IHSN

INTERNATIONAL HOUSEHOLD SURVEY NETWORK

RIEH (IHSN) - Red Internacional de Encuestas de Hogares

Difusión de archivos de microdatos

Principios, procedimientos y prácticas

Olivier Dupriez y Ernie Boyko

IHSN Documento de trabajo n° 005
Agosto de 2010

www.ihsn.org

Difusión de archivos de microdatos
Principios, procedimientos y prácticas

Olivier Dupriez y Ernie Boyko

IHSN Documento de trabajo n° 005

Agosto de 2010

Resumen

Los productores de datos de todos los países se enfrentan a una creciente demanda de microdatos. Decidir cuál es la mejor forma de difundir estos datos constituye un verdadero desafío. Este desafío es de naturaleza técnica, ya que implica poner en marcha procedimientos de documentación, catalogación y difusión de los datos, pero también implica acciones de índole jurídica y ética. Aunque los productores de datos son conscientes del poder y la importancia de los microdatos, también deben considerar esta demanda junto con la necesidad de garantizar la confidencialidad de la información suministrada por los informantes. Esta obligación viene impuesta por la legislación nacional en materia estadística y de confidencialidad de los datos y a menudo se materializa en un compromiso contraído con los informantes y que se comunica en el momento del levantamiento de la información. En este contexto, la difusión de microdatos implica la creación de políticas y procedimientos formales que definan las condiciones de acceso a los microdatos. El presente documento contiene una descripción general de estas políticas y procedimientos e identifica buenas prácticas existentes en este ámbito.

Los autores

Ernie Boyko es uno de los fundadores de Statistics Canada, donde ha dirigido variasdivisiones de estadísticas, entre ellas, Agricultura, Planificación Corporativa, Difusión y la División de operaciones del censo de 1991. También supervisó los trabajos de la Iniciativa para la Democratización de los Datos (IDD). Ernie Boyko es miembro activo de la Asociación Canadiense de Usuarios de Datos Públicos (ACUDP) y de la Asociación Internacional para los Servicios y Técnicas de Información en Ciencias Sociales (IASSIST).

Olivier Dupriez un experto economista y estadístico es miembro del Grupo de Gestión de Datos para el Desarrollo en el Banco Mundial. Además, es coordinador de la Red Internacional de Encuestas de Hogares (RIEH). También coordina los programas de asistencia técnica de un gran número de países en los ámbitos relacionados con la documentación y la difusión de microdatos.

Agradecimientos

El presente documento ha sido elaborado por la Red Internacional de Encuestas de Hogares (IHSN por sus siglas en inglés) con el apoyo económico del mecanismo DGF (Development Grant Facility) del Banco Mundial, cuya subvención (nº 4001009-06) ha sido gestionada por la Secretaría de PARIS21 en la OCDE. El documento ha sido preparado por Ernie Boyko y Olivier Dupriez con la colaboración de las personas que figuran a continuación, que han provisto de insumos o han realizado observaciones y sugerencias al texto: François Fonteneau (PARIS21, OCDE), Julia Lane (National Opinion Research Center, Universidad de Chicago), Johan Mistiaen (Banco Mundial), Dennis Trewin y Wendy Watkins (Universidad de Carleton, Canadá).

El presente documento incorpora también las conversaciones mantenidas con numerosos colegas de las agencias pertenecientes a la red RIEH, así como de los responsables estadísticos oficiales de varios países. Ha sido preparado para su publicación por John Wright y la tipografía fue realizada por Rhommell Rico.

.

El presente documento (o una copia actualizada) se encuentra disponible en la página web de la IHSN, en www.ihsn.org.

Referencia

Dupriez, Olivier y Boyko. 2010. Difusión de archivos de microdatos. Definición de políticas y procedimientos, Red Internacional de Encuestas de Hogares (IHSN), documento de trabajo No 005.

Índice

Anexos

Lista de gráficos

Lista de cuadros

Siglas y acrónimos

ABS	Australian Bureau of Statistics: Oficina de Estadística de Australia
ACS	American Community Survey: Encuesta anual realizada en EE. UU. a muestras de población
API	Application Programming Interface: Interfaz de programación de aplicaciones
ASCII	American Standard Code for Information Interchange: Formato normalizado para el intercambio de información
AUP	Archivo de microdatos de uso público
CDR	Siglas en francés de Centro de Datos de Investigación (Statistique Canada)
CENEX	Centre of Excellence for Statistical Disclosure Control: Centro de Excelencia para el Cotrol de la Divulgación de Estadísticas, un proyecto europeo que elabora una lista de las prácticas habituales en Europa en materia de protección de datos
CEPE-ONU	Comisión Económica para Europa de las Naciones Unidas
CESSDA	Council of European Social Science Data Archives: Consejo de Archivos Europeos de Datos de Ciencias Sociales
CSO	Central Statistics Office: Oficina Central de Estadísticas de Irlanda
CURF	Confidentialised Unit Record Files: Archivos de Registros Unitarios Anonimizados
DCMI	Dublin Core Metadata Initiative: Organización que impulsa las actividades relacionadas con la creación de un esquema de metadatos genérico
DDI	Data Documentation Initiative: Proyecto centrado en el establecimiento de normas de documentación técnica en el ámbito de las ciencias sociales
DHS	Demographic and Health Surveys: Programa mundial de encuestas demográficas y de salud
DSNU	División de Estadística de las Naciones Unidas
EE. UU.	Estados Unidos de América
GPS	Global Positioning System: Sistema de geolocalización operativo en todo el mundo
HTML	HyperText Markup Language: Formato de datos concebido para mostrar las páginas web
HTTP	Hypertext Transfer Protocol: Protocolo de transferencia de hipertexto
ICPSR	Inter-University Consortium for Political and Social Research: Consorcio Interuniversi tario para la Investigación en Ciencias Políticas y Sociales
IDD	Initiative de Démocratisation des Données (Statistique Canada): Iniciativa para la de mocratización de los datos de la oficina de estadística de Canadá
IHSN	International Household Survey Network: Red Internacional de Encuestas de Hogares (RIEH)
ISO/CEI	Organización Internacional de Normalización/Comisión Electrotécnica Internacional
JSI	Job Submission Interface: Interfaz que permite publicar y consultar ofertas de empleo
LIS	Luxemburg Income Study: Base de datos procedente de un estudio sobre las rentas
MCRDC	Michigan Census Research Data Center: Centro de Investigaciones sobre el Censo de Michigan
MICS	Multiple Indicator Cluster Surveys: Metodología de encuestas de indicadores múltiples
MIT	Instituto Tecnológico de Massachusetts
NCHS	National Center for Health Statistics: Centro Nacional de Estadísticas de Salud de EE. UU.
NCSA	National Center for Supercomputing Applications: Centro estadounidense para la investigación y explotación de aplicaciones de alto rendimiento
NDE	National Data Enclave: Centro de datos seguro
NORC	National Opinion Research Center (Universidad de Chicago)

NSD	Norwegian Social Science Data Services: Centro de almacenamiento de datos de Noruega
OAIS	Open Archival Information System: Modelo de referencia para un sistema abierto de archivo de la información
OCDE	Organización para la Cooperación y el Desarrollo Económicos
OCLC	Online Computer Library Center: Organismo internacional de investigación sin ánimo de lucro que ofrece servicios a las bibliotecas que favorecen el acceso de estas a la información en todo el mundo
ONE	Oficina Nacional de Estadística
ONG	Organización no Gubernamental
ONU	Organización de las Naciones Unidas
PDF	Portable Document Format: Formato de archivo desarrollo por Adobe
PUMA	Public Use Microdata Areas: Unidades de microdatos de uso público (UMUP)
PUMS	Public Use Microdata Sample: Muestras de microdatos de uso público (MMUP)
SAS	Statistical Analysis System (software)
SDC	Statistical Disclosure Control: Control de Divulgación de Estadísticas
SEN	Sistema Estadístico Nacional
SNZ	Statistics New Zealand: Instituto estadístico de Nueva Zelanda
SOAP	Simple Object Access Protocol: Protocolo orientado a objetos basado en XML
SQL	Structured Query Language: Lenguaje de bases de datos
UKDA	United Kingdom Data Archive: Centro de almacenamiento de datos del Reino Unido alojado por la Universidad de Essex
UNF	Universal Numeric Fingerprint: Firma electrónica universal
UNICEF	Fondo de las Naciones Unidas para la Infancia
UPM	Unidad Primaria de Muestreo
URL	Uniform Resource Locator: Cadena de caracteres utilizada para resolver las direcciones de los recursos en Internet
USB	Universal Serial Bus: norma referente a un bus informático de transmisión en serie que sirve para conectar periféricos a un ordenador
XML	eXtensible Markup Language: Lenguaje extensible de marcado
XSL	Extensible Stylesheet Language: Lenguaje de estructuración de datos

Introducción

El levantamiento de datos estadísticos, destinado a fundamentar los procesos de toma de decisiones en el ámbito privado y público de un país, es un emprendimiento enorme que generalmente se financia con fondos públicos. Es responsabilidad de todos los productores de datos que reciben estos fondos, de los investigadores y de los organismos financiadores garantizar la máxima rentabilidad de la inversión impulsando el aprovechamiento de estos datos.

Los datos socioeconómicos, que constituyen el eje central de este documento, proceden de los censos, de las encuestas por muestreo y de los sistemas de registros administrativos. Estas actividades generan microdatos (observaciones al nivel del informante). Estos microdatos pueden ser tratados posteriormente (editados, analizados y tabulados) antes de ponerse a disposición de los usuarios. Habitualmente, los resultados consisten en datos agregados que se presentan en forma de tablas, gráficos, resúmenes, informes descriptivos y documentos analíticos. El contenido de estas tablas e informes depende de su importancia para los productores de datos y los auspiciantes. La mayoría de las actividades de levantamiento de datos se llevan a cabo con fines específicos; alcanzar estos fines específicos suele ser la prioridad y a menudo el único objetivo del productor de datos o del auspiciante. No obstante, los microdatos recogidos con un fin particular a menudo pueden ser útiles para muchos otros fines, incluyendo algunos que no pueden anticiparse al momento del levantamiento de los datos. Dicho de otro modo, los microdatos pueden servir para otros propósitos distintos de los inicialmente previstos. Proveer y facilitar el acceso a los datos a la comunidad investigadora es una forma económica y eficaz de multiplicar y diversificar el análisis y la explotación de la información existente. Una explotación exhaustiva de estos datos ofrece oportunidades cuasi ilimitadas para generar nuevos conocimientos.

El aumento constante de la potencia de las computadoras y las aplicaciones informáticas desde la década de 1980 ha acentuado el interés de los investigadores en los microdatos. Los productores de datos de todos los países se enfrentan a una creciente demanda de acceso a los microdatos que constituyen la base de las estadísticas que se publican. El acceso a los microdatos no solo permite efectuar nuevas investigaciones y más diversificadas, sino también permite desarrollar métodos innovadores para usar, procesar y publicar la información, sin olvidar la creación de nuevas bases de datos al combinar datos provenientes de distintas y múltiples fuentes.

Sin embargo, decidir cuál es la mejor forma de difundir los microdatos plantea un importante desafío a los productores. Este desafío es tanto de orden técnico como organizacional, ya que implica instaurar procedimientos apropiados para la documentación, catalogación y difusión de los microdatos. En este contexto, la comunidad de archivistas de datos ha elaborado estándares y buenas prácticas en la materia para responder a estos temas. Sin embargo, el desafío es también de naturaleza jurídica y ética: si bien los productores de datos son plenamente conscientes del poder y de la importancia de la difusión de los microdatos, deben considerar esta demanda junto con la necesidad de garantizar la confidencialidad de la información suministrada por los informantes. Esta obligación viene impuesta por las legislaciones nacionales en materia estadística y de confidencialidad de los datos y a menudo se materializa en un compromiso contraído con los informantes al momento del levantamiento de la información. Los institutos nacionales de estadística y otro tipo de productores de datos deben actuar de forma que se preserve la confianza de los informantes, ya que se corre el riesgo de reducir la disposición a cooperar con las encuestas y consecuentemente la calidad de las estadísticas disminuiría. La difusión de los microdatos implica, pues, la elaboración de políticas y procedimientos que definan formalmente las condiciones de acceso a dicha información.

El marco previsto para tal efecto varía de un país a otro. No obstante, «independientemente de las diferencias que puedan existir entre las prácticas y políticas relativas a la difusión de los datos, así como de las restricciones legítimas a las que pueda estar sometido el acceso a los datos, una difusión más sistemática sería beneficiosa para prácticamente todos los tipos de investigación». [17]

La presente guía tiene por objeto ayudar a los productores y depositarios o compiladores de los microdatos a elaborar sus propias políticas y procedimientos en materia de difusión de archivos de microdatos. Es primordial que estas políticas y procedimientos sean formales y transparentes. Una adecuada difusión de los microdatos conlleva no solo la entrega de los datos y la documentación correspondiente, sino también la definición de las condiciones que rigen el uso y explotación de estos datos. Esta información debe ser pública y de fácil acceso, preferentemente a través de Internet.

Aunque la mayor parte de este documento es una guía de carácter genérico, está destinado principalmente para los productores de datos oficiales (INEs, Ministerios y similares) de los países en desarrollo. Los datos a los que nos referimos corresponden generalmente a los microdatos extraídos de encuestas por muestreo, censos y sistemas de recolección de datos administrativos.

Esta guía ha sido elaborada bajo el auspicio de la Red Internacional de Encuestas de Hogares (IHSN, por sus iniciales en inglés). Está basada fundamentalmente en los trabajos de la Comisión Económica para Europa de las Naciones Unidas (CEPE-ONU), específicamente por el Grupo de trabajo a cargo de la Gestión de la Confidencialidad Estadística y del Acceso a los Microdatos designado/administrado por la Conferencia de Técnicos Estadísticos Europeos y por la Oficina de Estadística de la Unión Europea (Eurostat) [5] [24] [25] [26]. La guía también se apoya en la experiencia de los institutos de estadística de regiones del mundo donde el acceso a microdatos es una práctica arraigada (con más de 40 años en algunos casos), pero también de diversos centros de datos académicos.

La información contenida en este documento responde a doce grandes preguntas que se plantean a la hora de formular una política de difusión de los archivos de microdatos:

1. ¿Qué son los microdatos?
2. ¿Qué son los metadatos?
3. ¿Por qué los productores de datos deberían difundir microdatos?
4. ¿Cuáles son los costes y los riesgos asociados a la difusión de microdatos y cómo se pueden abordar?
5. ¿Para quienes deberían estar disponibles los microdatos?
6. ¿Qué condiciones deben cumplirse para la difusión de microdatos?
7. ¿Qué se entiende por «anonimización» de los microdatos?
8. El acceso a los microdatos, ¿debe ser pagado o gratuito?
9. ¿En qué momento del ciclo de difusión deben publicarse los microdatos?
10. ¿Cuáles son los requisitos en cuanto a infraestructura técnica para diseminar microdatos?
11. ¿Cuáles son los requisitos institucionales relacionados con la difusión de microdatos?
12. ¿Cómo promover la utilización de archivos de microdatos?

La presente guía trata principalmente los aspectos relacionados con políticas de difusión de los microdatos. No obstante, para garantizar una difusión apropiada y segura de los microdatos, también se necesitan soluciones técnicas para la documentación, anonimización, catalogación y archivo de los datos y los metadatos. En este documento estos temas se abordan brevemente, pero son tratados con mayor detalle en otros trabajos publicados por la red IHSN y otros organismos.

1. ¿Qué son los microdatos?

1.1 ¿Qué son los microdatos?

Las encuestas/censos o la recogida de datos administrativos permiten a los institutos de estadística u otros productores de datos recabar información sobre cada una de las unidades de observación: un hogar, un individuo, una empresa, una explotación agraria, una escuela, un hospital, etc. Los *microdatos* a los que nos referimos aquí son archivos de datos digitales que contienen la información de cada una de estas unidades de observación. Al respecto, los microdatos son lo opuesto a los *macrodatos* (o *datos agregados*), que son una síntesis de la información individual recopilada y que se presentan en forma de medias, porcentajes, frecuencias u otras estadísticas de síntesis.

Los microdatos generalmente se estructuran en archivos de datos en donde cada línea (o *registro*) contiene información sobre una unidad de observación. Esta información se registra en forma de *variables* de diferentes tipos (numéricas o alfanuméricas, discretas o continuas, etc.) y pueden obtenerse directamente de la persona encuestada, a través de un cuestionario, o mediante observación, medición (localización por GPS, por ejemplo), imputación o cálculo.

La información contenida en los archivos de microdatos estadísticos está codificada. Así, el sexo de la persona informante puede estar registrado en forma de una variable con el nombre «*H01a*», y que podría adoptar el valor 1 o 2 (1 para masculino, 2 para femenino). Por consiguiente, los microdatos deben estar acompañados de un *diccionario de datos* que contenga la lista de las variables utilizadas, una descripción de su contenido y el significado de cada código. Estos *metadatos* constituyen la documentación mínima requerida. Sin embargo, como veremos en el capítulo 2, el número de metadatos necesario es, en realidad, mucho más elevado.

Un conjunto de datos de una encuesta o un censo generalmente comprende varios archivos de datos, a menudo fruto de varios niveles de observación producto de una misma operación de levantamiento de datos. En la mayor parte de los casos, los censos y otras encuestas realizadas a los hogares permiten recabar datos al menos a dos niveles: el hogar (con variables que describen las características de la vivienda, por ejemplo) y el individuo (con información sobre la edad, el estado civil, el nivel de educación y la actividad económica, por ejemplo). Un conjunto de datos puede estar formado por uno o varios archivos para cada uno de estos niveles. Estos contienen *variables clave (identificadores únicos o llaves)* que permiten a los usuarios vincular la información contenida en diferentes archivos de datos. A los conjuntos de datos organizados de este modo se les denomina *jerarquizados*.

1.2 ¿En qué formato se almacenan y difunden los archivos de microdatos?

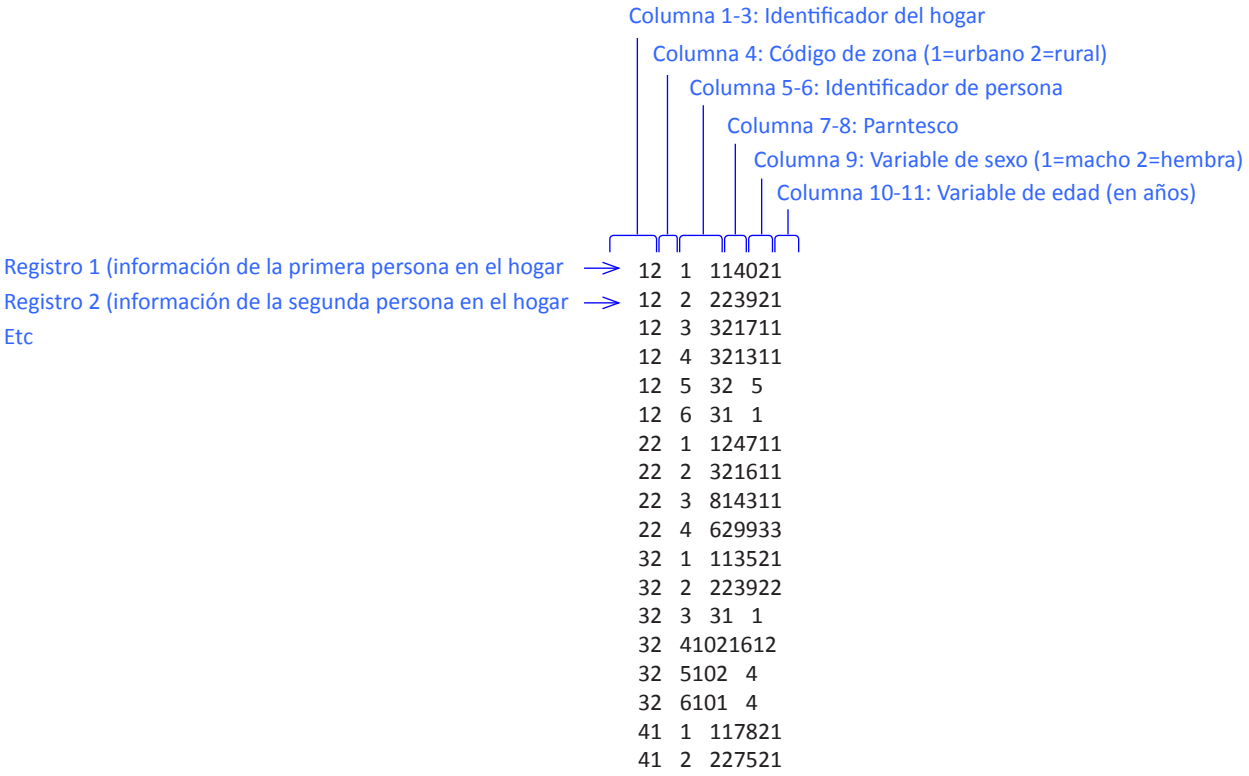
Los archivos de microdatos pueden almacenarse en diferentes formatos. Algunos de los formatos incluyen el formato no licenciado ASCII y los formatos propios creados por los programas estadísticos especializados como SAS, SPSS y Stata Los microdatos también pueden guardarse en formato SQL o en otros formatos de bases de datos. No obstante, estos formatos son menos frecuentes y su funcionalidad es limitada debido a que los programas de bases de datos no están diseñados específicamente para la creación de tablas y análisis estadísticos.

