados en internet (como el protocolo SSL/TSL o el protocolo IPSec) se encuentran documentados en forma pública y en constante proceso de revisión y discusión por la comunidad de analistas en seguridad informática (Stapleton, 2014).

En el caso de Redatam, hemos dado con un caso límite de seguridad por oscuridad: la confianza en que el esquema de guardado de los datos iba a mantenerse oculto parece haber conducido ya no a una implementación de encriptación débil, sino a ninguna encriptación.

En este sentido, cabe señalar que los resultados de esta exploración resultaron en parte inesperados, en la medida en que el equipo de Redatam afirmara, al menos desde el año 2002, que el software trabajaba comprimiendo y encriptando los datos. Según se pudo constatar, ninguna de ambas afirmaciones es exacta.

Respecto al uso del espacio (la compresión), sólo puede afirmarse que Redatam guarda los datos en forma normalizada,[6] es decir, guarda los datos sin repetir por ejemplo los datos de vivienda en cada hogar, o los datos de cada hogar en cada persona. En este sentido se comporta a la manera de cualquier base de datos relacional, almacenando una tabla para cada tipo de entidad y alojando los datos en función de su tamaño. Sin embargo, ni en bibliografía reciente (Román González, 2012) ni en bibliografía antigua (Coello y León, 1994) normalizar una base de datos constituye específicamente un método de compresión de datos.

En relación con la encriptación –y éste es quizás el aspecto más problemático– no se encontró durante el análisis ninguna estrategia explícita de protección de los datos. Cada registro se encontraba al-

[6] Para una definición precisa de la noción de normalización, véase Silberschatz, Korth y Sudarshan, 2002.

macenado uno debajo del otro, sin alteraciones ni en los textos ni en los números que representaban los valores, ni en el orden de los datos individuales de cada registro. Desde la estrategia más rudimentaria de encriptación –tal como tener una tabla de sustituciones– hasta el uso de algoritmos validados que permiten cifrar o firmar la información, nada de ello era parte de los datos consultados en las bases accedidas de Redatam. Así pues, como consecuencia de la ausencia de estrategias de cifrado, los microdatos en las bases de datos Redatam pueden ser leídos en forma directa.[7] Asimismo, como consecuencia de la ausencia de estrategias de firmado de los datos, los mismos pueden ser modificados intencional o accidentalmente sin que Redatam o sus usuarios puedan validarlo.

Retomando los planteamientos hechos al inicio de este artículo, cabe preguntarse cómo afectan estos hallazgos al estado actual de la tensión entre protección y difusión de datos censales. Tal como se ha dicho anteriormente, Redatam ha permitido ampliar la capacidad disponible de análisis de la comunidad científica sobre microdatos censales al producir una publicación generalizada de bases de datos. Sin embargo, tras veinte años de progreso en esta dirección, nos hallamos en una coyuntura que pone límites de importancia a esta estrategia: por un lado –con el facilitamiento del uso de técnicas de estadística avanzada–, el software Redatam no resulta tan flexible como muchos de sus usuarios lo requieren. Por otro lado, ya no es posible afirmar que el paquete Redatam proteja los microdatos como se sostuvo hasta aquí: es posible, de manera trivial, convertir una base de datos Redatam a listados de hogares y personas en formatos estándar de base de datos. Ambos hechos sugieren la necesidad de revisar las políticas de publicación y distribución de la información estadística de cara a los censos por venir.

[7] Cabe señalar aquí que si bien es un problema de importancia que el software publicite capacidades que no despliega, la salvaguarda de la privacidad individual se encuentra cubierta en gran medida por el hecho de que los institutos de estadística remueven de sus bases de datos las columnas que involucran datos personales tales como los nombres, teléfonos y direcciones de personas antes de convertirlas al formato Redatam. Un país que adopta como política esta perspectiva es Uruguay, el cual distribuye sus bases de datos censales a nivel de microdatos en forma pública (en formato DBF y SPSS), considerándolos suficientemente anónimos como para permitir su difusión.

Conclusiones

En síntesis, se ha avanzado hacia una especificación preliminar del formato Redatam. Se ha destacado la necesidad de transparentar los procesos de investigación, incluidos la circulación y el uso de la información estadística. Como parte de esta investigación se produjo una herramienta portable, extensible y de código abierto (De Grande, 2015) que permite validar supuestos respecto al formato Redatam. Esta herramienta ha podido leer y exportar con éxito la totalidad de las bases de datos evaluadas hasta la fecha. La exportación de los datos en formato Redatam emerge como un paso crucial para un análisis en profundidad de la información censal disponible y de la situación real frente al equilibro entre accesibilidad y confidencialidad.

Bibliografía

CEPAL (2015), *Tutoría básica R+SP Process*, Santiago de Chile, Comisión Económica para América Latina y el Caribe <http://www.redatam.org/cdr/Tutoriales/Process_Esp.html> (30 de junio de 2015).

Coello, C. y H. Hernández de León (1994), "Compresión de bases de datos", *Actas del VIII Simposio Internacional en Aplicaciones de Informática*, Antofagasta, 21 a 25 de noviembre, pp. 87-94.

De Grande, P. (2015), *Conversor Redatam (software)*, Buenos Aires, Discontinuos. Disponible en: <http://www.aacademica.org/conversor.redatam> (13 de enero de 2016).

De Grande, P. y A. Salvia (2008), "Segregación residencial socioeconómica y espacio social: deserción escolar de los jóvenes en el área metropolitana de Gran Buenos Aires", en Agustín Salvia (comp.), *Jóvenes promesas. Trabajo, educación y exclusión social de jóvenes pobres en la Argentina*, Buenos Aires, Miño y Dávila. Disponible en <http://www.aacademica.com/pablo.de.grande/5> (12 de abril de 2015).

Faijer, D. y S. Poulard (2002), "El software REDATAM para divulgación y análisis de datos censales", *Notas de Población*, vol. 75, pp. 321-341. Disponible en: <http://repositorio.cepal.org/bitstream/handle/11362/12742/np75321341_es.pdf?sequence=1> (18 de mayo de 2015).

Eilam, E. (2005), *Reversing: secrets of reverse engineering*, Indianapolis, Wiley.

Ferguson, N., B. Schneier y T. Kohno (2010), *Cryptography Engineering. Design Principles and Practical Applications*, Indianapolis, Wiley Publishing.

Katz, J. e Y. Lindell (2007). *Introduction to Modern Cryptography: Principles and Protocols,* Boca Raton, CRC Press.

Román González, A. (2012), "Clasificación de datos basado en compresión",

Revista ECIPerú, vol. 9, núm. 1, pp. 69-74. Disponible en: <https://hal.archives-ouvertes.fr/hal-00697873/document> (18 de mayo de 2015).

Silberschatz, A., H. Korth y S. Sudarshan (2002), *Fundamentos de base de datos*, Madrid, McGraw-Hill.

Stapleton, J. (2014), *Security without Obscurity. A Guide to Confidentiality, Authentication, and Integrity*, Boca Raton, CRC Press.