El formato de archivo ASCII no es propio de ningún programa informático o plataforma en particular. Los archivos ASCII contienen datos que pueden ser leídos por la mayoría de los programas informáticos. Al no estar vinculados a un programa informático que podría terminar quedando obsoleto, son la solución óptima para garantizar la conservación de los datos a largo plazo. Sin embargo, los datos en este formato no pueden descifrarse ni explotarse sin un diccionario de datos (archivo o documento aparte). El gráfico 1 presenta un extracto de un archivo tipo de datos estadísticos en formato ASCII fijo.

Para crear una tabla estadística o analizar datos en formato ASCII, primero hay que importarlos a otro programa informático. Todos los programas de estadística y bases de datos contienen herramientas y comandos para realizar esta tarea. La página siguiente contiene un ejemplo de *script* de Stata que permite importar los datos ASCII del gráfico 1 y asociar etiquetas (*labels*) a las variables y a los códigos para hacer que el contenido sea aprovechable para el usuario. Es evidente que para elaborar este tipo de *script*, el usuario debe disponer de un diccionario de datos que describa el contenido y la estructura del archivo de datos ASCII.

Una vez elaborado el *script* e importado en Stata, el archivo ASCII del gráfico 1 aparecerá como se muestra en el gráfico 2. Los formatos de archivos propios de SAS, SPSS, Stata o programas similares engloban los datos, las variables y las etiquetas correspondientes.

Gráfico 1: Extracto de un archivo de datos en formato ASCII fijo



Ejemplo de configuración (set-up) de la importación de datos ASCII en Stata

```
* . Read the ASCII data found in file test.dat and import the values in new variables;
. . infix hhid 1-3 area 4 pid 5-6 relat 7-8 sex 9 age 10-11 using test.dat;

* . Add a label to describe each new variables;
. . label variable hhid “Identificador del hogar”;
. . label variable area “Área”;
. . label variable pid “Identificador de la persona”;
. . label variable relat “Parentesco con el jefe de hogar”;
. . label variable sexo “Sexo”;
. . label variable Edad “Edad en el último cumpleaños”;

* . Add label to each code used by the variables;
. . label define relatcod 1 “Jefe de hogar” 2 “Cónyuge” 3 “Hijo/Hija” 4 “Nuera/Yerno ”
. . 5 “Nieto/Nieta” 6 “Padre/Madre” 7 “Suegro” 8 “Hermano/Hermana” 9 “Otro familiar”
. . 10 “Ningún parentesco”, add;
. . label values relat relatcod;
. . label define areacod 1 “Urbano” 2 “Rural”, add;
. . label values area areacod;
. . label define sexcod 1 “Masculino” 2 “Femenino”, add;
. . label values sexo sexcod;

* . Save the file as a Stata file;
. . save “test.dta”, replace;
```

Figura 2: Extracto de un archivo de datos Stata

	hhid	area	pid	relat	sexo	edad
1	1	Rural	1	Jefe de hogar	Masculino	40
2	1	Rural	2	Cónyuge	Femenino	39
3	1	Rural	3	Hijo/Hija	Femenino	17
4	1	Rural	4	Hijo/Hija	Femenino	13
5	1	Rural	5	Hijo/Hija	Femenino	5
6	1	Rural	6	Hijo/Hija	Masculino	1
7	2	Rural	1	Jefe de hogar	Femenino	47
8	2	Rural	2	Hijo/Hija	Femenino	16
9	2	Rural	3	Hermano/Hermana	Masculino	43
10	2	Rural	4	Padre/Madre	Femenino	99
11	3	Rural	1	Jefe de Hogar	Masculino	35
12	3	Rural	2	Cónyuge	Femenino	39
13	3	Rural	3	Hijo/Hija	Masculino	1
14	3	Rural	4	Ningún parentesco	Femenino	16
15	3	Rural	5	Ningún parentesco	Femenino	4
16	3	Rural	6	Ningún parentesco	Masculino	4
17	4	Urbano	1	Jefe de hogar	Masculino	78
18	4	Urbano	2	Cónyuge	Femenino	75

Los conjuntos de datos de encuestas pueden contener cientos de variables, incluso miles. La elaboración de *scripts* para importar y documentar estos archivos de datos a partir del formato ASCII lleva tiempo y puede conllevar errores. Para minimizar el riesgo de cometer errores y para una mayor facilidad de uso, es conveniente que los productores de datos entreguen sus archivos en formato ASCII, con modelos de *script* para SPSS, SAS y Stata, o bien en los formatos estadísticos propios más comunes. Existen aplicaciones informáticas especializadas, como StatTransfer de Stata Corporation para convertir los ficheros de datos automáticamente de un formato de paquete a otro.

1.3 ¿Qué versión de los archivos de datos se debe difundir?

Los productores de datos suelen crear varias versiones de un mismo archivo de microdatos. Estas versiones se diferencian por su calidad, contenido y el número de registros y van desde archivos de microdatos brutos/crudos que contienen todas las respuestas dadas por cada encuestado y que se obtienen directamente después de la transcripción de los datos, hasta archivos de datos depurados y validados para uso público

La figura 3 representa el ciclo de vida tipo de una encuesta o censo.

1.4 ¿Cuáles son los elementos sensibles del contenido de los microdatos?

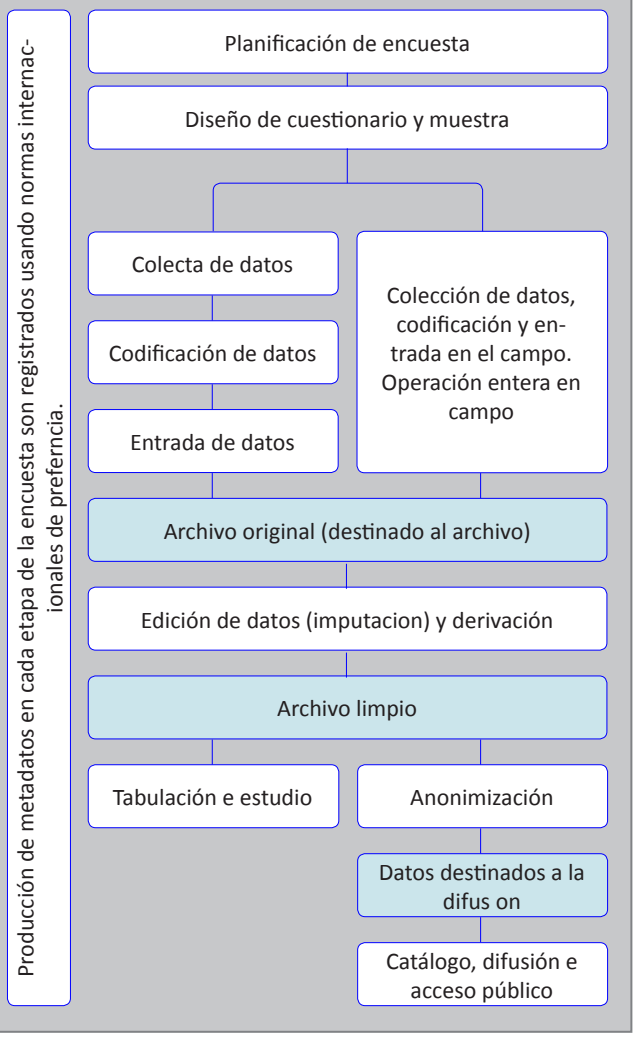
Los datos procedentes de censos y encuestas por muestreo se utilizan exclusivamente para fines estadísticos o de investigación. Por razones prácticas, a menudo el nombre y la dirección de los encuestados se mencionan en los cuestionarios, pero rara vez figuran en los archivos de datos correspondientes. Por regla general, estos últimos no contienen variables que constituyan *identificadores directos*. Por el contrario, los archivos de datos administrativos contienen con frecuencia los nombres, direcciones, números de teléfono, números de seguridad social, etc.

No obstante, la mayor parte de los conjuntos de datos comprenden *identificadores indirectos*. Tales como, la desagregación geográfica, la composición

del hogar por edad y sexo más información sobre la actividad económica pueden servir para identificar a los encuestados.

Estas variables se consideran sensibles, ya que pueden posibilitar que se identifique a las unidades de observación. Así también, existen otras variables que son sensibles por la naturaleza de la información que encierran, como el estado de salud de un individuo, sus hábitos sexuales, sus ingresos, etc. Las encuestas realizadas a empresas, por ejemplo, son sensibles por naturaleza, en la medida en que la información recogida puede ser explotada por la competencia.

Gráfico 3: Ciclo de vida de una encuesta



1.5 ¿Cuáles son los principales tipos de archivos de microdatos que se difunden?

Los archivos de microdatos destinados a ser difundidos son casi siempre diferentes de los que están reservados para el uso del personal de los organismos productores. La preparación de archivos de microdatos brutos con miras a su difusión comprende procedimientos de ajuste del contenido y/o del número de registros. La modificación del *contenido de los registros* de los archivos de microdatos para difundir consiste en suprimir los identificadores directos e indirectos para proteger el anonimato de los encuestados. Esto no se traduce forzosamente en una supresión de variables. A veces basta con agrupar (recodificar) varias variables en categorías menos detalladas y consecuentemente menos informativas. Esto puede requerir una reducción del *número de registros* contenido en el archivo de microdatos difundido (como en el caso de los datos de censos de población, por ejemplo). Los procedimientos que tienen como objetivo proteger la identidad de los encuestados se designan comúnmente con los términos de *anonimización* o *control de divulgación de estadísticas* (CDE).

Los archivos de microdatos elaborados con miras a producir estadísticas oficiales son difundibles con la condición de que pueda garantizarse correctamente el anonimato de los encuestados. En el marco de la definición de una política de difusión, deben considerarse tres tipos de archivos: los archivos de uso público, los archivos protegidos mediante licencia y los archivos accesibles únicamente en centros seguros (datos en enclave). Estos tres tipos de archivos se diferencian por su nivel de accesibilidad y su grado de anonimización.

«Ninguna persona […] puede pretender obtener datos identificables o acceder a ellos [...] en virtud únicamente de su categoría profesional. El acceso a datos identificables no viene determinado exclusivamente por la categoría profesional, por la afiliación a un organismo o por un compromiso/acuerdo financiero. La necesidad de contar con datos de identificación, el uso que se hará de ellos, así como la función y la responsabilidad del demandante en materia de recogida de datos son criterios más importantes. Debido a que el acceso a datos de identificación nunca está exento de riesgos, éste será objeto de una evaluación y de un seguimiento profundos una vez concedida la autorización correspondiente». [14]

Archivos de microdatos de uso público (AMUP)

Los archivos de microdatos de uso público (AMUP) pueden ser consultados por todas aquellas personas que acepten cumplir varias condiciones básicas. Se trata de condiciones de uso (prohibición de venta, por ejemplo) y no de acceso a los datos. Algunos AMUP se difunden sin condiciones y generalmente pueden encontrarse en Internet. Estos datos son fácilmente accesibles, ya que el riesgo de identificación de los encuestados se considera mínimo. Eso implica la supresión de todo el contenido que pueda permitir identificar directamente a los encuestados, como nombres, direcciones y números de teléfono. Además implica la eliminación de los identificadores indirectos. Estos varían en función del diseño (la metodología) de la encuesta, pero generalmente se eliminan variables con información geográfica con mayor desagregación que el nivel para el cual la encuesta es representativa. A menudo se eliminan de los AMUP algunos registros debido a la existencia de datos atípicos o variables caracterizadas por una distribución muy asimétrica. No obstante, existen otros métodos de CDE que permiten minimizar el riesgo de divulgación y elevar al máximo el contenido informativo de los datos (agrupación de los valores extremos superiores e inferiores, supresión de las variables que proceden de varios encuestados o incluso técnicas de perturbación de los datos). Los AMUP generalmente se crean a partir dearchivos de datos de censos (subconjuntos de registros en lugar de archivos completos) y encuestas realizadas a los hogares. Aunque técnicamente resulta posible crear AMUP a partir de encuestas a empresas, esta tarea plantea unos desafíos concretos que se describirán por separado.

Los AMUP deben tener un carácter lo más informativo posible. Según el NCHS (el Centro Nacional de Estadísticas de Salud de EE. UU.), «el objetivo es poner a disposición los microdatos con la mayor difusión posible y en la forma más detallada posible, siendo las únicas restricciones los recursos disponibles, las exigencias cualitativas, las limitaciones técnicas y la necesidad de proteger la confidencialidad de los datos». [14]

Archivos protegidos mediante licencia

Los *archivos protegidos mediante licencia* (o *archivos de investigación*) se diferencian de los AMUP por el hecho de que su difusión se limita a los usuarios a los que se ha autorizado el acceso después de presentar una solicitud debidamente justificada y firmar un acuerdo en el que se estipulan las condiciones de uso de los datos. Por regla general, los archivos protegidos mediante licencia también están anonimizados con el fin de reducir al mínimo el riesgo de identificación de los individuos en caso de utilización aislada. No obstante, pueden contener datos que podrían ser identificables en conjunción con otros archivos.

Los identificadores directos, como los nombres de los encuestados, deben suprimirse de los conjuntos de datos protegidos mediante licencia. Sin embargo, a veces estos archivos contienen variables indirectas que pueden servir para identificar a los encuestados cuando se cotejan con otros conjuntos de datos (por ejemplopadrones electorales, catastros o expedientes escolares).

La difusión de los archivos protegidos mediante licencia se fundamentará preferentemente en la redacción y firma de un contrato entre el productor de los datos y usuarios *de buena fe*, es decir, con necesidades legítimas de acceso a los datos. Es conveniente que este acuerdo rija el acceso y la utilización de los archivos de microdatos. Los contratos de licencia a veces se firman únicamente con usuarios miembros de un organismo promotor (o auspiciante) apropiado (centro de investigación, universidad o socio en materia de desarrollo).

Además, se recomienda a los productores de datos que pidan a los solicitantes, antes de la firma del contrato de utilización y acceso, que rellenen un formulario de solicitud que legitime su necesidad de consultar el archivo protegido mediante licencia (en lugar del AMUP correspondiente, si estuviera disponible) para un fin estadístico o de investigación declarado. Se ofrecen modelos de contrato y de formulario de solicitud de acceso a ficheros protegidos mediante licencia en el capítulo 6, que aborda las condiciones en las que debe concederse el permiso de acceso a los archivos de microdatos.

Archivos consultables en centros de datos seguros (o enclaves de datos)

Algunos archivos pueden estar disponibles para su consulta en centros seguros (o «enclaves de datos») en condiciones muy estrictas. Se trata de instalaciones equipadas con ordenadores que no están conectados a Internet ni a una red externa y de los que no se puede extraer ningún dato a través de puertos USB, unidades de CD/DVD u otros tipos de lectores. Estos «enclaves» custodian datos especialmente sensibles o que permiten identificar directa y fácilmente a los encuestados.

Pueden ser, por ejemplo, conjuntos de datos completos procedentes de censos de población, encuestas realizadas a empresas o incluso historiales médicos que contienen datos altamente confidenciales. Los usuarios que deseen consultar estos conjuntos de datos no tienen necesariamente un acceso integral, sino que podrían acceder a ellos con restricciones (acceso al subconjunto de datos necesarios). A estos usuarios se les pide que rellenen un formulario de solicitud en el que deben justificar que el acceso a los datos tenga fines estadísticos o de investigación (véase el ejemplo del capítulo 6). Los resultados obtenidos deben ser objeto de un examen minucioso en el marco de un procedimiento completo de control antes de su entrega al solicitante.

La explotación de un centro de datos seguro supone unos costes considerables, ya que implica la adecuación de un local y la compra de equipos informáticos especiales. También requiere la contratación de personal que disponga de las competencias y el tiempo necesarios para realizar los controles que permiten eliminar los riesgos de divulgación. Estas personas deben estar familiarizadas con los métodos de análisis de datos y deben saber procesar las peticiones y gestionar servidores de archivos.

Habida cuenta de los sustanciales costes de explotación que conllevan y las competencias técnicas especializadas que requieren, algunos institutos de estadística u otros productores de datos oficiales han optado por colaborar con entidades universitarias o centros de investigación para crear y gestionar los «enclaves de datos». A continuación figuran

Cuadro 1: Una encuesta, varios productos

Los productores de datos pueden optar por crear varios productos a partir de un mismo conjunto de datos de censos o encuestas. Así, los AMUP pueden generarse a partir de pequeñas muestras o subconjuntos de variables. Estos archivos pueden difundirse ampliamente sin riesgo de divulgación de la identidad de los encuestados. Así también podría crearse una versión más amplia con una muestra más extensa y con acceso sujeto a la obtención de una licencia. Por último, el archivo completo (con o sin identificadores) que podría consultarse en un centro seguro. Al respecto, cabe señalar que la Oficina del Censo de EE. UU. ha creado dos archivos de uso público distintos a partir del conjunto de datos procedentes del censo de 2000, con tasas de muestreo del 1% y el 5%.

En vista de la rápida evolución de las tecnologías informáticas y la creciente accesibilidad de los datos de censos para la comunidad de usuarios, la Oficina del Censo se vio obligada a adoptar medidas más restrictivas para proteger la confidencialidad de los microdatos de uso público y para ello recurrió a técnicas que limitan la divulgación de la información. Al mismo tiempo, la Oficina del Censo es consciente de que los usuarios necesitan datos más detallados y más específicos en el plano geográfico. Así pues, la entidad suministra dos conjuntos de archivos: uno que contiene un número mayor de características detalladas (archivo nacional, con una tasa de muestreo del 1%) y otro que contiene datos geográficos más precisos, pero características menos detalladas (archivos de los estados, con una tasa de muestreo del 5%).

Fuente: http://www.census.gov/population/www/cen2000/pums/index.html
(en inglés), consulta el 8 de abril de 2010.

Los datos pueden consultarse íntegramente en diferentes centros de datos seguros en EE. UU., como el MCRDC de la Universidad de Michigan.

El Centro de Investigaciones sobre el Censo de Michigan (Michigan Census Research Data Center, MCRDC) permite a los investigadores cualificados que trabajan en proyectos aprobados por la Oficina del Censo EE. UU. explotar los datos no publicados recopilados en el marco de los programas económicos y demográficos de la Oficina del Censo, así como por el Centro Nacional de Estadísticas de Salud (NCHS). Todas las investigaciones del MCRDC se llevan a cabo dentro de un laboratorio seguro situado en Ann Arbor, dentro del Instituto de Investigaciones Sociales de la Universidad de Michigan.

Fuente: http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/restricted/index.jsp (en inglés), consulta el 7 de mayo de 2010.

algunos ejemplos, acompañados de la dirección web: el MCRDC (siglas en inglés de Centro de Investigaciones sobre el Censo), un proyecto conjunto de la Oficina del Censo estadounidense y la Universidad de Michigan (www.isr.umich.edu/src/mcrdc/); el *National Opinion Research Center* (NORC) de la Universidad de Chicago (www.norc.org/DataEnclave); el programa CDR de Statistics Canada (www.statcan.gc.ca/rdc-cdr/index-fra.htm); y el *Research Data Center* del Centro Nacional de Estadísticas de Salud de EE. UU. (http://www.cdc.gov/nchs).

1.6 ¿Existen alternativas para difundir los archivos de microdatos?

En los casos mencionados arriba, al usuario se le ofrece un acceso directo a los ficheros de microdatos. Sin embargo, existen otras formas de acceso a los microdatos, tales como el envío de trabajos (scripts) a distancia y el acceso remoto. Los requerimientos específicos de las políticas que rigen este tipo de acceso no se describen aquí, sino que simplemente se mencionan y se incluyen referencias a fuentes de información complementarias. Se debe destacar que estos sistemas generalmente son costosos y técnicamente complejos.

Envío de trabajos (scripts)

Uno de los medios de los que disponen los usuarios para analizar datos confidenciales es crear un procedimiento que les permita enviar programas (o scripts) de tratamiento y análisis de datos confidenciales a distancia a los depositarios de datos. En este caso, el usuario recibe una base de datos sintetizado que reproduce la estructura y el contenido de los originales. En este marco, los investigadores pueden elaborar programas con ayuda de herramientas como SAS, SPSS o Stata. A continuación, estos programas (o scripts) se transmiten al personal de la entidad depositaria de los datos, que ejecuta la aplicación sobre el conjunto de datos original. Los resultados obtenidos se verifican y ajustan antes de transmitirlos al usuario.

A modo de ejemplo, cabe citar el proyecto LIS (*Luxembourg Income Study*), que ofrece la posibilidad de acceder a bases de microdatos a través de un sistema de envío de *scripts* automatizado denominado LISSY (Cuadro 2).

Si bien la confidencialidad de los datos está protegida, dar cobertura técnica a los servicios de envío de trabajos puede ser costoso. Además, si los recursos asignados a estos servicios son insuficientes, el procedimiento podría ser lento desde la perspectiva de los usuarios.

Acceso remoto

Aquí, los usuarios tienen acceso a aplicaciones informáticas vía web de tabulación y análisis de datos, pero no pueden descargar ninguna base de datos ni crear tablas que permitirían revelar información individual o un pequeño número de registros.

En el mercado se pueden encontrar varias aplicaciones informáticas, como Nesstar, Beyond 20/20, SuperCross, Redatam, PcAxis, etc. Algunos centros que se sitúan a la vanguardia de los avances en la materia desarrollan sus propias plataformas. El centro de almacenamiento UKDA gestiona un servicio de datos seguro (*Secure Data Service*, SDS) «cuyo objetivo es fomentar la excelencia en la investigación ofreciendo a los investigadores de buena fe acceso remoto seguro a datos considerados sensibles, detallados, confidenciales o potencialmente identificables con el fin de transmitirlos a través de acuerdos de licencia y difusión estándar». [12]

Este sistema es adecuado para la creación de tablas estadísticas (en particular, para censos de población y vivienda), pero no para un análisis en profundidad.

Contratación de un investigador como empleado temporal

Algunos productores de datos se esfuerzan por que los investigadores tengan acceso a los microdatos contratándolos como empleados temporales. De este modo, están sometidos a las mismas normas de confidencialidad que el personal permanente del organismo. No conviene recurrir a esta fórmula, salvo en el caso de que la actividad del investigador suponga una aportación verdadera al trabajo del productor de datos en cuestión; de lo contrario, podría parecer una simple simulación. [24]

Cuadro 2: *Luxemburg Income Study* – LISSY

LISSY es un sistema remoto de ejecución de trabajos (scripts) totalmente automatizado que funciona 24 horas al día, 7 días a la semana. Permite a los investigadores enviar programas de tratamiento estadístico por lotes (creados con SAS, SPSS o Stata) sin tener que desplazarse. LISSY ejecuta automáticamente las tareas solicitadas y después transmite los resultados de forma sintetizada en cuestión de minutos.

Los archivos de microdatos no pueden descargarse y no se permite ningún acceso directo a los datos. Lo único que se envía a los usuarios son los resultados de las consultas estadísticas.

Inscripción obligatoria

Si bien las cifras clave del sistema LIS (*key figures*) se ponen a disposición del público en general, el acceso a las archivos de microdatos se reserva a los usuarios inscritos. La autorización de acceso se concede únicamente por un periodo de un año, renovable anualmente.

Dos modos de envío de trabajos (o scripts)

LISSY ofrece un acceso remoto seguro a los microdatos a través de dos modos de envío:

- software JSI (*Job Submission Interface*);
- Software de mensajería, como Outlook, Thunderbird, etc.

Los dos sistemas conducen a los mismos resultados, pero la página web de LIS recomienda encarecidamente a los usuarios que recurran a JSI para acceder a LISSY. La interfaz es más fácil de usar y contiene funciones adicionales, como el acceso al historial de trabajos del usuario.

Instrucciones de envío

Con independencia del modo de envío de los trabajos, existen varias características específicas en relación con la sintaxis de programación estándar que deben conocerse para que las consultas de los usuarios puedan ser procesadas correctamente por LISSY.

Las aplicaciones estadísticas actualmente disponibles en el sistema LISSY son SAS 9.2, Stata 11.0 y SPSS 11.5.

Asistencia técnica

Todas las dudas relacionadas con el uso y el contenido de las bases de datos de LIS deben remitirse al departamento de asistencia técnica (*LIS User Support*) en lugar de los empleados de LIS. El objetivo es poder llevar un registro coordinado del conjunto de las dudas planteadas.

Fuente: http://www.lisproject.org/dataaccess/dataaccess.html (consulta el 9 de abril de 2010).

2. ¿Qué son los metadatos?

Los metadatos se definen generalmente como «datos que describen otros datos». En el capítulo anterior se aborda la importancia que reviste la publicación de un diccionario de datos adecuado que describa el contenido de todas las variables comprendidas en un conjunto de datos. Sin embargo, los metadatos de calidad suministran información que va más allá de la que contiene un simple diccionario de datos.

Los *metadatos* sirven para ayudar a los investigadores a **comprender** lo que miden los datos y la forma en que se han obtenido. Sin una buena descripción del diseño de la encuesta y los métodos empleados para recoger y procesar los datos, el usuario corre un riesgo elevado de malinterpretarlos o incluso utilizarlos incorrectamente.

Una documentación adecuada reduce la cantidad de asistencia que los estadísticos deben prestar a los usuarios externos de sus microdatos.

Por otro lado, los metadatos tienen como finalidad ayudar a los usuarios a **evaluar** la calidad de los datos. Es importante que los investigadores que deseen juzgar el interés de determinados datos para su trabajo conozcan las normas en vigor en materia de recogida de datos y estén en disposición de identificar las desviaciones frente a dichas normas.

Por último, los metadatos son necesarios para desarrollar herramientas de **investigación de datos** como los catálogos de encuestas, que ayudan a los investigadores a encontrar conjuntos de datos relevantes a sus intereses.

Conviene señalar que los datos deben documentarse incluso cuando no están destinados a ser difundidos. Producir metadatos de calidad favorece la creación de una «memoria institucional» en materia de recogida de datos, puede contribuir a la formación del personal nuevo y puede mejorar la coherencia de los datos a lo largo del tiempo.

¿Qué son metadatos de calidad?

La descripción de los metadatos de calidad que figura a continuación se ha extraído de la guía *Good Practices in Data Documentation* (Buenas prácticas en materia de documento de datos) publicada por la UKDA. [20] La página web de la Red Internacional de Encuestas de Hogares (www.ihsn.org) y su guía práctica dirigida a los responsables de repositorios de datos (*Quick Reference Guide for Data Archivists*) contienen también información muy interesante a este respecto. [4]

«Un elemento esencial en la creación de un conjunto de datos de calidad y explotable a largo plazo es garantizar que los datos sean fáciles de comprender y analizar. Eso conlleva una descripción y una documentación de los datos amigables con el usuario que sean claras y detalladas, al tiempo que exhaustivas». (http://www.data-archive.ac.uk)

La documentación ideal de un conjunto de datos contiene esencialmente tres tipos de documentos:

1. Documentos explicativos

Estos documentos constituyen el mínimo imprescindible para garantizar la viabilidad y la funcionalidad de los datos a largo plazo; si no existe esta información, resulta imposible comprender correctamente el conjunto de bases de datos y su contenido.

Información sobre los métodos de levantamiento de los datos

Esta parte trata sobre el proceso de recogida de datos: encuesta, compendio de datos administrativos o transcripción de un documento fuente. Normalmente describe los instrumentos utilizados, los métodos empleados y la forma de elaboración de estos últimos. Si procede, se informará sobre el método y el marco de muestreo. La información sobre un eventual sistema de vigilancia de la recogida de datos, así como de control de calidad, es también de grandísima utilidad.

Información sobre la estructura del conjunto de datos

Consiste esencialmente en un documento detallado que describa la estructura del conjunto de bases de datos, incluidas las relaciones entre determinados archivos o elementos de información (registros) que contiene. Generalmente establece las variables clave necesarias para identificar de forma inequívoca los individuos (o unidades de observación) en los diferentes archivos. Normalmente, también se indica el número de casos y variables que contiene cada archivo, así como el número total de archivos que constituyen el conjunto de datos. En el caso de modelos relacionales, debería informarse sobre la estructura y sobre los vínculos entre los registros.

Información de tipo técnico

Esta información se refiere a la infraestructura técnica y debería incluir:

- El sistema informático utilizado para crear los archivos.
- Los programas informáticos utilizados para crear los archivos.
- El medio para el almacenamiento de los datos.
- La lista completa de los archivos contenidos en el conjunto de datos.

Variables y valores, sistemas de codificación y clasificación

Es deseable que en la documentación conste la lista completa de las variables (o de los campos) que aparecen en el conjunto de datos, con una descripción exhaustiva y detalles sobre los sistemas de codificación y clasificación. Es especialmente importante explicar si existen variables sin información o datos perdidos. También resulta útil indicar cuáles son las variables a las que se aplica una nomenclatura estándar, precisando la versión del sistema de clasificación empleado y preferiblemente consignando las referencias bibliográficas correspondientes.

Información sobre las variables derivadas

Numerosos productores de datos crean nuevas variables a partir de datos originales. El método puede ser muy sencillo y materializarse, por ejemplo, en forma de agrupación de datos por rangos edad (en número de años cumplidos) de acuerdo con las clases de edad de la encuesta en cuestión. Se pueden aplicar otros métodos más complejos que recurren a algoritmos mucho más elaborados. Conviene explicar la lógica que preside una agrupación o una derivación de variables. En el diccionario de datos se puede explicitar una simple agrupación de datos por edad. Sin embargo, las derivaciones más complejas deben documentarse de forma separada, preferentemente indicando de forma precisa los organigramas o las expresiones booleanas. La idea es que la información suministrada permita establecer fácilmente el vínculo entre las variables principales y las variables resultantes. Por otro lado, se recomienda que los algoritmos informáticos utilizados para crear las variables derivadas se conserven y consignen junto con la información relativa al software empleado.

Ponderación y extrapolación

Debe suministrarse una lista completa de las variables extrapoladoras y ponderadoras (o factores de expansión) , junto con información sobre el modo de obtención de estas variables y estipulando claramente las condiciones de aplicación de estas. Esto es especialmente importante cuando deben aplicarse varios factores de ponderación en función del objetivo que desea alcanzarse.

Fuente de los datos

Información sobre la fuente de donde los datos se derivan debería ser incluida con detalle. Por ejemplo, en el caso de una fuente de datos formada por respuestas a cuestionarios de encuestas, todas las preguntas deberían estar inventariadas minuciosamente en la documentación. Idealmente, deberían mencionarse las referencias de cada variable generada. Además, resulta interesante explicar las condiciones en las que se formulan las preguntas y, si fuera posible, indicar las situaciones en las que se aplican, preferentemente con una síntesis de las estadísticas de respuesta.

Confidencialidad y anonimización

Es importante precisar si los datos contienen información confidencial sobre los individuos, hogares, organizaciones o instituciones. De ser así, se recomienda mencionar estos datos, así como cualquier contrato que rija las condiciones de uso de dichos datos (por ejemplo, información sobre los participantes de la encuesta). Las cuestiones relativas a la confidencialidad pueden restringir los análisis que pueden efectuados o los resultados a ser publicados, sobre todo si los datos se destinan a un uso secundario. En el caso de que se anonimicen los datos con el fin de proteger la identidad de los encuestados, se deberá precisar el procedimiento utilizado en este marco y su efecto sobre los datos. En la medida en que dichas modificaciones sean susceptibles de limitar los análisis posteriores, resulta útil indicar esta circunstancia.

2. Información contextual

Se trata de suministrar a los usuarios información sobre el contexto del levantamiento de los datos, así como el uso que se ha hecho de ellos. Esta información enriquece y da mayor profundidad a la documentación. Permite a los usuarios secundarios conocer íntegramente el contexto y los procedimientos de recogida de datos.

Lo que es más: constituye un testimonio histórico fundamental para los futuros investigadores.

Descripción del proyecto original

Es deseable que la documentación incluya información sobre la génesis del proyecto o sobre el proceso que originó el conjunto de datos, describiendo el marco intelectual y formal de referencia. Por ejemplo, una descripción podría cubrir temas como:

- Razones que motivaron el levantamiento de datos.
- Metas y objetivos del proyecto.
- Unidad de observación y características (o variables) investigadas.
- Cobertura geográfica y temporal.
- Publicaciones a las que ha contribuido el estudio o políticas que se han definido a partir de dicho estudio.
- Cualquier otra información que se considere pertinente.

Procedencia del conjunto de datos

Se trata de información relativa a aspectos como el origen del proceso de levantamiento de datos, los cambios introducidos y la evolución de los datos propiamente dichos, así como la metodología y los ajustes efectuados. También puede suministrarse la siguiente información:

- Lista de los datos erróneos.
- Problemas encontrados en el momento de la recogida, el ingreso, la crítica y la validación de los datos.
- Conversión realizada para garantizar la compatibilidad con otro software o sistema de explotación.
- Referencias bibliográficas de los informes o publicaciones en los que se apoya el estudio.
- Cualquier otra información sobre el ciclo de vida del conjunto de datos que se considere pertinente.

Bases de datos de datos de panel y series de tiempo, nuevas ediciones

Para conjuntos de datos provenientes de estudios de corte transversal repetidos, panel o series de tiempo, resulta extremadamente útil disponer de información complementaria, por ejemplo sobre la evolución de la formulación de las preguntas, las etiquetas (o nombres) de las variables o de los procedimientos de muestreo.

3. Sistemas de catalogación

Estos sistemas cumplen un doble objetivo. En primer lugar, constituyen una referencia bibliográfica del conjunto de datos, lo que permite identificarlo y citarlo correctamente en publicaciones. También permite registrarlo oficialmente para su conservación a largo plazo. En segundo lugar, es la herramienta básica utilizada para descubrir e investigar recursos. En este sentido, el conjunto de datos puede identificarse frente a otros de forma inequívoca si se precisan las informaciones que ayudarán a los usuarios secundarios a determinar la utilidad del estudio para su actividad.

Si no existen títulos, resúmenes, palabras clave y otros elementos de metadatos importantes, será difícil para los investigadores identificar los conjuntos de datos y las variables que responden a sus necesidades. Todos los sistemas de catalogación e identificación de recursos, ya sean físicos o electrónicos, se basan en metadatos.

«Resulta mucho más sencillo obtener documentación de calidad cuando su composición se planifica desde el comienzo del proyecto y se piensa en ella en todas las etapas de la investigación (a lo largo de todo el ciclo de vida de los datos). Planificar con antelación puede ayudar a reducir sensiblemente los plazos y el presupuesto necesarios para preparar la documentación»

(http://www.data-archive.ac.uk, traducción del inglés). Véase también [21].

Estándares de metadatos y buenas prácticas

«La interoperabilidad tecnológica y semántica es esencial para facilitar y fomentar la accesibilidad y la utilización de los datos de investigación en un contexto internacional e interdisciplinario. Los dispositivos de acceso deberían tener debidamente en cuenta las normas internacionales aplicables en materia de documentación de los datos». [17]

Deseosa de favorecer el intercambio de datos entre las organizaciones y los sistemas informáticos y de mejorar la calidad de la información estadística que se suministra a los usuarios, la comunidad internacional de repositorios de datos ha elaborado un conjunto de normas sobre metadatos. Estas normas ofrecen un marco estructurado para la organización y difusión de

información sobre el contenido y la estructura de los datos estadísticos.

ISO 11179 – Tecnologías de la información – Registros de metadatos (RM)

La norma internacional ISO/CEI 11179-1 ha sido creada por el comité técnico JTC 1 de ISO (tecnologías de la información), subcomité SC 32 (servicios de gestión de datos).

«La norma ISO/CEI 11179 define la estandarización y el registro de los datos con el fin de facilitar su comprensión e intercambio. La estandarización y el registro de datos que se describen en la norma ISO/CEI 11179 permiten crear un entorno de datos que se comparten con mayor rapidez y facilidad que con los métodos de gestión de datos clásicos». [9] Algunas organizaciones han basado el diseño de sus bases de datos de conceptos y definiciones en la norma ISO 11179. No obstante, conviene señalar que dicha norma no suministra ninguna herramienta de documentación y difusión práctica, contrariamente a las normas que figuran a continuación, basadas en el lenguaje XML.

Data Documentation Initiative (DDI)

Tradicionalmente, los productores de datos redactaban los libros de códigos en formato texto. Con el fin de explotar lo máximo posible la tecnología de Internet, la mayoría de las normas se definen ahora en lenguaje XML. La especificación *Data Documentation Initiative* (DDI) es un estándar para la documentación de los microdatos. [13]

La iniciativa DDI desarrolló estándares que ofrecen un marco estructurado para organizar los contenidos, la presentación, la difusión y la conservación de metadatos en ciencias sociales y del comportamiento. Este marco permite documentar los archivos de microdatos más complejos de forma sencilla, pero rigurosa.

La iniciativa DDI busca establecer un estándar internacional basado en el lenguaje XML para la documentación de los microdatos. Su objetivo es proponer medios prácticos para el registro y la comunicación de los aspectos esenciales de los conjuntos de microdatos. La especificación DDI es una gran evolución del libro electrónico de códigos tradicional: en esencia, ofrece las mismas posibilidades, pero mejora ostensiblemente la extensión y el rigor de la información suministrada. La especificación DDI para metadatos fue lanzada por el ICPSR (Consorcio Interuniversitario para la Investigación en Ciencias Políticas y Sociales), una organización formada por más de 500 universidades y centros de enseñanza superior de todo el mundo. Ahora, el proyecto está en manos de una asociación de instituciones norteamericanas y europeas entre cuyos miembros se encuentran algunos de los productores y repositorios de datos más importantes del planeta.

La especificación DDI ha sido diseñada para abarcar todos los tipos de datos procedentes de encuestas, censos, archivos administrativos, experimentos, observaciones directas y otras metodologías sistemáticas utilizadas para obtener mediciones empíricas. Dicho de otra forma, la unidad de análisis puede estar constituida por individuos, hogares, familias, empresas, transacciones, países o incluso otros sujetos de interés científico. Del mismo modo, las observaciones pueden consistir en mediciones efectuadas de forma puntual en el tiempo (durante una semana, dentro de una muestra representativa de un país, por ejemplo). También puede tratarse de observaciones repetidas realizadas en el tiempo y en múltiples circunstancias (datos longitudinales y transversales de diferentes países, así como datos cronológicos y agregados). La especificación DDI contiene asimismo una descripción exhaustiva de la metodología del estudio (forma de levantamiento de los datos, métodos de muestreo, si aplica, universo de estudio, zonas geográficas cubiertas, organismo y personas responsables, etc.).

Estructura

La especificación DDI permite describir en detalle todos los aspectos de una encuesta: metodología, responsabilidades, archivos y variables. Brinda una lista estructurada y exhaustiva de varios cientos de elementos y atributos susceptibles de constituir la documentación de un conjunto de datos. Aunque es poco probable que en ella figuren todos los elementos, algunos de ellos, como el título (*Title*), son obligatorios e imperativamente inequívocos. Otros elementos, como el autor y el investigador principal (*Authoring*

Cuadro 3: ¿Quién utiliza la norma DDI?

El estándar de metadatos DDI se utiliza en una extensa comunidad de repositorios de datos, sobre todo bibliotecas de datos de universidades, gestores de datos de INEs y otros productores de datos oficiales, así como en organizaciones internacionales.

Ejemplos de usuarios en el medio universitario:

- DataFirst, Universidad del Cabo (www.datafirst.uct.ac.za).
- UKDA, Universidad de Essex (www.data-archive.ac.uk).
- ICPSR, Universidad de Michigan (www.icpsr.umich.edu).
- Universidades canadienses que participan en el programa sobre control de la divulgación de estadísticas de Statistics Canada (http://www.statcan.gc.ca/rdc-cdr/network-reseau-eng.htm).
- Red DataVerse del Centro de Datos Harvard-MIT y de la biblioteca de la Universidad de Harvard (http://thedata.org).
- Agencias miembros de la CESSDA (http://www.cessda.org).

Productores de datos oficiales de más de 50 países, tales como:

- Statistics Canada, a través de su Iniciativa para la Democratización de los Datos (IDD) (http://www.statcan.gc.ca/dli-ild/dli-idd-fra.htm).
- Instituto Nacional de Estadística de Bolivia (http://www.ine.gob.bo/anda).
- Oficina Nacional de Estadística, Departamento de Estadísticas Laborales y de Empleo (Labour and Employment Statistics) de Filipinas (www.census.gov.ph, www.bas.gov.ph, www.bles.dole.gov.ph)
- Departamento de Censos y Estadísticas de Sri Lanka (http://statistics.sltidc.lk).
- Agencia Estadística Central de Etiopía (www.csa.gov.et).
- Y muchas otras (véase www.ihsn.org/adp).

Organizaciones internacionales:

- UNICEF, para su metodología de encuestas de indicadores múltiples (MICS) (http://www.childinfo.org/mics3_surveys.html).
- Banco Mundial (http://data.worldbank.org).
- Global Fund (http://www.theglobalfund.org/html/5YEdata).

La adopción del estándar de metadatos DDI se ha visto enormemente facilitada por la existencia de programas informáticos fáciles de usar, como DDI Nesstar Publisher, y otras herramientas de catalogación compatibles con DDI suministradas por la IHSN (véase www.ihsn.org/toolkit y www.ihsn.org/nada).

Entity y *Primary Investigator*) son opcionales y pueden aparecer varias veces, ya que hacen referencia a la persona o personas y/o la entidad o entidades responsables de la encuesta.

Los elementos del estándar DDI (versión 2.4) se organizan en cinco secciones:

Sección 1.0: Descripción del documento

El organismo que se encarga de la documentación y la difusión del estudio (encuesta, censo u otro) no siempre es el productor de los datos. Por consiguiente, es importante suministrar información (metadatos) no solo sobre el estudio propiamente dicho, sino también sobre el proceso de documentación. La sección «Descripción del documento» es una presentación general que describe el documento XML en formato DDI, es decir, «metadatos de los metadatos».

Sección 2.0: Descripción del estudio

Esta sección consiste en una presentación general la operación estadística. En ella se indica la referencia de la encuesta y la identidad de la organización o institución que ha recogido, compilado o difundido los datos. También resume (sintetiza) el contenido de los datos, informa sobre los métodos de levantamiento, procesamiento, etc.

Sección 3.0: Descripción de los archivos de datos

Esta sección describe el contenido de cada archivo de datos, indica el número de registros y variables, consigna la versión, el nombre del productor, etc.

Sección 4.0: Descripción de las variables

Esta sección presenta detalles de cada variable, como: la formulación de las preguntas, el universo de la pregunta, las etiquettas de la variable y sus valores, los métodos de derivación e imputación, etc.

Sección 5.0: Otros documentos

Esta sección permite incluir la descripción de otros documentos relativos al estudio, por ejemplo: recursos documentales (cuestionarios, notas sobre la codificación, informes técnicos y analíticos, manual del entrevistador, etc.), programas de procesamiento y análisis de los datos, fotografías y mapas.

Norma de metadatos Dublin Core (DCMI)

El contenido de esta sección fue extraído de la página web de la DCMI (*Dublin Core Metadata Initiative*): http://dublincore.org (únicamente en inglés).

El conjunto de los elementos de metadatos Dublin Core (norma ISO 15836), también denominado estándar de metadatos Dublin Core, permite describir recursos digitales. Esta norma resulta especialmente útil para describir los recursos sobre los que se sustentan los microdatos: cuestionarios, informes, manuales, *scripts* y programas de procesamiento de los datos, etc. La idea de elaborar esta norma nació en 1995, bajo los auspicios del OCLC (*Online Computer Library Center*) y del NCSA (*National Center for Supercomputing Applications*), reunidos en Dublin (Ohio). Progresivamente, Dublin Core se ha convertido en la norma más extendida para la descripción de recursos digitales en Internet. Se elevó al rango de estándar ISO en 2003. La DCMI garantiza su mantenimiento y trabaja en el desarrollo del estándar. Esta organización internacional tiene como objetivo promover la creación de estándares de metadatos interoperables.

El éxito de la norma de metadatos Dublin Core se debe principalmente a su simplicidad. Desde el comienzo, sus creadores se esforzaron por que el conjunto de elementos fuera lo más resumido y simple posible, para que así el estándar pudiera ser utilizado por no especialistas. El objetivo de esta norma es favorecer la creación cómoda y barata de notas descriptivas simples de los materiales de catalogación, permitiendo al mismo tiempo investigar eficazmente estos mismos recursos en Internet o en cualquier otro entorno en red. En su forma más simple, la norma Dublin Core está formada por 15 elementos de metadatos, todos ellos opcionales y utilizables varias veces. Estos 15 elementos son:

• Título	• Relación	• Derechos
• Tema	• Cobertura	• Fecha
• Descripción	• Creador	• Formato
• Tipo	• Editor	• Identificador
• Fuente	• Colaborador	• Idioma

Cuadro 4: ¿Quién utiliza el estándar Dublin Core?

Dublin Core es un estándar de metadatos extremadamente sencilla y flexible. No tiene el grado de detalle del estándar DDI (sobre todo, DC no contiene elementos relacionados con información sobre los archivos de datos y las variables), pero puede servir para describir las características generales de un conjunto de datos. Además, ofrece una alternativa que permite describir los conjuntos de datos cuyos recursos no pueden ser objeto de una documentación detallada. La Open Data Initiative de EE. UU. utiliza una versión adaptada de Dublin Core (www.Data.gov).

Cuadro 5: El lenguaje XML

XML son las siglas de eXtensible Markup Language. Este lenguaje ofrece una forma de etiquetar un texto con ayuda de etiquetas que determinan el significado y no solo la presentación. Dicho de otro modo, XML permite estructurar un texto con ayuda de etiquetas que contienen indicadores semánticos. Conceptualmente, en términos de organización estas etiquetas son idénticas a los campos de una base de datos. A diferencia de los archivos de bases de datos, los archivos XML pueden visualizarse y modificarse con ayuda de un editor de texto estándar. En estos archivos pueden realizarse búsquedas de la misma forma que en una base de datos clásica, con ayuda de las herramientas apropiadas. Del mismo modo que el contenido de una base de datos puede servir para crear un informe, los documentos XML pueden publicarse y convertirse con ayuda de otras aplicaciones informáticas para obtener formatos más sencillos (hojas de cálculo, PDF o páginas web).

El ejemplo que figura debajo permite hacerse una idea de la forma en que la información textual sobre una encuesta puede presentarse en formato XML:

Considere la siguiente información: Entre enero y marzo de 2005, la oficina nacional de estadística (ONE) de Popistán llevó a cabo una encuesta de indicadores múltiples (MICS) financiada por UNICEF. Se seleccionaron aleatoriamente un total de 5.000 hogares representativos de la población del país para participar en la encuesta, de acuerdo con un plan de muestreo en dos niveles. De ellos, 4.900 suministraron información.

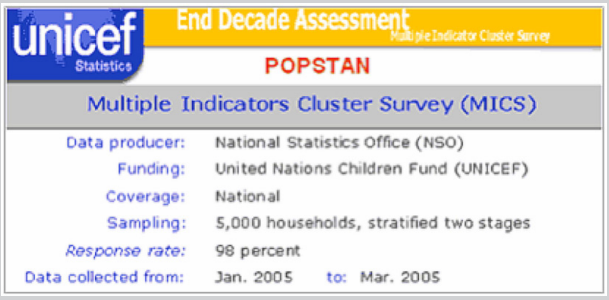
Esto mismo convertido al formato DDI-XML tendría el aspecto siguiente:

<titl>Encuesta de indicadores múltiples 2005</titl>
<altTitl>MICS</altTitl>
<AuthEnty>Oficina Nacional de Estadística (ONE)</AuthEnty>
<fundAg abbr="UNICEF">Fondo de las Naciones Unidas para la Infancia</fundAg>
<collDate date="2005-01" event="start"/>
<collDate date="2005-03" event="end"/>
<nation>Popistán</nation>
<geogCover>Nacional</geogCover>
<sampProc>5.000 hogares, estratificados en dos niveles</sampProc>
<respRate>98 por ciento</respRate>

El uso de las etiquetas es tanto más eficaz cuando la comunidad de usuarios adopta un conjunto común de etiquetas (estándar DDI o Dublin Core, por ejemplo). La adopción de un conjunto común de etiquetas XML ofrece grandes ventajas a la hora de documentar microdatos:

- Creación de una lista exhaustiva de los elementos de metadatos utilizados.
- Posibilidad de evaluar el contenido de un archivo verificando si contiene o no ciertas etiquetas.
- Creación de un catálogo para los conjuntos de datos que permita investigar elementos de metadatos claves.
- Posibilidad de convertir el fichero en formatos más accesibles.

Los ficheros XML pueden convertirse en HTML, PDF u otros tipos de documentos usando transformaciones XSL. Además pueden intercambiarse entre diferentes redes o en Internet a través de servicios web o el protocolo SOAP (basado en el lenguaje XML y que permite el intercambio de información entre aplicaciones a través de HTTP). La página HTML que figura debajo es un ejemplo de transformación XSL de la versión en inglés del documento XML anterior



3. ¿Cuáles son los argumentos a favor de la difusión de microdatos?

La difusión es una de las principales responsabilidades de un instituto de estadística. Este capítulo resume las numerosas ventajas que brinda la difusión de microdatos. Una revisión sobre la misión y visión, las políticas de difusión y la experiencia de varios productores de datos de todo el mundo subraya la importancia y los argumentos a favor del acceso a los archivos de microdatos.

3.1 Apoyar la investigación

El motivo principal (y a veces el único motivo explícito) que lleva a los productores de datos a compartir sus microdatos es el apoyo a la investigación. Al final de una encuesta, los responsables del levantamiento de datos generalmente producen una serie de tablas destinadas a presentar los aspectos esenciales y a ofrecer una visión general de los resultados a los usuarios. No es lógico esperar de estas agencias que identifiquen todos los temas de investigación que estos datos pueden contribuir a tratar, y tampoco disponen de presupuesto para ello. Los archivos de microdatos ofrecen a los investigadores una flexibilidad considerable para identificar relaciones e interacciones entre los fenómenos cubiertos por una encuesta, lo que favorece y promueve la diversidad y la calidad de los trabajos de investigación.

A continuación se ofrecen algunos ejemplos que ilustran la forma en que algunos institutos nacionales de estadística describen el objetivo de sus políticas de difusión de datos.

> «El acceso a los microdatos permite y favorece una toma de decisiones informada mediante un uso amplio de los datos de la ABS [*Australian Bureau of Statistics*] con fines de análisis y de investigación económica y social. Desde 1985, la ABS suministra microdatos en forma de archivos de registros individuales confidenciales (CURF) en determinadas condiciones y con fines estadísticos. Actualmente, la ABS recibe numerosas peticiones de acceso a registros individuales más detallados de formas más sencillas y en un abanico de conjuntos de datos más amplio (datos de empresas y conjuntos de datos longitudinales vinculados, por ejemplo). La imposibilidad de satisfacer estas demandas tendrá un efecto cada vez más negativo sobre la razón de ser de la ABS, su relevancia y, en última instancia, sobre la coherencia del sistema estadístico nacional. Esta evolución, ligada a una serie de factores adicionales que justifican un cambio, tales como el riesgo creciente de identificación, han llevado a la ABS a proponer una nueva estrategia de acceso a sus microdatos en el futuro» (Oficina de Estadísticas de Australia, http://www.abs.gov.au)

> «El principal objetivo de la CSO [*Central Statistics Office*] en materia de acceso a los microdatos es dar apoyo a la comunidad investigadora y garantizar que los datos que recojamos pueden ser explotados al máximo. Este enfoque promueve la elaboración de políticas fundamentadas en evidencias y, además, puede contribuir a reducir los costes de investigación y a evitar las duplicidades en el levantamiento de datos». (Oficina Central de Estadística de Irlanda, http://www.cso.ie) (Traducción del inglés)

> «El objetivo del archivo nacional de microdatos de Sri Lanka es responder a las necesidades de la comunidad investigadora nacional e internacional, que se esfuerza por encontrar respuestas a los problemas socioeconómicos que se plantean en el mundo. Con ayuda del *Depart*amento de Censos y Estadísticas de Sri Lanka, LankaDatta difunde datos estadísticos pertinentes, fiables y actuales producidos por las agencias del sistema estadístico nacional, preservando la confidencialidad de los encuestados». (Department of Census and Statistics, Sri Lanka http://statistics.sltidc.lk)

> «La IDD [Initiativa de Democratización de los Datos] es un ejemplo claro de aprovechamiento de las tecnologías de la información en Canadá, ya que permite a los centros de enseñanza superior ofrecer por primera vez una completa gama de servicios de datos tanto a los alumnos como a los profesores. Parece cada vez más evidente que esta iniciativa está contribuyendo decisivamente a la enseñanza y a la investigación en Canadá. (...) La IDD ha contribuido a instaurar una cultura de los datos en Canadá» (Statistique Canada, http://www.statcan.gc.ca/dli-ild/about-apropos-fra.htm, traducción del francés; véase también Watkins [31]).

3.2 Reforzar la credibilidad de las estadísticas oficiales

Ofrecer un acceso más amplio a los microdatos da muestra de la confianza de los productores en sus datos, en la medida en que autorizan la reproducción de los cómputos (indicadores) o su rectificación por parte de organismos independientes.

3.3 Mejorar la fiabilidad y la pertinencia de los datos

Crear una relación más estrecha entre los productores de datos y los usuarios informados puede generar otras ventajas. Con mucha frecuencia, es el propio uso de los datos el que permite hacerse una idea de las posibles mejoras, sobre todo en el plano del diseño de las encuestas y la difusión de los microdatos. El proceso de difusión de los microdatos puede comprender un procedimiento formal de retroalimentación por parte de los usuarios, como ocurre con la Oficina del Censo estadounidense. Los aportes de los usuarios pueden llevar a una mejora progresiva de las encuestas.

3.4 Reducir las duplicidades en los datos

l hecho de que se pongan archivos de microdatos a disposición de los usuarios les disuade a menudo de recabar ellos mismos los datos que necesitan. Así, se recurre menos veces a los encuestados y se reduce el riesgo de obtener estudios incoherentes sobre un mismo tema.

3.5 Incrementar la rentabilidad de la inversión

«El acceso abierto a los datos de las investigaciones financiadas con fondos públicos y su difusión contribuyen no solo a reforzar el impacto de las nuevas tecnologías y las nuevas redes digitales en el potencial de la investigación, sino que también permite rentabilizar mejor la inversión pública en investigación. [...]

«Los investigadores y los centros de investigación financiados con fondos públicos recogen volúmenes cada vez mayores de datos. Este conjunto de datos de investigación que no para de crecer constituye al mismo tiempo una enorme inversión de fondos públicos y una de las fuentes potenciales de los conocimientos necesarios para afrontar los múltiples desafíos a los que se enfrenta la humanidad. […]

Para elevar el rendimiento científico y social de las inversiones públicas en datos de investigación, los países miembros de la OCDE, [por ejemplo], han promulgado diferentes leyes, políticas y prácticas relacionadas con el acceso a los datos de investigación a escala nacional. En este contexto, contar con directrices internacionales constituirá una ganancia importante para fomentar los intercambios y el uso a escala mundial de los datos de investigación». [17]

3.6 Recaudar fondos para estudios estadísticos

Cuanto más se difundan y exploten los archivos de datos, más valor tendrán a los ojos de los donantes.

Este argumento puede animar a los auspiciantes a financiar el levantamiento de datos. Algunos donantes exigen, en efecto, pruebas de uso de los datos.

Un mejor uso de los datos se traduce en una mayor rentabilidad de la inversión para los auspiciantes, que estarán más dispuestos a sostener las actividades de levantamiento de datos. Cada vez con más frecuencia, la financiación de encuestas por parte de los organismos internacionales está condicionada a la difusión de los conjuntos de datos obtenidos.

3.7 Reducir los costes de la difusión de los datos

Por último, los organismos encargados del levantamiento de datos podrían beneficiarse de una mayor eficiencia, en el sentido de que podrían reducir la cantidad de tablas predefinidas que producen y dedicar más esfuerzos a la elaboración análisis más profundos. Hacer hincapié en este aspecto (sobre todo en los medios de comunicación y en el sector de la enseñanza) puede despertar el interés de un público más amplio y favorecer la financiación de las actividades de los organismos que recaban datos, ya sea las ONE u otro tipo de entidades. A menudo, estas tablas resultan ser insuficientes para llevar a cabo análisis en profundidad. Por último, es preciso colocar la cuestión de la eficiencia en el contexto de los costes de producción y difusión de los archivos de microdatos. Estas cuestiones se abordan en la siguiente sección.

3.8 Respetar las obligaciones legales y contractuales

En algunos países, los organismos públicos tienen la obligación de difundir una parte de sus microdatos.

El levantamiento de datos a menudo está financiado por el contribuyente a través de los impuestos y, por lo tanto, se considera de dominio público. En otros, los donantes son los organismos auspiciantes, que exigen que los datos obtenidos se pongan a disposición de los investigadores (véase Cuadro 6). Esta obligación de difusión de los microdatos no entra en contradicción con la obligación de respetar la confidencialidad y la privacidad. El máximo responsable estadístico de cada organismo, o eventualmente el comité de comunicación y difusión, debe decidir el contenido de los microdatos que se publicarán, así como los procedimientos que se emplearán para crear archivos de uso público.

3.9 Fomentar el desarrollo de nuevas herramientas para el uso de los datos

Desde hace algunos años, se está extendiendo un nuevo movimiento denominado «datos abiertos». El concepto fundamental de esta tendencia es que los datos recogidos gracias a fondos públicos o bajo la supervisión de un organismo estatal son de dominio público.

Varios gobiernos se han sumado a este movimiento: véase la *Open Government Initiative* en EE. UU. (www.data.gov) o su homóloga británica (http://data.gov.uk). Este tipo de tendencia anima al público en general a que aporte valor a los datos existentes. Al abrir el acceso a los microdatos sin restricción, estas iniciativas fomentan el desarrollo de nuevas herramientas informáticas, sobre todo aplicaciones innovadoras para la Web 2.0.

Estas aplicaciones web innovadoras, que utilizan datos abiertos, a menudo se conocen con el nombre de *mashups* (o aplicaciones web híbridas):

«En el ámbito del desarrollo de páginas web, un *mashup* es una aplicación o una página web que utiliza o combina datos o funciones de varias fuentes externas con el fin de crear un nuevo servicio. Eso implica una integración fácil y rápida, generalmente a través de API (aplication programming interface) y utilizando fuentes de datos abiertas, con el objetivo de generar resultados enriquecedores que no se correspondan necesariamente con la motivación que impulsó la producción de los datos brutos iniciales.

«Con el fin de poder acceder permanentemente a los datos de otros servicios, los *mashups* se diseñan generalmente como aplicaciones clientes o alojadas en línea. Los *mashups* desempeñan sin duda un papel activo en la evolución de las aplicaciones informáticas sociales y de la Web 2.0 (véase también http://fr.wikipedia.org/wiki/application_composite, en francés ó también en inglés http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)

Con el fin de favorecer el desarrollo de estas aplicaciones web híbridas, los productores de datos no se limitan a difundir datos, sino que también suministran con frecuencia aplicaciones informáticas denominadas *interfaces de programación* (API). Estas aplicaciones permiten a los programadores utilizar los datos más fácilmente.

«Una interfaz de programación de aplicaciones (*Application Programming Interface* o *API*) es una interfaz suministrada por un programa informático. Hace posible la interacción de los programas, de forma análoga a una interfaz hombre-máquina, que hace posible la interacción entre las personas y las máquinas». Las API son implementadas por una aplicación, por una biblioteca de software o por un sistema operativo y permiten determinar la sintaxis y las convenciones de llamada que el programador debe respetar para utilizar los servicios que se ofrecen. Pueden contener especificaciones relativas a las rutinas de ejecución, las estructuras de datos, las clases de objetos y los protocolos de comunicación entre el usuario y el componente que implementa la API (http://fr.wikipedia.org/wiki/Interface_de_programmation, en francés, consulta el 9 de enero de 2011; o también disponible en inglés:

http://en.wikipedia.org/wiki/Application_programming_interface; véase también el Cuadro 7).

Cuadro 6: Obligación legal de difusión de los microdatos: el ejemplo del Centro Nacional de Estadísticas de Salud de EE. UU. (NCHS)

El Centro Nacional de Estadísticas de Salud de EE. UU. (NCHS), adscrito a los Centros de Control y Prevención de Enfermedades (US Centers for Disease Control and Prevention), nos ofrece un ejemplo de legislación sobre difusión de microdatos.

«Por su condición de instituto de estadística federal, el NCHS debe hacer todo lo que esté en su mano para potenciar la disponibilidad de los datos, sobre todo acortar el intervalo entre el levantamiento y la difusión de los datos, con el fin de optimizar su calidad, pero también para reducir al mínimo el riesgo de divulgación de información privada. La legislación que rige las actividades del NCHS estipula que los datos deben difundirse con la mayor amplitud posible [...].

«No obstante, la obligación de difusión de los datos debe entenderse desde la condición de instituto de estadística federal del NCHS y estar en consonancia con la necesidad de garantizar la protección de la confidencialidad de los encuestados y la calidad de los datos. […]

«La misma ley que obliga al NCHS a difundir sus datos también le impone la obligación de proteger la identidad de las personas y organizaciones que aparecen en sus sistemas de datos». [14]

Cuadro 7: Fomentar las aplicaciones web híbridas difundiendo datos abiertos y API

EE. UU.: data.gov – Descubrir, participar y comunicar

Uno de los objetivos subyacentes de la Open Government Initiative es modificar los hábitos de difusión de la información institucionalizando una comunicación más amplia de los datos federales en formatos más accesibles. Como la punta de lanza de la Open Government Initiative, la página web Data.gov está destinada a facilitar el acceso a los conjuntos de datos federales que ayuden al público en general a comprender mejor el funcionamiento de las agencias federales y sus actividades, que ponen de relieve sus objetivos, generan oportunidades económicas y que mejoran la transparencia y la apertura del gobierno federal (eso es, conjuntos de datos de alto valor añadido).

Felicitaciones a los programadores

Estamos muy sorprendidos por el número de descargas registrado y las aplicaciones innovadoras para integrar datos de la administración que han aparecido desde el lanzamiento de Data.gov. Uno de los principales objetivos de Data.gov es favorecer la innovación. A este respecto, nuestra comunidad de programadores ha destacado especialmente. Animamos a todos los desarrolladores y programadores a explorar los conjuntos de datos recogidos en Data.gov y a participar activamente en esta comunidad dinámica en pleno crecimiento.

Fuente: http://www.data.gov/open, en inglés, consulta el 3 de abril de 2010

Reino Unido: data.gov.uk – La llave de la innovación – Enséñanos el camino...

Sabemos que fuera del ámbito de la administración existen numerosas personas como usted que poseen las competencias y las capacidades necesarias para sacar el máximo partido a los datos públicos. Ese es para nosotros el punto de partida de la relación de colaboración que deseamos establecer con usted.

Fuente: http://data.gov.uk, en inglés, consulta el 3 de abril de 2010

La iniciativa de datos abiertos gira en torno a la idea de accesibilidad. Con el lanzamiento de donnees.banquemondiale.org, estamos intensificando los esfuerzos del Banco Mundial para abrir sus catálogos de datos por medio de un acceso fácil y directo a través de Internet. Durante 2010, vamos a desplegar sucesivamente dos conjuntos de funciones web que servirán como plataforma para este esfuerzo. […]

La fase 2 se centrará en la promoción del uso de datos del Banco Mundial por medio de la API en Internet. Para ello será necesario realizar mejoras en la API actual, pero también crear un lugar para las comunicaciones en torno a la API y un umbral de entrada más bajo, sobre todo para los usuarios que no son programadores informáticos.

Esta fase marcará el inicio de una evolución que hará que la API de la web deje de ser una interfaz destinada exclusivamente a programadores informáticos para convertirse en una interfaz orientada tanto a los investigadores y responsables políticos como a los desarrolladores informáticos.

Fuente: http://donnees.banquemondiale.org/developpeurs, en francés, consulta el 7 de mayo de 2010.

4. ¿Cuáles son los costes y los riesgos asociados a la difusión de microdatos y cómo se pueden controlar?

Los INEs y otros productores y recopiladores de datos deben tener en cuenta una serie de factores a la hora de definir y poner en práctica sus políticas y sus programas de difusión de microdatos: costes y competencias necesarias, cuestiones relativas a la calidad de los datos, posibilidad de una explotación abusiva o de una mala interpretación de los datos por parte de los usuarios, aspectos jurídicos y éticos, sin olvidar la preservación de la confianza y la protección de los encuestados.

4.1 Cuestiones éticas y preservación de la confianza de los encuestados

Cuando las oficinas de estadística y otros productores de datos recogen información de los individuos, de entidades o de otras organizaciones, generalmente garantizan a los encuestados que la información facilitada será utilizada únicamente con fines estadísticos. Se trata de una obligación moral y ética.

«Los INEs deben conservar la confianza de los encuestados para que estos sigan colaborando en la recogida de datos. La protección de la confidencialidad es la clave de esta confianza. Si los encuestados piensan o sienten que el INE no es capaz de garantizar la confidencialidad de los datos, se mostrarán menos dispuestos a cooperar y a proporcionar información precisa. Incluso un incidente aislado, sobre todo si se acompaña de una intensa cobertura mediática, puede tener un impacto significativo en la colaboración de los encuestados y, en consecuencia, en la calidad de las estadísticas oficiales. La protección de la confidencialidad es la preocupación principal de los INEs, pero no la única. Los INEs ponen también especial empeño en conseguir la autoridad suficiente, bien sea por medio de disposiciones legales o de cualquier otra forma de autorización, para poder facilitar el acceso de los investigadores a los microdatos». [24] Para obtener el máximo provecho en la utilización de los microdatos, las ONE y otros productores de datos tienen que encontrar un equilibrio entre el cumplimiento de las exigencias de confidencialidad y la concesión de derechos de acceso. Una de las soluciones consiste en utilizar diferentes tipos de microdatos (véanse capítulos anteriores). Existe otra opción, que, sin embargo, no puede aplicarse en todas las situaciones: conseguir que cada encuestado apruebe formalmente la comunicación de los datos recogidos.

Obtener el consentimiento de los individuos

«El consentimiento puede ser expreso o tácito. En el primer caso, se entrega al encuestado un documento escrito que describe las condiciones de uso de la información que se le solicita y se le invita a dar su consentimiento mediante la firma de dicho documento. No siempre se informa al encuestado por escrito, sino que, a veces, solo se lo hace verbalmente. Si el encuestado accede a proporcionar la información solicitada, el encuestador «deduce» que está dando su consentimiento para la utilización y la comunicación de datos previstas por las partes, que constan por escrito o que se mencionan verbalmente. El encuestador estará, por tanto, autorizado a utilizar los datos, exclusivamente de acuerdo con las condiciones descritas al encuestado». [15]

Obtener el consentimiento de personas jurídicas

«En el caso de personas jurídicas, el método varía dependiendo de si la información se solicita en persona o por correo.

1. A. Si la información la solicita en persona un empleado o un representante del encuestador, estos deben, primeramente, establecer quién está autorizado o quién es la contraparte para proporcionar dicha información de la entidad. Cuando esta persona responde a las preguntas, tras haber sido informada sobre la utilización que se hará de los datos recibidos, el representante [del organismo encuestador] interpreta que la entidad da su consentimiento.

2. B. Si los datos se recogen por correo, la solicitud podrá estar dirigida a la propia entidad, a su administrador, o a cualquier otra persona a la que el organismo encuestador haya identificado previamente como autorizada para proporcionar la información solicitada. La carta de solicitud de información deberá precisar la utilización que se hará de los datos. Tras

la recepción de los datos, el personal del organismo encuestador considera que la entidad aprueba su utilización conforme a las condiciones previamente indicadas». [15]

4.2 Aspectos jurídicos

¿Está un productor de datos legalmente capacitado para difundir archivos de microdatos? No existe una respuesta única a esta pregunta. La legislación que regula las actividades de los productores de datos es específica de cada país y de cada programa marco (Cuadro 8).

Como ya hemos mencionado, la difusión de microdatos es, en ocasiones, una obligación legal. No obstante, en la mayoría de los casos la legislación prevé simples restricciones. Por lo tanto, la política de cada país en materia de difusión de microdatos vendrá determinada por su marco legislativo. Es primordial que los productores de datos «se aseguren de que disponen de una sólida base jurídica y ética (así como de las herramientas técnicas y metodológicas necesarias) para proteger la confidencialidad. Esta base jurídica y ética requiere que haya un cuidado equilibrio entre el interés público en una sólida protección de la confidencialidad, por una parte, y los beneficios que aporta la investigación, por otra. La decisión de saber si conviene o no permitir a un investigador tener acceso a los datos dependerá del valor de su proyecto de investigación y de su credibilidad, y habrá que tenerlo en cuenta, de una manera o de otra, en las disposiciones legales». [24]

«Las modalidades de acceso a los datos deberían respetar los derechos e intereses legítimos de todos los intervinientes en la actividad de investigación pública. El acceso a ciertos datos de investigación y su utilización deberán estar necesariamente limitados por diferentes tipos de disposiciones legales, que pueden imponer restricciones por motivos de:

- Seguridad nacional: ciertos datos relativos a la inteligencia, a las actividades militares o a la toma de decisiones políticas pueden estar clasificados y, por tanto, estar sujetos a un acceso limitado.

- Protección de la privacidad y de la confidencialidad: los datos relativos a personas y otros datos personales están sujetos a un acceso limitado de conformidad con las leyes y políticas nacionales en materia de protección de la confidencialidad y la privacidad. Conviene, sin embargo, que los propietarios de estos datos tengan en cuenta los procedimientos de anonimización o de confidencialidad que permitan asegurar un nivel de confidencialidad apropiado con el fin de preservar, en la medida de lo posible, la utilización de los datos por parte de los investigadores.

- Secretos comerciales y derechos de propiedad intelectual: los datos concernientes a empresas u otras entidades, o provenientes de estas, que contengan información confidencial podrían no ser accesibles con fines de investigación. [...]

La adhesión a códigos profesionales de conducta facilita el cumplimiento de las disposiciones legales». [17]

Principios fundamentales de las Naciones Unidas sobre Estadística Oficial

Numerosos países se basan en los *Principios Fundamentales de las Naciones Unidas sobre Estadística Oficial* para elaborar sus leyes. Así pues, es interesante examinar estos principios en lo referente a la confidencialidad de las estadísticas.

El sexto principio fundamental de las Naciones Unidas sobre estadística oficial establece lo siguiente: «Los datos individuales que recogen los organismos responsables para la elaboración de estadísticas, ya sean referidos a personas físicas o jurídicas, deberán ser estrictamente confidenciales y se utilizarán exclusivamente con fines estadísticos». [27]

Cualquier principio relativo al acceso a los microdatos debe ser coherente con este principio fundamental o con las disposiciones legales aplicables a las ONE. A la hora de tratar la confidencialidad de los microdatos, deben tenerse en cuenta los principios que figuran a continuación.

Principio 1: Utilización apropiada de los microdatos.

«Los microdatos recogidos para elaborar estadísticas oficiales pueden utilizarse en análisis estadísticos para apoyar investigaciones, siempre y cuando se proteja la confidencialidad de dichos datos. [...]

La comunicación de los microdatos a los investigadores no está en contradicción con el sexto principio fundamental de la ONU, siempre y cuando

no sea posible identificar los datos que se refieren a un individuo. El mencionado Principio 1 no constituye una obligación de difundir los microdatos. Es potestad de la ONE decidir si conviene o no proporcionar los microdatos. Puede haber otras consideraciones, por ejemplo la calidad, que hagan que resulte inadecuado proporcionar acceso a los microdatos. O puede haber ciertas personas o instituciones a quienes no sería conveniente proporcionar los microdatos». [24]

Principio 2: Los microdatos podrán ser proporcionados únicamente para fines estadísticos.

«Según el Principio 2, conviene hacer la distinción entre la utilización de los datos con fines estadísticos o analíticos y la utilización administrativa de los mismos. En el primer caso, el objetivo consiste en establecer estadísticas relativas a un grupo (que pueden ser personas físicas o jurídicas). Por el contrario, el objetivo en el uso administrativo de los datos es obtener información sobre una persona física o jurídica para tomar una decisión que pueda beneficiar o perjudicar a un individuo, como por ejemplo, la solicitud de datos por orden judicial. Para asegurar la confianza del público en el sistema estadístico oficial, estas solicitudes, aunque sean legales, son contrarias a este principio y deberían ser rechazadas sistemáticamente.

Si la utilización de microdatos es incompatible con fines estadísticos o analíticos, no debería permitirse el acceso a los microdatos. Los comités de ética —o cualquier sistema similar— pueden prestar su colaboración cuando exista duda acerca del acceso a los microdatos.

Los investigadores acceden a los microdatos con fines de investigación. Aquí se incluye la elaboración de agregados estadísticos de diversa naturaleza, la creación de distribuciones estadísticas, el ajuste de modelos estadísticos o el análisis de las diferencias estadísticas entre subpoblaciones. Estos usos son compatibles con los fines estadísticos. En este sentido, los microdatos se consideran útiles para reforzar la investigación». [24] (Traducción del inglés)

Principio 3: La entrega de los microdatos debería realizarse de conformidad con las disposiciones jurídicas y de otra índole que garanticen la confidencialidad de los microdatos proporcionados.

«En relación al Principio 3, sería muy recomendable aplicar las disposiciones jurídicas dirigidas a proteger la confidencialidad antes de publicar cualquier microdato. No obstante, las disposiciones legales deben completarse con medidas administrativas y técnicas que regulen el acceso a los microdatos y que garanticen que los datos personales no puedan divulgarse. La existencia y visibilidad de tales medidas —ya sean de orden jurídico o estén incluidas en normativas complementarias, ordenanzas, etc.— son indispensables para aumentar la confianza del público en el uso adecuado que se hará de los microdatos. Obviamente, es preferible la introducción de disposiciones legales. En el país donde no fuera posible, convendría establecer cualquier otra disposición administrativa. Sería igualmente necesario que, allí donde ya existan, las autoridades competentes en materia de confidencialidad validaran dichas disposiciones legales —u otras— antes de que entraran en vigor. Si no hubiera autoridad en esta materia, algunas ONG pueden ejercer una función de vigilancia en cuestión de confidencialidad. Sería conveniente conseguir el apoyo de dichas organizaciones a cualquier disposición jurídica o a otras medidas, o que, al menos, indiquen las inquietudes importantes que puedan tener.

No todos los países cuentan con una base legislativa. Como mínimo, conviene que la divulgación de los microdatos esté amparada por alguna autoridad. Es preferible, sin embargo, constituir una legislación ». [24]

Principio 4: Es necesario garantizar la transparencia en los procesos de acceso de los investigadores a los microdatos, así como los usos y los usuarios de los microdatos, y hacerlos públicos.

«El Principio 4 es importante para que el público tenga confianza en que los microdatos están siendo utilizados con buen criterio y para demostrar que las decisiones que afectan a la difusión de microdatos se toman de acuerdo con una base objetiva. La ONE debe determinar si los microdatos pueden ser divulgados, en qué condiciones y a qué usuario. No obstante, conviene garantizar la transparencia de sus decisiones. La página web de la ONE constituye un instrumento eficaz para garantizar el cumplimiento de las normas establecidas e informar sobre cómo acceder a los informes de los estudios basados en los microdatos proporcionados». [24]

4.3 Exposición a la crítica y a la contradicción

«A algunas ONE les preocupa que la calidad de sus microdatos no sea suficiente y no permita difundirlos más ampliamente. Aunque se compruebe que la calidad es suficiente para permitir la elaboración de agregados estadísticos, a veces esta puede ser insuficiente para un análisis detallado. En algunos casos, se realizan ajustes a los agregados estadísticos en el momento de la validación de resultados, sin que los microdatos sean editados. En consecuencia, pueden aparecer incoherencias entre los resultados de investigaciones basadas en microdatos y las que se basan en datos agrupados publicados» [24]. Los grupos de datos que no se consideren suficientemente fiables pueden suprimirse antes de la difusión. Conviene que los productores de datos sean claros y transparentes en lo que a la calidad de los datos se refiere.

Por otra parte, la comunicación de los microdatos a los investigadores puede conllevar la posibilidad de que sus resultados publicados no se ajusten a las estimaciones del organismo que ha producido los datos. Si se trata de un organismo estadístico oficial, esto puede acarrear un conflicto entre las estimaciones contradictorias (oficiales y no oficiales) que ponga en tela de juicio los datos, o que incluso tenga repercusiones políticas. Existen varios factores que pueden causar tales diferencias. En primer lugar, es posible que las estimaciones oficiales sean erróneas, en cuyo caso sería recomendable un control externo. En segundo lugar, las diferencias pueden estar relacionadas con el uso de diferentes versiones de los datos (archivo principal completo o versión pública anonimizada/restringida, otros cambios realizados por el investigador, etc.). Sin embargo, estas diferencias deberían ser mínimas y fácilmente justificables.

Por último, las diferencias pueden deberse a disparidades metodológicas. Esto es, a menudo, más problemático para los productores de datos, en la medida en que el público no siempre es capaz de comprender explicaciones técnicas complejas. Es importante que los productores de datos sepan defender sus estimaciones. Esto implica que el levantamiento, el procesamiento y el análisis de datos estén documentados de principio a fin y que, además, esta información sea de fácil acceso. En ocasiones, los resultados publicados se elaboran por o con ayuda de expertos externos que ya no están disponibles para responder a las preguntas que surgen. Los productores de datos pueden evitar este riesgo adoptando y poniendo en práctica sistemas rigurosos

Cuadro 8: Ejemplos de legislación en materia de confidencialidad

A continuación se exponen algunos ejemplos de leyes en materia de estadística y de difusión de microdatos:

1. La actividad de la Oficina del Censo de EE. UU. se rige por las disposiciones de la sección novena del título XIII del Código de los Estados Unidos. «En virtud de las disposiciones de la sección novena del título XIII del Código de los Estados Unidos, está prohibido publicar o difundir cualquier información que permita identificar a una entidad, a un individuo o a un hogar concretos. El control de divulgación incluye medidas que garanticen que los datos mencionados en el título XIII sean preparados, antes de su publicación, para protegerlos contra una divulgación indebida. Estas medidas incluyen métodos que limitan la divulgación y un procedimiento de control dirigido a garantizar que los métodos empleados ofrecen una protección adecuada de los datos».

Fuente: http://www.census.gov/srd/sdc/wendy.drb.faq.pdf.

2. La Ley sobre Estadística de Canadá estipula lo siguiente: «Ninguna persona que haya prestado juramento en virtud del Artículo 6 podrá revelar ni hará revelar conscientemente, por ningún medio, la información obtenida en virtud de la presente ley, de manera que, a causa de dicha revelación, puedan relacionarse los datos obtenidos con un particular, con una empresa o con una entidad».

Fuente: http://www.statcan.gc.ca/about-aperçu/act-loi-fra.htm.

Statistics Canada difunde archivos de microdatos. Su política sobre difusión de microdatos establece que la publicación de los ficheros de microdatos está sujeta a las siguientes condiciones:

(a) la difusión debe mejorar sensiblemente el valor analítico de los datos recogidos; y

(b) el organismo estadístico debe tomar todas las medidas posibles para impedir la identificación de las unidades de observación de la encuesta.

3. En Tailandia, la Ley contiene las siguientes disposiciones: «Los datos personales obtenidos en virtud de esta ley serán tratados con la más estricta confidencialidad. Toda persona que tenga una obligación conforme a la presente o sea responsable de la conservación de este tipo de información deberá abstenerse de revelarla a cualquier persona que no tenga ninguna obligación en virtud de esta ley, salvo en los siguientes casos:

(1) Que la información se divulgue en el curso de una investigación o procedimiento judicial por un delito en virtud de la presente.

(2) Que la información divulgada se utilice para la preparación y el análisis de estadísticas o estudios estadísticos, siempre que tal divulgación no perjudique a los titulares de los datos y preserve la identidad de los mismos». (Traducción del inglés)

Fuente: http://web.nso.go.th/eng/en/about/about0.htm, sección 15.

de documentación y conservación, de acuerdo con las normas de reproducción de datos. Básicamente, estas normas se describen de la siguiente manera: «[...] La única forma de entender y evaluar un análisis empírico es conocer con exactitud el proceso de creación y de análisis de los datos. [...] Las normas de reproducción prevén que, para que un tercero pueda replicar los resultados, la cantidad de información disponible sea suficiente para comprender, evaluar y hacer referencia a trabajos anteriores sin necesidad de disponer de ninguna información del autor» (véase también [11]).

4.4 Costes

«A los INEs les puede preocupar también la cuestión de los costes, no solo aquellos derivados de la creación y documentación de los archivos de microdatos, sino también los relacionados con el desarrollo de herramientas de acceso y de protección, y aquellos derivados del apoyo y la autorización de encuestas solicitadas por la comunidad investigadora; de hecho, los nuevos usuarios de sus archivos de datos necesitan ayuda para orientarse por las estructuras de archivos complejos y las definiciones de variables. Aunque dichos costes son asumidos por los INEs, por lo general estas no tienen una partida presupuestaria para financiar los trabajos adicionales que se lleven a cabo en este contexto. En lo que respecta a los investigadores, por lo general no disponen de los medios para asumir una parte considerable de estos costes» [24]. Por lo tanto, en la medida de lo posible estos costes deben incluirse en el presupuesto de la encuesta y deben servir para obtener el máximo aprovechamiento de los resultados. Es de interés general que las conclusiones extraídas de los datos se comuniquen con el fin de informar a los responsables políticos y al público. Además, si los datos de la encuesta se utilizan más ampliamente, proporcionarán una barrera adicional contra la reducción de las partidas presupuestarias asignadas a los programas estadísticos. Las encuestas que ofrecen poca ayuda a la hora de elaborar políticas públicas son las que corren un mayor riesgo.

4.5 Pérdida de la exclusividad

Al divulgar los microdatos, sus propietarios pierden su derecho exclusivo de consulta de los datos. Este aspecto es más problemático para los investigadores académicos que para los productores de datos oficiales, aunque estos últimos (o algunos miembros de su personal) a veces aprovechan su acceso exclusivo a los datos para ofrecer servicios de consultoría. Cada vez más a menudo, los auspiciantes de encuestas establecen que el productor tenga acceso exclusivo a los datos durante un tiempo «razonable». Al finalizar ese periodo, los datos deben ser accesibles para otros usuarios.

4.6 Capacidad técnica

Se requieren algunos recursos técnicos para poder garantizar la difusión de archivos de microdatos. Los archivos deben ir acompañados de la documentación detallada (preferiblemente conforme a los estándares de metadatos DDI) y deben conservarse adecuadamente. Además, deben examinarse para evaluar el riesgo de divulgación de información personal y la reducción de ese riesgo a través de la utilización de diferentes herramientas. Los requisitos técnicos relativos a la difusión de archivos de microdatos se describen con más detalle en el capítulo 10.

5. ¿A quién están destinados los microdatos?

Los archivos de microdatos están destinados a los especialistas dotados de competencias cuantitativas consolidadas. Ejemplos:

- Responsables políticos e investigadores empleados por la administración y los servicios de planificación.
- Organizaciones internacionales y otros organismos promotores.
- Institutos universitarios y centros de investigación que desarrollen estudios económicos y sociales.
- Personal universitario y estudiantes.
- Otros usuarios que participen en la investigación científica.

Uno de los fundamentos de la difusión de microdatos es la *equidad*. Cuando la difusión es jurídicamente posible, y según el principio de libre acceso, los microdatos obtenidos a través de procesos de recogida financiados con fondos públicos deben ponerse a disposición de todos los usuarios potenciales, acompañados de los metadatos más completos posibles. «Este libre acceso debe otorgarse en igualdad de condiciones a la comunidad científica internacional, al más bajo coste posible y preferiblemente sin superar el coste marginal de la difusión. El acceso libre a los datos de la investigación financiada con fondos públicos debe ser cómodo, rápido, sencillo y preferiblemente a través de Internet». [17]

Los INEs suelen proporcionar diferentes productos dirigidos a un público muy diverso. Los resúmenes de datos (tablas, gráficos, análisis) están dirigidos a un público más amplio y se publican en las páginas web de los organismos. Generalmente, los archivos de microdatos están destinados a los investigadores de diferentes instituciones (por ejemplo, organismos gubernamentales y ministerios, ONG, centros de investigación, universidades y organizaciones internacionales) a las que comúnmente se denomina «comunidad investigadora».

«¿Qué es la comunidad investigadora? Obviamente, incluye a aquellos que trabajan en los centros universitarios, pero también a los investigadores que realizan su labor para las organizaciones no gubernamentales (ONG) y para las organizaciones internacionales. Asimismo, algunos investigadores que solicitan el acceso a los microdatos trabajan en organismos e instituciones financiadas por el gobierno». [24]

Existen numerosas leyes en materia de estadística que establecen que los datos deben estar disponibles para fines estadísticos legítimos y para respaldar la investigación. El uso comercial dentro de un marco contractual se considera generalmente incompatible con esta norma. Enfatizar la utilización potencial de los microdatos es un aspecto fundamental a la hora de cumplir las leyes que rigen la actividad de las ONE.

Eurostat, la oficina estadística de la Unión Europea, define el acceso a los microdatos de la siguiente manera:

«Los microdatos se ponen fundamentalmente a disposición de la comunidad científica, los centros de investigación o las universidades únicamente con fines de investigación y previa presentación de un proyecto de investigación. En casos muy excepcionales, el acceso se hace extensible a estudiantes y a un público no científico más amplio». Los conjuntos de microdatos anonimizados puestos a disposición del público se denominan archivos de microdatos de uso público. Aunque los organismos de estadística los difunden amplia y gratuitamente, tienen un interés y valor limitados para la actividad pública. «Por norma general, las empresas privadas no tienen permitido trabajar con microdatos».

La difusión de los microdatos se centra principalmente en la comunidad investigadora. Sin embargo, existen otras consideraciones asociadas. Así pues, en ocasiones se hace referencia a los usuarios *fiables*, es decir, aquellos investigadores que actúan de buena fe.

En realidad, los investigadores no se convierten en «propietarios» de los archivos de datos, sino que se les otorga una licencia de uso intransferible. Esto permite a los INEs averiguar el propósito de la investigación. Es recomendable que el INE elabore una política que le permita establecer las condiciones de acceso a sus datos.

Conviene que esta política sea lo más general posible y que contenga información para todos los tipos de usuarios. Se pueden aplicar condiciones específicas a determinados usuarios, en función de su nacionalidad o del organismo al que pertenezcan.

Ejemplos:

Usuarios nacionales:

- Los funcionarios que estén expresamente sometidos a la legislación que rige las estadísticas se pueden beneficiar del acceso a los microdatos oficiales, puesto que están sujetos a la misma normativa que los INEs. Se les puede tomar juramento fácilmente y sancionarlos si fuera necesario, lo que sirve a los objetivos del gobierno.

- Otros usuarios nacionales cuya actividad esté regulada por la legislación sobre estadísticas tienen la posibilidad de utilizar los archivos de datos, siempre y cuando firmen el pertinente contrato de licencia de uso.

Usuarios internacionales:

- En numerosos casos, es obligatorio comunicar los datos a las organizaciones internacionales. Esta obligación puede estar relacionada con la pertenencia del organismo a un grupo internacional o con contratos de financiación de proyectos importantes, o incluso con la participación del país en proyectos de desarrollo internacional.

- La concesión de licencias de uso a los investigadores que trabajan para las universidades o los centros de investigación extranjeros es más complicada. En esos casos es más difícil hacer respetar las disposiciones aplicables. No obstante, puede resultar interesante compartir los datos con estos investigadores, ya que poseen una valiosa experiencia. Se puede reducir el riesgo otorgando la licencia a un centro —por ejemplo, a una universidad— en lugar de a una persona en particular (véase la información complementaria que se detalla más adelante).

Asistentes técnicos:

El personal de los INEs recibe a menudo la colaboración de consultores, extranjeros o no, para el procesamiento y/o análisis de los datos de encuestas. Puesto que esta costumbre es compatible con los objetivos del INE, se recomienda hacer mención de la misma en la política de difusión de datos, a condición de que los consultores firmen el mismo contrato que los otros investigadores. Las cláusulas de ese contrato deberán establecer que los datos no pueden ser comunicados sin el consentimiento previo de sus productores. Conviene igualmente exigir la firma de una declaración de confidencialidad.

El INE puede hacer que los investigadores asuman parte de los riesgos tomando las siguientes medidas:

(i) «Solicitarles que demuestren su legitimidad como investigadores, el interés público de sus investigaciones y que aporten pruebas de que la utilización de los microdatos es necesaria.

(ii) Procurar que los investigadores firmen un compromiso legalmente vinculante que contemple sanciones, semejantes a las que se aplican al personal del INE, si infringen la normativa en materia de confidencialidad.

(iii) Explicar las razones por las que el INE toma estas precauciones. Asegurarse de que los investigadores son plenamente conscientes de sus obligaciones informándoles al respecto convenientemente. Establecer métodos eficaces de seguimiento y de supervisión. Sería de utilidad crear un código de conducta en colaboración con la comunidad investigadora.

(iv) Si se cometiera alguna infracción, dejar de prestar servicio durante un tiempo al investigador y, eventualmente, al organismo que representa (por ejemplo, hasta que dicho organismo haya impuesto medidas disciplinarias al investigador infractor. Es muy importante que entiendan que las críticas públicas podrían poner en peligro la divulgación futura de microdatos a la comunidad investigadora. Emprender acciones legales si fuera oportuno». [24]

Cuando un investigador presenta una solicitud de acceso en nombre de una entidad, sería recomendable que pudiera hacerlo únicamente en nombre de dicha entidad (o de su organismo, empresa, etc.), y no en su propio nombre. Estas organizaciones tienen una reputación que proteger y, por lo tanto, es más fácil hacer que respeten los compromisos adquiridos. Es más difícil hacer que se cumplan estos contratos cuando la solicitud de acceso se realiza desde otro país. Para solucionar este problema, se puede colaborar con los repositorios de datos o el INE del país solicitante.

Cuadro 9: Ejemplo de declaración de confidencialidad

Por la presente, me comprometo a:

1. No hacer ninguna copia parcial o total de los archivos a los que pudiera tener acceso por autorización del personal del centro nacional de datos seguro (en lo sucesivo el Centro) y a no sacar del Centro ninguna información o datos personales consultados u obtenidos en virtud de mi puesto de investigador en este Centro.
2. Devolver al personal del Centro la totalidad de documentos confidenciales que me sean confiados durante mi periodo de investigación en el Centro, así como todo documento solicitado.
3. No intentar identificar a ninguna persona, organización o unidad de muestreo cuya identidad no sea mencionada en los archivos de datos de uso público.
4. Preservar la más estricta confidencialidad sobre la identidad de organizaciones o de personas descubierta inintencionadamente en cualquier documento, durante una conversación o como parte de un análisis. Cualquier descubrimiento involuntario de identidades en el transcurso de mi trabajo de análisis será inmediatamente comunicado al personal del Centro.
5. No retirar copias impresas, ni archivos electrónicos, ni cualquier otro documento en cualquier soporte que no haya sido previamente examinado por el personal del Centro para eliminar el posible riesgo de divulgación. Entiendo que el Centro llevará a cabo un control de divulgación y que la retirada de cualquier dato, ya esté en formato electrónico o en papel, estará sujeta a la autorización previa del Centro.
6. No retirar ninguna información manuscrita sobre la identidad de organizaciones o personas o sobre coordenadas geográficas establecidas en el transcurso de mi investigación en el Centro.
7. Adoptar un comportamiento conforme a los principios y normas de buena conducta acordes a un centro de investigación científica.

He sido informado de que el incumplimiento deliberado de cualquiera de estas condiciones anteriormente descritas conllevará la anulación del permiso de acceso a los datos. Asimismo, entiendo que una infracción de este tipo podría suponer la prohibición permanente de acceso al Centro siempre que el director lo considere necesario para proteger la integridad de la confidencialidad del Centro.

_____ _____

Nombre y firma del investigador Fecha

_____ _____

Nombre y firma del testigo Fecha

Esta declaración de confidencialidad se basa en la versión 2008 del contrato que regula las condiciones de acceso a los datos del Centro de Datos de Investigación del Centro Nacional de Estadísticas de Salud de EE. UU. (NCHS) (Acuerdo sobre las condiciones de acceso a datos confidenciales en el Centro de Investigación del Centro Nacional de Estadísticas de Salud de EE. UU.) [16]

6. ¿Qué condiciones deben cumplirse para la difusión de microdatos?

La mayoría de los productores de datos financiados con fondos públicos tienen la misión de difundir lo más ampliamente posible los datos que recogen. No obstante, todos estos organismos están sometidos a obligaciones morales y legales que imponen modalidades de difusión restringidas. «Es primordial difundir al máximo las estadísticas útiles, pero es también importante hacerlo de un modo que no perjudique a los organismos que las han producido» [15] . «Habrá [pues] que establecer un conjunto de medidas legales, administrativas y técnicas para ganarse la confianza del público y mantenerla». [24]

Las condiciones para la difusión de los microdatos deben ser formales y transparentes. Idealmente, se deberían definir tanto a través de una política como de notas de procedimientos o protocolos. Este capítulo contiene información sobre la formulación de esas políticas y esos procedimientos. De todos modos, cabe señalar que la información aquí incluida no constituye una norma recomendada, sino que corresponde a cada productor de datos definir su política y sus procedimientos propios basándose en consideraciones tanto técnicas como legales y éticas.

- Normalmente, una política de difusión contiene disposiciones de carácter general, en especial:

- El objetivo de la política, cuya finalidad es «definir qué tipo de archivos de microdatos se destinarán a la difusión, para qué fines podrán utilizarse y en qué condiciones se publicarán».

- Una breve declaración de principios, a través de la cual la organización explica por qué considera importante difundir microdatos (véase el capítulo 3).

- La legislación vigente en materia de difusión de microdatos (véase la sección 6.1).

- Los grandes principios que guían la preservación de la vida privada y de la confidencialidad de los datos.

- Los plazos de publicación de los datos. Ejemplo: «Consciente de la importancia de satisfacer las necesidades de los usuarios y de proporcionar datos recientes, el INE se esforzará por difundir los archivos de microdatos en un plazo de 6 a 12 meses a partir de la primera publicación de los datos de la encuesta en cuestión. La publicación de los datos de encuestas podría efectuarse en varias fases, de modo que el INE pueda revisarlos y analizarlos antes de anonimizarlos con miras a un posterior uso por parte de otra organización» (véase http://www.surveynetwork.org/home/index.php?q=tools/dissemination/policy, en inglés).

- Las principales funciones y responsabilidades relativas a la definición e implementación de la política de difusión y de los correspondientes procedimientos. Ejemplos:

Directores de encuestas, con las siguientes atribuciones:

- Identificar las necesidades de los principales interesados (o contrapartes) y velar por la creación de un archivo anonimizado que permita responder a las necesidades de la comunidad de usuarios respetando la legislación estadística.

- Iniciar un primer control del archivo de microdatos, identificar los posibles problemas y presentar una versión preliminar del mismo al comité de difusión de los microdatos.

Comité de difusión de los microdatos, con las siguientes atribuciones:

- Examinar todas las solicitudes de difusión de archivos de microdatos anonimizados presentadas por los directores de encuesta de acuerdo con los criterios preestablecidos.

- Aprobar los archivos para su difusión o indicar a los directores de encuestas cómo podrían mejorarse.

- Supervisar la concesión de licencias y dar solución a los riesgos derivados de posibles infracciones.

- Actualizar las directrices destinadas a los directores de encuestas para la creación de archivos de microdatos anonimizados.

Comisión de comunicación y difusión, con las siguientes atribuciones:

- Examinar las solicitudes de acceso de los investigadores a archivos de datos protegidos mediante licencia.
- Proporcionar el acceso a los archivos de datos a los usuarios autorizados.
- Responder a las peticiones de asistencia e información complementaria de los usuarios.

El **director o responsable general de la organización** valida toda entrega de archivos de microdatos anonimizados a los usuarios basándose en los consejos y las recomendaciones del comité de difusión.

- Una descripción general de la política tarifaria de la organización. Idealmente, esta política de precios favorecerá un uso lo más amplio posible de los datos producidos haciéndolos asequibles. Los productores de datos deberían procurar, pues, que los costes vinculados a la creación de archivos de microdatos anonimizados sean considerados en el presupuesto de las encuestas. En paralelo, el INE podrá legítimamente intentar recuperar los costes vinculados a los servicios especiales prestados exclusivamente a una categoría concreta de usuarios.

En los protocolos y las notas de procedimientos se ofrece información más detallada sobre los procedimientos y las condiciones de difusión de los archivos de microdatos, en especial:

- Modalidades de solicitud de acceso a los datos (solicitudes en línea, formularios de solicitud utilización, etc.).
- Autorizaciones y restricciones para los distintos tipos de conjuntos de datos (véase la sección 6.2).
- Entidad responsable de conceder autorizaciones de acceso y de facilitar información práctica sobre los procedimientos de control.
- Métodos de control de la divulgación de estadísticas.
- Información que deben aportar los investigadores y posible uso de dicha información.
- Política tarifaria detallada.
- Tipo y grado de asistencia técnica disponible para los usuarios.
- Otra información práctica.

La UK Statistics Authority ha publicado una guía de buenas prácticas en materia de estadísticas oficiales [22]. Aunque no trata específicamente de la difusión de archivos de microdatos, constituye un buen ejemplo de principios y protocolos formulados con gran claridad.

6.1 Base legislativa

Se precisa una base legislativa por varios motivos:

«(i) Para lograr la confianza del público en las disposiciones establecidas o, dicho de otro modo, para que el público sepa que existen limitaciones jurídicas que determinan lo que está autorizado y lo que no.

(ii) Para que los INEs y los investigadores comprendan mutuamente sus posiciones en lo relativo a estas disposiciones.

(iii) Para garantizar una mayor coherencia en el tratamiento de los proyectos de investigación.

(iv) Para sentar las bases de un sistema de sanciones en caso de infracción». [24]

Algunos INEs no pueden ampararse en ninguna disposición legal sobre estadísticas para la difusión de los archivos de microdatos. Hay legislaciones que prohíben expresamente su difusión, y otras no tratan de forma explícita este tema, con lo cual la difusión queda a criterio de las propias organizaciones. Esto ocurre sobre todo cuando la legislación se remonta a un periodo en el que ni se contemplaba una disposición de este tipo. En tales casos, la difusión de los archivos de microdatos puede depender de una revisión previa de la legislación en vigor. La Ley de Estadística canadiense, por ejemplo, data de principios del siglo XX, pero fue revisada en 1971, entre otras cosas para facultar al INE a difundir archivos de microdatos.

No obstante, «no es indispensable que las disposiciones queden reflejadas en una ley. Los detalles

[...] pueden precisarse más fácilmente en reglamentos, decretos, etc., que tienen también efecto legal. A falta de disposiciones legales, es esencial obtener una forma u otra de autorización, ya que el prestigio del INE puede quedar en entredicho si no existe ninguna forma de autoridad sobre la difusión de los microdatos, aunque sean anonimizados. Es importante que la legislación (o la autorización) tenga en cuenta los siguientes aspectos:

(i) Qué se autoriza y para qué fines, y qué no se autoriza.

(ii) Las condiciones de divulgación.

(iii) Las consecuencias de no respetar dichas condiciones». [24]

6.2 Condiciones aplicables a los AMUP

En principio, los datos considerados de dominio público los puede consultar cualquier persona que tenga acceso al sitio web de un INE. No obstante, se recomienda precisar las condiciones de uso de esos datos, así como las precauciones que hay que tomar al utilizarlos. Aunque este tipo de disposiciones no siempre tienen fuerza legal, permiten sensibilizar al usuario sobre los temas que tratan. Así, en las «condiciones de uso» puede estipularse que se prohíbe cualquier intento de cotejar la información con otras fuentes y hacer que el usuario tenga que aceptar esas condiciones en línea para poder descargar los datos deseados.

En el sitio web de Statistics Canada hay un ejemplo de archivo de microdatos de uso público (AMUP). Es fruto de la *Encuesta conjunta de salud Canadá/Estados Unidos* (véase también la *Encuesta Social Europea*). Para acceder a los archivos de microdatos, los usuarios solo tienen que registrarse. El objetivo es que puedan estar informados de posibles modificaciones en los archivos existentes y de la disponibilidad de nuevos archivos de datos.

La difusión de los archivos de microdatos supone el cumplimiento de determinados principios y reglas. En el cuadro 10 se indican los principios básicos aplicables a los AMUP.

6.3 Condiciones aplicables a los archivos protegidos mediante licencia

Las condiciones generales relativas a los **archivos de microdatos protegidos mediante licencia** deben retomar los grandes principios básicos, complementados con disposiciones aplicables al organismo al que pertenezcan los investigadores. Puede haber dos variantes: primera, que se proporcionen datos a un investigador o equipo de investigadores con unos objetivos específicos; y segunda, que se proporcionen datos a una organización para que los use internamente de acuerdo con un contrato marco (por ejemplo, una organización internacional o un centro de investigación). En ambos casos deberá mencionarse el organismo al que pertenece el investigador y sus representantes deberán firmar la licencia de uso.

Acceso concedido a un investigador o equipo de investigadores con unos objetivos específicos

Cuando se publican datos con vistas a un proyecto concreto de investigación, el equipo de investigadores debe ser identificado. Para ello, debe presentar una solicitud formal de acceso a los datos (véase modelo en el Anexo 1). En las condiciones de obtención de los datos (véase un ejemplo en el Cuadro 12) se estipulará que los archivos no deberán entregarse a personas ajenas a la organización y que deberán almacenarse de forma segura. En la medida de lo posible, se hará constar la finalidad del uso de los datos, así como una la lista de los resultados previstos y la política de difusión de la organización. El acceso a conjuntos de datos protegidos mediante licencia se reserva a las entidades reconocidas legalmente como auspiciantes (ministerio competente, universidad, centro de investigación, organización nacional o internacional).

Contrato marco suscrito con una organización

En este caso se acuerda que los datos estarán reservados para uso interno en la organización, para los fines que ésta considere útiles, y que deberán adoptarse las medidas de seguridad oportunas. Deberá garantizarse la conformidad con los principios en vigor y se nombrará a una persona formalmente responsable. Cada usuario debe ser informado de las condiciones generales aplicables a los archivos de datos y podrá exigirse un compromiso de confidencialidad. Si existen un acuerdo de este tipo y un sistema de seguridad apropiado, no se exigirá la destrucción de los datos tras su uso. En el Cuadro 13 se incluye un modelo de contrato marco.

6.4 Condiciones específicas para centros de datos seguros

Los **centros de datos seguros** se crean para los datos especialmente sensibles o muy detallados, cuya

Cuadro 10: Condiciones de acceso y de uso aplicables a los AMUP

1. Los datos y otros documentos proporcionados por el INE no pueden, bajo ningún concepto, revenderse o cederse a terceras personas, instituciones u organizaciones sin el consentimiento por escrito del INE.
2. Los datos se utilizarán exclusivamente con fines estadísticos o en el marco de trabajos de investigación científica. Servirán solo para crear datos agregados, incluyendo la modelización de información, y queda prohibida cualquier investigación sobre particulares u organizaciones individuales.
3. Queda prohibido cualquier intento de reidentificación de los encuestados. Es más, no se utilizará la identidad de ninguna persona u organización que se descubra fortuitamente y cualquier descubrimiento de este tipo deberá comunicarse de inmediato a la ONE.
4. En ningún caso se intentarán cotejar conjuntos de datos distribuidos por el INE, ni datos del INE con otros conjuntos de datos que pudiesen dar lugar a la identificación de particulares u organizaciones.
5. Los libros, artículos, documentos de trabajo, tesis, disertaciones, informes o cualquier otra publicación que se base en los datos obtenidos del INE deberán citar su fuente, de conformidad con la obligación de cita asociada al conjunto de datos proporcionado.
6. Se enviará al INE una copia en formato digital de todas las publicaciones basadas en los datos solicitados.
7. El recopilador inicial de los datos, el INE y los organismos financiadores correspondientes se deslindan de toda responsabilidad en cuanto al uso y la interpretación de los datos o en cuanto a las conclusiones que se deriven de ellos.

Observaciones:

- Los puntos 3 y 6 implican que los usuarios deben tener la posibilidad de comunicarse fácilmente con el productor de los datos. Es recomendable indicar un número de teléfono o una dirección de correo electrónico y, si es posible, ofrecer un sistema de retroalimentación en línea.
- Para el punto 5, véase el Cuadro 11.

anonimización no puede ser suficiente para permitir un uso público.

También se emplean los términos «laboratorio de datos» o «centro de datos de investigación». Un centro de datos seguros puede estar situado en la sede del INE o en lugares relevantes como universidades en contacto con la comunidad investigadora. Permiten a los investigadores el acceso a archivos de datos completos sin el riesgo de difusión de información confidencial. Por regla general, en estos centros el personal del INE supervisa el acceso a los datos y su uso, los ordenadores no deben estar conectados a una red externa y es obligatorio que un analista de la ONE controle los resultados de los investigadores para garantizar que se respeta la confidencialidad de los datos. En el Anexo 2 se ofrece un modelo de política de acceso a un centro de datos seguros. El Anexo 3 incluye un formulario tipo de solicitud de acceso a un centro de datos seguros.

Los centros de datos seguros tienen la ventaja de que ofrecen el acceso a microdatos detallados, pero también el inconveniente de que obligan a los investigadores a desplazarse de su lugar de trabajo habitual.

Además, los costes de creación y explotación de un centro de este tipo son elevados. Aun así, varios países han optado por que los investigadores puedan acceder in situ a los microdatos. Se trata de investigadores a los que se toma juramento conforme a las leyes estadísticas, al igual que los empleados permanentes de los INEs. Este sistema tiende a favorecer a los investigadores que viven cerca del INE.

6.5 Gestión de las infracciones de los investigadores

La experiencia de los países avezados en la práctica de difundir microdatos muestra que los casos de violación de la confidencialidad de los archivos de datos son excepcionales. A los investigadores no les interesa cometer este tipo de infracciones, puesto que ponen en peligro su reputación y la del organismo para el que trabajan. No obstante, sean intencionadas o accidentales, hay que señalar que existen.

Cuadro 11: Cómo citar un archivo de datos digita l

No existen reglas universales para citar conjuntos de microdatos. Los elementos indispensables que deben incluir una referencia escrita son: nombre del productor de los datos, título del conjunto de datos y año de referencia (seguidos de una indicación especificando que se trata de microdatos), número de referencia (idealmente con el número de versión incluido), nombre del distribuidor de los datos y fecha de obtención de los archivos (para más información sobre la cita de productos estadísticos, véase también [18]).

Ejemplo 1

Departamento de Comercio de los Estados Unidos, Oficina del censo. AMERICAN COMMUNITY SURVEY (ACS): MUESTRA DE MICRODATOS DE USO PÚBLICO (MMUP), 2005 [archivo digital]. ICPSR04587-v1.Washington, D.C.: Departamento de Comercio de los Estados Unidos, Oficina del Censo [productor], 2005. Ann Arbor, Institute of Massachussets: Consorcio Interuniversitario para la Investigación en Ciencias Políticas y Sociales (ICPSR) [distribuidor], 08-08-2007.

Ejemplo 2

Archivo de microdatos de uso público (AMUP): Encuesta sobre la dinámica del trabajo y la renta, 8º sondeo, 2000: Familias económicas [archivo digital] / Canadá, Statistique Canada, Departamento de Encuestas a Hogares, versión 2, Ottawa, Ontario: Statistics Canada [productor]; Statistics Canada. Iniciativa para la Democratización de los Datos [distribuidor], 28-08-2003.

Estas referencias, si bien válidas, no cumplen todas las exigencias y no serán satisfactorias para todos los centros de datos universitarios. «Pueden indicarse URL, generalmente no se mantienen. […] Las versiones revisadas de los datos suelen difundirse con el mismo nombre y sin una indicación estándar del número de versión. Los revisores no siguen reglas fijas; en realidad, no siguen ninguna regla. Las fuentes se citan a veces en las referencias bibliográficas, a veces integradas en el texto o bien ni se mencionan. En cualquier caso, las referencias no suelen incluir información suficiente para garantizar el futuro acceso al mismo conjunto de datos». [1]

Para resolver este problema (y otros), M. Altman y G. King (Harvard-MIT Data Center) proponen la siguiente solución: «Las referencias a los datos digitales deben incluir seis elementos básicos. Los tres primeros son clásicos, ya que son los que se usan para los documentos en papel. Se trata del nombre del autor o de los autores, la fecha de publicación y el título del conjunto de datos. La presentación de estos elementos debe adaptarse al tipo de artículo u obra en el que se cite la referencia. El autor, la fecha y el título permiten ver rápidamente qué tipo de datos se han consultado y cuándo. No obstante, esta información no basta para identificar un conjunto de datos de forma inequívoca, ni para localizarlo, recuperarlo o examinarlo de manera fiable, por lo que hemos decidido añadir otros tres elementos basados en la tecnología moderna, cada uno pensado para que no pierda validez aunque evolucione la tecnología. El objetivo es también sacar el máximo partido del formato digital de los datos cuantitativos.

El cuarto elemento es un identificador único. Se trata de una designación sintética o una cadena de caracteres única que permita identificar de forma permanente el conjunto de datos con independencia del lugar en donde esté almacenado. […] Los identificadores únicos garantizan la continuidad del vínculo entre la referencia y el objeto en cuestión. De todos modos, hay que garantizar y comprobar que el objeto en cuestión no varíe de manera significativa en caso de que cambie el formato de almacenamiento. Añadimos, pues, una firma electrónica universal (Universal Numeric Fingerprint, UNF).

Se trata de una pequeña cadena de caracteres alfanuméricos con una extensión fija que resume el contenido del conjunto de datos. Cualquier modificación de los datos, por mínima que fuese, supondría una nueva firma electrónica universal. […] El último elemento de una referencia estándar sería un gestor de soportes. […] Un ejemplo de referencia completa con esos elementos básicos estándar sería:

> Sidney Verba. 1998. U.S. and Russian Social and Political Participation Data. hdl:1902.4/00754; UNF:4:ZNQRI14053UZq389x0Bffg?==;http://id.thedata.org/hdl%3A1902.4%2F00754» [1] (Traducción del inglés)

Cuadro 12: Condiciones de acceso y de uso aplicables a los archivos de datos protegidos mediante licencia

Observación: los puntos 1-7 son idénticos a las condiciones aplicables para los archivos de uso público. Los puntos 8 y 9 se adaptarán según el marco del contrato.

1. Los datos y otros documentos proporcionados por el INE no pueden, bajo ningún concepto, revenderse o cederse a terceras personas, instituciones u organizaciones sin el consentimiento por escrito de la ONE.
2. Los datos se utilizarán exclusivamente con fines estadísticos o en el marco de trabajos de investigación científica. Servirán solo para crear datos agregados, incluyendo la modelización de información, y queda prohibida cualquier investigación sobre particulares u organizaciones individuales.
3. Queda prohibido cualquier intento de reidentificación de los encuestados. Es más, no se utilizará la identidad de ninguna persona u organización que se descubra fortuitamente y cualquier descubrimiento de este tipo deberá comunicarse de inmediato a la ONE.
4. En ningún caso se intentarán cotejar conjuntos de datos distribuidos por el INE, ni datos del INE con otros conjuntos de datos que pudiesen dar lugar a la identificación de particulares u organizaciones.
5. Los libros, artículos, documentos de trabajo, tesis, disertaciones, informes o cualquier otra publicación que se basen en los datos obtenidos del INE deberán citar su fuente, de conformidad con la obligación de cita asociada al conjunto de datos proporcionado.
6. Se enviará al INE una copia en formato digital de todas las publicaciones basadas en los datos solicitados.
7. El INE y los organismos financiadores correspondientes se deslindan de toda responsabilidad en cuanto al uso y la interpretación de los datos o en cuanto a las conclusiones que se deriven de ellos.
8. Debe indicarse el nombre de la organización, del investigador principal y también del resto de los investigadores que utilizarán los datos. El investigador principal deberá firmar un contrato de licencia en nombre de su organización. Si no está autorizado a firmar en nombre de la organización receptora, habrá que indicar el nombre del representante de la organización.
9. Debe precisarse el uso al que se destinarán los datos, así como una lista de resultados previstos y la política de difusión de la organización.

(Las condiciones 8 y 9 pueden variar para los centros de enseñanza superior).

Teniendo en cuenta la importancia de disponer de una política de difusión de microdatos viable, los INEs deberán prever procedimientos de ejecución de las correspondientes disposiciones:

- Cualquier infracción deberá tratarse de inmediato, lo cual es de vital importancia para mantener la confianza de los encuestados y la credibilidad de la organización.

- En caso de que se infrinja la ley, se podrán emprender acciones legales.

- Si los investigadores no respetan sus compromisos, el INE puede verse obligada a suspender los derechos de acceso tanto individuales como del organismo para el que trabajan.

- Si el compromiso contractual lo asume una organización representada por un investigador, las sanciones que hubiese que aplicar respecto a uno de sus miembros las establecerá la propia organización en lugar del INE. La pérdida de un derecho de acceso individual puede suponer la misma sanción para el conjunto de la organización.

- El INE tomará las medidas necesarias para evitar otras infracciones.

- Para infracciones menores, puede bastar con un aviso.

Información complementaria sobre la gestión de las infracciones respecto a los contratos de uso:

> «Es recomendable que este tipo de acuerdo se fundamente en una cierta base jurídica, por ejemplo incluyéndolo en una legislación *ad hoc*. De este modo podrían tomarse medidas legales contra aquellos que cometan infracciones. Esto no impide que se tomen otras medidas contra las infracciones, como no prestar más servicios al investigador implicado y/o al organismo del que depende». [24]

Si el INE desea recibir comentarios de los usuarios, debería implementar procedimientos regulares de seguimiento a los investigadores. Puede aprovechar la oportunidad para recordarles la obligatoriedad de comunicar sus resultados y también pedirles que hagan sugerencias para mejorar el programa de encuestas.

Cuadro 13: Contrato marco

Contrato entre [el organismo proveedor] y [el organismo beneficiario] relativo al depósito y uso de microdatos:

A. **El presente contrato se refiere a los siguientes conjuntos de microdatos**:

1. _____
2. _____
3. _____
4. _____
5. _____

B. **Términos del contrato**:

En calidad de propietario de los derechos de autor relativos a los productos mencionados en la sección A, o en virtud de la adecuada autorización otorgada por este, el representante de [el organismo proveedor] acepta suministrar a [el organismo beneficiario] los conjuntos de datos mencionados en la sección A, quedando reservado su uso a los empleados de [el organismo beneficiario], siempre que se respeten las siguientes condiciones:

1. Los microdatos (incluidos los subconjuntos de datos) y otros elementos protegidos por derechos de autor publicados por [el organismo proveedor] no podrán revenderse o cederse a terceras personas, instituciones u organizaciones sin la autorización por escrito de [el organismo proveedor]. En cambio, los elementos protegidos por derechos de autor que no contengan microdatos (por ejemplo cuestionarios de encuestas, manuales, listas de códigos o diccionarios de datos) podrán distribuirse sin necesidad de autorización previa. [El organismo proveedor] sigue siendo el propietario del conjunto de los productos suministrados.
2. Los datos se utilizarán exclusivamente con fines estadísticos o en el marco de trabajos de investigación científica. Servirán solamente para crear datos agregados, incluyendo la modelización de información, y queda prohibida cualquier investigación sobre particulares u organizaciones individuales.
3. Queda prohibido cualquier intento de reidentificación de los encuestados. Es más, no se utilizará la identidad de ninguna persona u organización que se haya descubierto fortuitamente y cualquier descubrimiento de este tipo deberá comunicarse de inmediato a [el organismo proveedor].
4. En ningún caso se intentarán cotejar conjuntos de datos distribuidos por [el organismo proveedor], ni datos de [el organismo proveedor] con otros conjuntos de datos que pudiesen dar lugar a la identificación de particulares u organizaciones.
5. Todos los libros, artículos, documentos de trabajo, tesis, disertaciones, informes o cualquier otra publicación que se basen en los datos obtenidos de este archivo o proveedor deberán citar su fuente, de conformidad con la obligación de cita asociada al conjunto de datos proporcionado.
6. Se enviará a [el organismo proveedor] una copia en formato digital de todas las publicaciones basadas en los datos solicitados.
7. [El organismo proveedor] y los organismos de financiación pertinentes declinan toda responsabilidad en cuanto al uso y la interpretación de los datos o en cuanto a las conclusiones que se deriven de ellos.
8. Los datos se almacenarán en un entorno seguro y con las restricciones de acceso apropiadas. [El organismo proveedor] se reserva el derecho de solicitar en cualquier momento información sobre las modalidades de almacenamiento y los sistemas de difusión implantados.
9. [El organismo beneficiario] se compromete a presentar a [el organismo proveedor] un informe anual sobre el uso y los usuarios de los conjuntos de datos antes mencionados, indicando el número de investigadores con acceso a cada conjunto de datos y los resultados de los trabajos de investigación llevados a cabo.
10. Se permite el acceso a los datos por un periodo [determinado o ilimitado, según el contrato].

C. **Comunicaciones entre las partes:**

[El organismo beneficiario] designará a un único interlocutor para el presente contrato. Si esa persona fuese sustituida, [el organismo beneficiario] se compromete a comunicar sin demora a [el organismo proveedor] el nombre y los datos de contacto de un nuevo interlocutor. La correspondencia administrativa o relativa a un procedimiento específico puede desarrollarse por correo electrónico, fax o correo postal indicando siempre los datos siguientes:

Comunicaciones enviadas por [el organismo proveedor] a [el organismo beneficiario]:

Nombre de la persona de contacto: _____
Cargo de la persona de contacto: _____
Dirección del destinatario: _____
_____
E-mail: _____
Fax: _____

Comunicaciones enviadas por [el organismo beneficiario] a [el depositario]:

Nombre de la persona de contacto: _____
Cargo de la persona de contacto: _____
Dirección del destinatario: _____
_____
E-mail: _____
Fax: _____

D. **Firmantes**

Los abajo firmantes han leído y aceptan los términos del presente contrato:

Representante de [el organismo proveedor]

Nombre _____

Firma _____ Fecha _____

Representante de [el organismo beneficiario]

Nombre _____

Firma _____ Fecha _____

7. ¿Qué se entiende por «anonimización» de los microdatos?

Este capítulo se basa en gran medida en el manual publicado por el CENEX sobre el control de la divulgación de estadísticas [3], que puede descargarse gratuitamente en la siguiente dirección: http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf. (en inglés).

La difusión de archivos de microdatos al público, a los investigadores o a otras organizaciones es una misión delicada para los INEs. Por un lado, se trata de publicar archivos de microdatos en los que se basarán un gran número de análisis estadísticos y, por otro lado, el de proteger la identidad de los encuestados. El conjunto de procesos relativos a este segundo aspecto recibe el nombre de control de la divulgación de estadísticas (CDE) o anonimización.

«Como la confidencialidad no puede garantizarse al 100%, los riesgos y las ventajas relacionados con el acceso a los datos deben valorarse en función de los riesgos de divulgación y de la sensibilidad de los datos» [14]. La protección de los datos es una preocupación totalmente legítima para los institutos de estadística y los demás productores de datos y se ha publicado mucho respecto a este tema. En toda esta literatura sobre la gestión de la confidencialidad de los datos con frecuencia aparecen referencias a lo que Peter Madsen [13] denomina la «paradoja de la confidencialidad». En el año 2003, en un taller sobre confidencialidad organizado bajo el auspicio de la organización estadounidense National Science Foundation, Peter Madsen señaló que «el esmero por garantizar la privacidad de los datos en el contexto de la investigación resulta paradójicamente en un menor, en lugar de mayor, beneficio para la sociedad».

7.1 Conceptos relacionados con el control de la divulgación de estadísticas (CDE)

La divulgación se produce cuando una persona u organización advierte o descubre una información inédita sobre otra persona u organización. Se distinguen dos tipos de riesgos de divulgación: la **divulgación de identidad** y la **divulgación de atributos**. La primera se produce cuando la identidad está directamente asociada a un registro del archivo de datos difundido. Ocurre cuando el registro en cuestión contiene variables que permiten identificar inequívocamente a un encuestado (por ejemplo, nombre, número de pasaporte/documento de identidad o número de teléfono). Es fundamental eliminar esas variables de cualquier archivo de microdatos antes de su difusión. La divulgación de atributos se produce cuando los valores (concretos o estimados) presentes en los datos difundidos pueden asociarse a una persona en particular.

Se denomina «clave» (o variable clave) a una combinación de variables de un registro de microdatos que permite identificar a un encuestado. Esta «reidentificación» es posible cuando (a) el encuestado pertenece a una clase de población poco común respecto a un determinado valor clave; y cuando (b) esta clave puede servir para cotejar un archivo de microdatos con otros archivos de datos susceptibles de contener identificadores directos o de otro tipo (padrones electorales, catastros o expedientes escolares, incluso utilizando motores de búsqueda públicos en Internet.)

En la mayoría de los países en desarrollo, el riesgo de divulgación relacionado con la posibilidad de cotejar un archivo de microdatos de una encuesta con otros archivos de datos es bastante limitado, bien porque todavía no existen o bien porque no están ampliamente difundidos. En realidad, el riesgo de divulgación inherente a la difusión de archivos de microdatos de encuesta puede minimizarse de forma satisfactoria sencillamente suprimiendo los identificadores directos de los registros, así como las informaciones geográficas demasiado precisas que van más allá del nivel de muestreo para el cual la encuesta es representativa. Aun así, es recomendable evaluar el riesgo de divulgación caso por caso, puesto que existen excepciones importantes a esta regla general que justifican medidas de control adicionales. Los archivos de microdatos de encuestas con registros relativos a grupos objetivo reducidos —por ejemplo una empresa— son más difíciles de anonimizar.

Por último, otra consideración importante para los INEs: es necesario que protejan sus marcos de muestreo, sobre todo el marco maestro de muestreo intercensal que se utiliza para diseñar las distintas encuestas a hogares. La difusión de los marcos de muestreo detallados es incompatible con las medidas encaminadas a proteger la identidad de los encuestados. Esta información podría emplearse para cotejar datos: así, por ejemplo, podrían establecerse detalles geográficos (tales como el nombre y localización de la unidad primaria de muestreo o UPM) cotejando los factores de expansión y muestreo (que deben aparecer

en archivos de microdatos para que los usuarios puedan sacar conclusiones estadísticas correctas). La protección de los marcos de muestreo se explica sobre todo por la voluntad de limitar la probabilidad de que los encuestados de una UPM sean seleccionados para otras encuestas al margen del INE. En los países en los que las UPM del marco de muestreo maestro se actualizan cada 10 años, esto podría llevar a mayores porcentajes de no respuesta (encuestados hastiados) y provocar distorsiones.

7.2 Escenarios de divulgación

La primera fase del proceso de anonimización de un archivo de microdatos para su difusión consiste en identificar qué partes del archivo presentan un riesgo de divulgación. Así, antes de implementar medidas de control de la divulgación de estadísticas, es interesante analizar los distintos casos en los que el usuario de un archivo de microdatos podría identificar a un encuestado. Se distinguen principalmente dos casos:

- **Conocimiento de la respuesta:** el usuario, a título personal, dispone de información suficiente sobre los atributos asociados a uno o varios registros. Dicho de otro modo, el usuario pertenece al círculo o está alrededor de una unidad estadística. Este hecho es más frecuente cuanto más reducido es el grupo objetivo (por ejemplo, encuestas dentro de una empresa o entidad y encuestas a hogares en países pequeños o en regiones con baja densidad de población).

- **Cotejo con archivos externos:** el usuario asocia determinados registros del archivo de microdatos difundido con otros conjuntos de datos (o registros) públicos que contienen identificadores directos, a pesar de la prohibición expresa estipulada en el contrato de uso de los datos. Para hacerlo, utiliza las variables clave disponibles en los dos conjuntos de datos y crea claves de identificación (combinación de datos).

Antes de anonimizar datos, es recomendable definir un escenario de divulgación para describir tanto la información que podría revelarse como los posibles métodos de identificación de particulares. Normalmente, este escenario se basará en hipótesis de máxima cautela con la finalidad de prever el peor de los casos. En ocasiones será necesario definir varios escenarios, ya que al usuario se le pueden proponer, simultánea o alternativamente, distintas fuentes de información.

7.3 Evaluación del riesgo de divulgación

La violación de la confidencialidad se produce en caso de reidentificación de un encuestado y cuando el infractor (un usuario no autorizado o que no ha respetado las condiciones estipuladas en el contrato relativo al acceso a los datos y a su uso) llega a asociar variables sensibles a un individuo.

Antes de difundir los microdatos, los INEs deben evaluar los datos «para determinar si su publicación pondría en peligro la identidad de determinados individuos u organizaciones. En este marco hay que considerar varios factores: 1) el nivel de detalle de los datos que se van a difundir (en especial precisiones geográficas y posibles variables en común con fuentes de datos externas, susceptibles de servir como claves y que aumentan el riesgo de identificación); 2) las variables o combinaciones de variables que permiten aislar a encuestados dentro de la muestra y que podrían facilitar su identificación por parte de personas externas; y 3) la existencia de otros datos accesibles en el exterior [del INE], como por ejemplo los ya publicados sobre la misma encuesta o sobre una encuesta relacionada, o incluso de información sobre el mismo encuestado en posesión de otras organizaciones». [14]

«Determinar el riesgo de divulgación de datos identificables es una labor muy compleja que requiere tanto un análisis estadístico empírico como muy buen juicio» [14]. Existen varios métodos, cada uno con sus ventajas, si bien ninguno se ha impuesto como «el mejor». Sin duda, el riesgo de divulgación no puede eliminarse por completo, aunque pueden preverse medidas para compensarlo que, no obstante, reducen la utilidad de los datos. Estas medidas implican establecer un umbral de riesgo (*threshold rule* en inglés) que permita determinar si la difusión del conjunto de datos es o no segura. Existen básicamente dos métodos de cálculo matemático del riesgo de reidentificación:

- **Cálculos individuales:** evaluación del riesgo para cada registro. Estos cálculos se expresan en general en forma de probabilidad de reidentificación de un encuestado o en términos de unicidad y singularidad dentro de la muestra.

- **Cálculos globales:** evaluación del riesgo para el conjunto del archivo. Estos cálculos son cuantificados como el número esperado de reidentificaciones y pueden derivarse agregando los cálculos individuales

La ventaja de los cálculos individuales es que solo hay que proteger aquellos registros que aparecen como sensibles respecto al umbral de riesgo establecido, con lo cual se minimiza la pérdida de información y de utilidad de los datos. Limitarse a un cálculo general del riesgo de reidentificación puede incitar a recurrir a técnicas de CDE para cada registro del archivo de microdatos. En ese caso, podría producirse una pérdida mayor de información, y potencialmente reducir la utilidad de los datos para el análisis estadístico.

Los métodos de cálculo del riesgo de reidentificación se caracterizan también por la utilización de claves. Los cálculos basados en claves de la muestra permiten identificar combinaciones únicas o singulares de variables categóricas (claves) dentro de la muestra. Una unidad o registro se considera bajo riesgo cuando su combinación de puntuaciones para las variables de identificación es inferior al umbral establecido. Por ejemplo, lo más probable es que un hombre de 30 años, médico de profesión y padre de cuatro niñas sea único dentro de la muestra de una encuesta. El registro correspondiente se considerará, pues, sensible, aunque en la población total pueda haber muchos otros individuos con esas mismas características. El riesgo para una unidad dada también puede determinarse combinando puntuaciones para las variables de identificación categóricas dentro de la población o por su probabilidad de reidentificación. Como la frecuencia de aparición dentro de la población normalmente se desconoce, esas probabilidades deben establecerse a través de modelización. Con variables de identificación continuas es imposible aplicar el criterio de singularidad de las claves, ya que la mayoría, si no todas, suelen ser únicas. El riesgo de divulgación a través de las variables continuas se calcula estimando la probabilidad de reidentificación por combinación de variables de dos conjuntos de datos distintos, a partir de la «proximidad» de sus valores.

Existen varias maneras de identificar en un archivo casos y variables que pueden provocar la divulgación de información. Uno de los métodos más habituales consiste en generar distribuciones de frecuencia y tablas multidimensionales para identificar celdas con pocos casos. La información geográfica detallada (o desagregada) es una de las principales fuentes de identificación de individuos, sobre todo por parte de usuarios de la misma región y que conocen bien las características de algunos encuestados.

Cuadro 14: Elementos de evaluación de los distintos escenarios y riesgos de divulgación de microdatos

Con miras a la revisión previa a la difusión de archivos de microdatos de encuestas y de censos para los que la Oficina del Censo de los Estados Unidos debe proteger la confidencialidad de los encuestados, existe una lista de elementos que deben verificarse (Checklist on Disclosure Potential of Proposed Data Releases, www.census.gov/srd/sdc/). Dicha lista facilita el proceso de control de la divulgación. El objetivo es ayudar a las personas encargadas del control de la divulgación a determinar si hay que optar por la publicación de archivos de microdatos de uso público o bien de datos en tablas. La sección 3 de la lista está dedicada a los archivos de microdatos. Los puntos que hay que verificar en este sentido son los relativos a elementos importantes como información geográfica, las variables con un riesgo inusual de divulgación de la identidad, información contextual y ecológica, la posibilidad de una asociación, comparación o cotejamiento con otros datos que podrían llevar a la identificación de una combinación única de atributos.

7.4 Técnicas de CDE específicas para archivos de microdatos

La primera gran fase del CDE de un archivo de microdatos es la supresión de todos los identificadores directos (variables que permiten identificar inequívocamente a un encuestado). A continuación, se procede a la anonimización del archivo de microdatos mediante **métodos de enmascaramiento** o a través de la creación de **microdatos sintéticos**. Un archivo de microdatos sintéticos se crea aleatoriamente mediante un proceso que permite conservar determinadas correlaciones estadísticas o las relaciones existentes al interior del archivo de microdatos original

En cuanto a los métodos de enmascaramiento, son técnicas que permiten crear una versión modificada del archivo inicial de microdatos brutos. Existen dos tipos de métodos de CDE por enmascaramiento. Los **métodos de perturbación de la información** consisten en modificar los datos antes de su publicación e introducir deliberadamente un elemento erróneo por motivos de confidencialidad. Los métodos de restricción de la información consisten en reducir el contenido informativo de los datos proporcionados suprimiendo una parte o por agregación.

A continuación se ofrece una presentación general de las técnicas de CDE más habituales, junto con el tipo de datos a los cuales se aplican: variables continuas

o categóricas, o ambas. Una variable se considera continua cuando es numérica y permite operaciones aritméticas (por ejemplo, ingresos, edad y número de miembros de una familia). Las variables que se definen únicamente por un conjunto finito y que no sirven para realizar cálculos matemáticos se denominan cualitativas o categóricas (variables clasificadas en una escala ordinal, como el nivel de estudios, o medidas en una escala nominal, como el estado civil, en que el orden de los valores es irrelevante).

Técnicas de restricción

Las técnicas que se basan en la restricción de los datos consisten en modificar los archivos de microdatos con la finalidad de eliminar las variables o los registros susceptibles de ser asociados de forma inequívoca a un individuo en particular. Otra solución consiste en crear categorías, de modo que aumente el número potencial de encuestados para una categoría concreta. Así, el INE puede decidir establecer un umbral que imponga un mínimo de respuestas por celda. Las seis principales técnicas de enmascaramiento por restricción son:

1. **Muestreo:** en vez de la difusión del archivo de microdatos completo, y esta técnica se recomienda para la publicación de datos del censo de población. Siempre que el muestreo sea lo suficientemente reducido (por ejemplo, 5% de la población) y se supriman todas las variables de identificación directas, esta técnica permite reducir el riesgo de divulgación de forma satisfactoria para los datos categóricos. No obstante, si el archivo de microdatos contiene variables continuas, susceptibles de ser combinadas más fácilmente con un archivo de datos externo, se precisan técnicas de CDE suplementarias.

2. **Recodificación global:** consiste en reagrupar determinados valores en función de clasificaciones predefinidas, de modo que desaparezcan las repuestas individuales. Este método se puede aplicar tanto a las variables continuas o discretas como a los códigos geográficos. Así, por ejemplo, la edad puede distribuirse en intervalos de edades, o la profesión y los códigos sectoriales pueden agruparse en categorías más amplias. Asimismo, puede suprimirse la información geográfica más precisa por debajo del nivel para el cual la encuesta, dado su diseño muestral, es representativa. El método de recodificación global es adecuado tanto para datos continuos como categóricos.

3. **Agrupación de valores extremos: es una** operación que se ejecuta en presencia de variables numéricas u ordinales cuyos valores máximos o mínimos son extremadamente inusuales y susceptibles de revelar la identidad de los encuestados. Para agrupar esos valores extremos se crean categorías «comodín» (por ejemplo, para valores superiores, «personas mayores de X años» o «ingresos superiores a Y»). Esta técnica es adecuada para variables continuas y variables categóricas medidas en escalas ordinales.

4. **Supresión local: es** técnica sencilla que se utiliza cuando la asociación de dos variables puede llevar a identificar a una persona en concreto. Dicho de otro modo, una combinación de las variables puede proporcionar una clave de reidentificación para un registro concreto. Por ejemplo, un registro que contiene la siguiente información: edad = 85 y nivel de estudios = actualmente inscrito en la escuela primaria. Es muy probable que este registro sea único, en la medida en que no hay muchos octogenarios inscritos en primaria. En este caso concreto, bastaría con suprimir una de esas dos informaciones (situación escolar o edad) para eliminar el problema. La supresión local es útil sobre todo para los datos categóricos.

5. **Supresión de determinadas variables:** método necesario para informaciones consideradas inadecuadas para su difusión por ser demasiado sensibles (por ejemplo, etnia o religión).

6. **Supresión de determinados registros:** a veces es necesaria esta técnica para preservar el anonimato de encuestados con un conjunto único de variables. Cuando se suprime por completo un registro del archivo de microdatos, es vital calcular e incluir factores de expansión ajustados. Este método debe utilizarse con moderación, puesto que la supresión de un número significativo de registros supone una distorsión de los datos.

Técnicas de perturbación

Las técnicas de perturbación consisten en modificar los datos de forma que el emparejamiento de datos sea menos preciso y más complejo. En caso de tentativa de reidentificación, los valores modificados de este modo

hacen que el emparejamiento sea incierto. Las siete principales técnicas de perturbación son:

1. **Adición de ruido (o perturbación aleatoria)**: consiste en añadir valores aleatorios a los reportados por un encuestado. Los métodos varían en función de si se añade ruido a una sola variable o a varias, o si se desea mantener las medias, las varianzas y covarianzas. Las técnicas de programación lineal permiten, además, minimizar las desviaciones entre los valores modificados y los reales.

2. **Intercambio de datos**: consiste en modificar un archivo de microdatos sustituyendo los valores correspondientes a variables confidenciales por los de otro registro. Los registros se «intercambian» de modo que se conserven los valores marginales o de mínima frecuencia de ocurrencia. Esta técnica puede utilizarse para variables continuas o categóricas.

3. **Intercambio de rangos: en esta aproximación** las variables que deben protegerse son ordenadas de manera ascendente y luego se construyen agrupaciones. Se seleccionan al azar parejas de registros en cada grupo y sus valores se intercambian con los de otras parejas dentro de un rango predefinido. La creación de grupos de distintos tamaños produce diferentes distribuciones (o presentaciones) de los datos.

4. **Microagregación**: consiste en sustituir un valor observado en una muestra por la media calculada para un grupo reducido de unidades (pequeña agregación o microagregación), entre las cuales está la controlada. En el archivo difundido, las unidades de un mismo grupo están representadas por el mismo valor. Cada grupo contiene un número mínimo de unidades predefinido (k). El valor mínimo admisible de k es 3. Para un determinado valor k, el objetivo es dividir todo el conjunto de las unidades en grupos de al menos k unidades, minimizando la pérdida de información, que por lo general se expresa como de pérdida de variabilidad. En consecuencia, las unidades se agrupan en función de su máxima similitud. El mecanismo de microagregación protege los datos garantizando que el archivo difundido contenga por lo menos k unidades con el mismo valor en el archivo de datos. Esta técnica se aplica en ocasiones a variables continuas.

5. **Redondeo:** se emplean distintos tipos de redondeo según el objetivo que se persiga. Existe el redondeo controlado, para conservar los totales así como ciertas propiedades de adición: o el redondeo aleatorio, destinado a garantizar que en las celdas de una tabla de datos agregados no revelen recuentos de una o dos observaciones.

6. **Remuestreo:** consiste en sacar distintas muestras independientes de los valores relativos a las variables que hay que enmascarar. Estas variables se ordenan siguiendo el mismo criterio de ordenamiento. Las variables enmascaradas se crean tomando como primer valor la media de todos los primeros valores de cada muestra, como segundo valor la media de los segundos, y así sucesivamente.

7. **Postaleatorización (o aleatorización a posteriori): es una** versión aleatoria de la técnica intercambio de datos. Esta técnica añade un elemento de incertidumbre a los valores de determinadas variables al sustituirlos según un mecanismo probabilístico. Como en el caso del intercambio de datos, la protección de los datos se garantiza porque los usuarios no pueden establecer con certeza la exactitud de un valor difundido. En consecuencia, cualquier tentativa de cotejar el registro con identificadores externos puede conducir fácilmente a errores de combinación o interpretación. Este método se utiliza básicamente para variables categóricas, pero también puede aplicarse a variables numéricas continuas.

Archivos de microdatos sintéticos

Los archivos de microdatos sintéticos se crean mediante un proceso aleatorio con la condición de que se mantengan las correlaciones estadísticas o internas del archivo inicial. Es tentador distribuir archivos totalmente construidos con ayuda de simulaciones y que, por lo tanto, no presentan ningún riesgo de divulgación, en lugar de los archivos de microdatos iniciales. A tal efecto se han desarrollado varias técnicas. Sin embargo, en comparación con los métodos de enmascaramiento de datos, las distintas técnicas de creación de archivos de microdatos sintéticos son extremadamente laboriosas y complejas. Los datos de encuestas suelen incluir cientos de variables,

cuyas distribuciones y relaciones no son fáciles de modelar con herramientas paramétricas estándar. Se sigue investigando para mejorar e implementar estas técnicas.

7.5 El equilibrio entre el riesgo de divulgación y la pérdida de información

La modificación de un archivo de microdatos mediante técnicas de control de la divulgación de estadísticas se traduce en una pérdida de información. Corresponde al INE hallar el correcto equilibrio entre esa pérdida de información y el riesgo de divulgación. Del mismo modo que evalúa el riesgo de divulgación, el INE puede evaluar la pérdida de información relacionada con las distintas técnicas de CDE. Para los datos categóricos, esa pérdida puede analizarse con métodos como la comparación directa, la comparación de tablas de contingencia o medidas basadas en entropía. En el caso de variables continuas, puede evaluarse por comparación de las medias cuadráticas, las medias absolutas o la variación entre medias. No obstante, como el potencial de utilización de los archivos de microdatos es tan amplio, resulta imposible calcular con precisión la pérdida de información. La mejor solución es identificar qué categoría de usuarios sería la más afectada por la aplicación de medidas de CDE, que por lo general son investigadores acostumbrados a realizar análisis estadísticos avanzados a partir de archivos de microdatos. Esta categoría de usuarios no suele ser muy amplia y sus trabajos pueden contribuir de forma significativa e importante al interés general, por lo que se subraya la necesidad no solo de difundir AMUP sino también los archivos con licencia menos anonimizados. Además, los INEs deberían hacerse una idea bastante precisa de la pérdida de información relacionada con el control de la divulgación de estadísticas (CDE) examinando los formularios de acceso a los archivos protegidos mediante licencia, donde figuran los motivos por los que la versión de uso público de un archivo de microdatos no se considera útil para determinados proyectos de investigación.

7.6 La documentación del proceso de divulgación de los datos estadísticos

Los métodos de CDE empleados deben permitir equilibrar la pérdida de información y la probabilidad de divulgación de información personal. Llegado el caso, se recomienda informar a los usuarios de que se ha evaluado el riesgo de divulgación del conjunto de datos y de los métodos de protección que se han aplicado. Sería también deseable informarles de la naturaleza y el alcance de cualquier modificación realizada para el control de la divulgación.

Podrían indicarse la técnica o técnicas empleadas, si bien el nivel de detalle no deberá permitir al usuario reconstruir los archivos de microdatos iniciales mediante ningún sistema.

Cuadro 15: Documentación de la Oficina del Censo estadounidense relativa a las medidas de CDE aplicadas a las muestras de microdatos de uso público del censo del año 2000

«La confidencialidad se protegerá mediante los siguientes procesos: intercambio de datos, agrupación de valores extremos superiores, umbrales geográficos, perturbación de los datos relativos a la edad en hogares con muchos miembros y restricción de los detalles difundidos para determinadas variables categóricas.

El intercambio de datos es un método de limitación de divulgación de información ideado para proteger la confidencialidad de los datos que figuran en las tablas de frecuencia (número de personas o porcentaje de la población que presenta determinadas características). Consiste en modificar los datos iniciales o en intercambiar registros para una muestra de casos. El intercambio se aplica a registros individuales y, por lo tanto, protege también los microdatos.

La agrupación de valores extremos superiores es un método de limitación de divulgación de información que consiste en incluir dentro de una sola categoría todos los casos iguales o superiores a una determinada distribución de frecuencia.

Los umbrales geográficos impiden la divulgación de los datos de individuos u hogares asociados a unidades geográficas con recuentos de población inferiores a un determinado nivel (véase la descripción de las unidades de microdatos de uso público (UMUP) y de las súper-UMUP en la sección III).

La perturbación de los datos relativos a la edad, es decir, la modificación de las edades de los miembros de un hogar, es necesaria para grandes hogares (10 o más miembros) por motivos de confidencialidad.

El grado de detalle de las variables categóricas se variará si la categoría en cuestión no cumple el umbral de población nacional mínimo establecido».

Fuente: http://www.census.gov/population/www/cen2000/pums/index.html, en inglés, consulta el 10 de agosto de 2010.

8. El acceso a los microdatos, ¿debe ser cobrado o gratuito?

Los productores de datos pueden considerar la posibilidad de vender microdatos como un medio para recuperar parte de sus costes de producción (no siempre presupuestados). Las oficinas de estadísticas, que cada vez disfrutan de una mayor autonomía, a menudo se ven obligadas a detectar y poner en marcha actividades que les reporten ingresos. En este sentido, los microdatos constituyen un producto muy valioso.

La publicación de los archivos de microdatos aumenta el valor de una encuesta, pero también acarrea unos costes relacionados con la documentación y la anonimización. Cuando la producción de microdatos para su posterior difusión no está incluida en el presupuesto de la encuesta, el INE no tiene otra opción que intentar recuperar ese coste.

Un servicio completo de difusión de microdatos también genera otros costes, sobre todo relacionados con el control de los archivos antes de su difusión, la concesión de licencias de uso, la explotación de un centro de datos seguro, la asistencia que se ofrece a los usuarios y la gestión de la infraestructura. Por norma general, para llevar a cabo estas tareas se precisa personal e infraestructura. Si la organización no dispone del presupuesto necesario o si no hay personal que pueda hacerse cargo de esas tareas, difícilmente podrá garantizarse un servicio viable y de calidad.

La venta de datos es una forma de trasladar a los beneficiarios del servicio de datos una parte de los costes.

8.1 Ejemplos de dos países

La recuperación de los costes por parte de los INEs tiene una larga tradición en todo el mundo. Se generalizó en la década de 1980, a menudo como respuesta a recortes presupuestarios y a las presiones de los organismos competentes para que los costes de las estadísticas se trasladasen del contribuyente, es decir de quienes pagan impuestos, al usuario. Aunque el objetivo del presente documento no es analizar detalladamente las actividades que realizan los INEs para recuperar los costes, a continuación se presentan algunas observaciones:

- El instituto de estadística neozelandés (SNZ) trató de recuperar el 25% de su presupuesto total a través de la venta de varios productos y servicios. La decisión se tomó en un periodo de austeridad administrativa en el país y la iniciativa permitió evitar la supresión de algunos programas. No obstante, la organización no logró su objetivo y la iniciativa se desestimó. El SNZ renunció entonces a recuperar los costes y puso en práctica una estrategia de reducción de los mismos. Según su propio sitio web, se puede acceder al 90% de su información de forma gratuita. Se perciben tasas por tabulaciones personalizadas y tablas detalladas. El acceso a los archivos de microdatos es cobrado, pero en el sitio web no se indican tarifas.

- Statistics Canada puso en marcha un programa completo de recuperación de costes en los años ochenta para hacer frente a la presión presupuestaria y política y se consagró a esa labor a pesar de las duras críticas de los usuarios. El programa tuvo un resultado moderado y fue en gran parte suprimido, o considerablemente reducido. Cuando el precio de acceso a los microdatos aumentó de forma considerable, los investigadores canadienses comenzaron a usar datos recogidos en Estados Unidos, fácilmente accesibles a través del Consorcio Interuniversitario para la Investigación en Ciencias Políticas y Sociales (ICPSR). Para paliar este problema, contrario a los intereses canadienses, se ideó la Iniciativa para la Democratización de los Datos (IDD), cuya creación surgió del siguiente análisis:

  > «[...] el verdadero ejercicio de la democracia significa que los ciudadanos tengan la posibilidad de acceder a información compleja y de adquirir las competencias necesarias para comprenderla». Aunque Paul Bernard es consciente de las presiones que se han ejercido sobre Statistics Canada para que reduzca sus costes y aumente sus ingresos, tiene la sensación de que se han traducido en un «acceso a los datos reservado a las categorías de usuarios que tienen medios para pagarlo». Según Paul Bernard, esta situación podría «obstaculizar la participación en el debate público de las categorías que disponen de recursos limitados, [así como de] las que tienen pocas posibilidades de sacar provecho o de obtener un beneficio tangible

> y relativamente inmediato del uso de los datos». Añade que «a largo plazo, esto podría provocar un desarrollo subóptimo y una simulación de democracia». [31]

Se creó entonces una red de 74 instituciones educativas que daba acceso a toda la colección de datos públicos de Statistics Canada. Estos datos incluían alrededor de 300 AMUP y miles de otros archivos, bases de datos y archivos geográficos.

Las cuotas de suscripción cubren los costes de asistencia técnica y la creación de una infraestructura técnica optimizada para los suscriptores y la organización. El objetivo no es que cubran los costes de producción de los datos.

El portal de acceso a los recursos de la IDD no está abierto a los ministerios, muchos de los cuales han firmado acuerdos con Statistics Canada y participan en la financiación de algunas encuestas. En general, esos acuerdos ya incluyen el acceso a los microdatos, de modo que el servicio propuesto no responde a una necesidad de los ministerios.

8.2 ¿Acceso cobrado o gratuito?

No hay una respuesta definitiva a esta pregunta. Muchos son los argumentos a favor de una minimización de los precios —o incluso de la gratuidad— de acceso a los archivos de microdatos.

Acceso gratuito

El principal argumento a favor de la gratuidad es que muchas ONE están reforzando sus prácticas de difusión de archivos de microdatos y temen que unas tarifas demasiado elevadas desanimen a los usuarios.

- El cobro por el acceso a los datos reduce considerablemente el número de usuarios potenciales y, por lo tanto, el valor de los datos.
- En los países en desarrollo, puede ser un obstáculo para los principales interesados: estudiantes, centros de investigación locales, universidades, etc.

El otro argumento en contra del pago por acceso a los datos es el coste relacionado con el cobro de los derechos de acceso. Además, según si esos derechos van a parar al INE o un organismo competente, la motivación del personal para percibirlos no será la misma: para un correcto funcionamiento, el sistema debe ser eficaz.

- El hecho de cobrar crea una exigencia de calidad y servicio.
- El cobro por acceso no genera pocos ingresos: la mayoría de las solicitudes proceden de la comunidad universitaria, que tiene recursos limitados y que fácilmente puede usar otros datos obtenidos en otra parte.

La experiencia de algunos países (la mayoría industrializados) muestra que la recuperación de costes es posible hasta cierto punto. Aun así, parece difícil poder recuperar todos los costes suplementarios relacionados con la difusión de los microdatos. También se puede demostrar que una estrategia de recuperación de costes agresiva no favorece el uso de los archivos de microdatos y, a largo plazo, reduce el valor potencial de una encuesta.

La solución ideal para un INE es que todos los costes estén reflejados en el presupuesto de la encuesta, lo cual refuerza la accesibilidad de los datos. Esos costes pueden asumirlos los auspiciantes, lo que potencia los beneficios de la encuesta. Este aspecto es particularmente importante en los países en que los investigadores disponen de medios limitados o los productores tienen pocos recursos que destinar al análisis de datos.

Acceso cobrado

Es muy probable que un sistema de acceso cobrado genere ingresos. No obstante, deben tenerse en cuenta otros factores antes de tomar la decisión de poner en práctica un sistema de este tipo:

- ¿El INE puede cobrar legalmente por los derechos de acceso a sus productos?
- ¿Qué costes desea recuperar el INE?
- ¿Son esos costes claramente identificables? ¿Los entenderán y aceptarán los usuarios?
- ¿Los usuarios tienen medios para poder pagar? ¿Se podría crear una asociación de usuarios para cubrir los costes iniciales? Esto supone una identificación clara de los costes que hay que recuperar y una distribución de dichos costes entre los organismos usuarios.

- ¿Pueden cobrarse de forma eficaz los derechos de acceso?

- Los archivos complejos disponibles de forma gratuita en un sitio web podrían consultarlos personas sin las competencias necesarias para usar microdatos, lo cual podría generar un aumento de las solicitudes de asistencia.

Otro factor importante que hay que tomar en consideración para la elaboración de una estrategia de precios sería su coherencia con la política tarifaria aplicada a otros productos y servicios, como las publicaciones en papel y el acceso a Internet. La mayoría de los sitios web de los INEs son accesibles de forma gratuita, en la medida en que los costes marginales son escasos o inexistentes. Pero no ocurre lo mismo con los productos en papel. Si las tarifas de esas publicaciones se fijan en función de los costes marginales de producción y distribución de las copias suplementarias, el mismo principio podría aplicarse a los archivos de microdatos y al sobrecoste generado por la ayuda o asistencia proporcionada a los usuarios adicionales.

9. ¿En qué momento del ciclo de difusión deben publicarse los microdatos?

La producción y la publicación de los archivos de microdatos deben enmarcarse en un ciclo de difusión. Es muy aconsejable que la información destinada a un público amplio sea difundida con prioridad, de forma que el INE pueda cumplir sus objetivos inmediatos e implementar un sistema de retroalimentación con el público. Esta información incluye principalmente una serie de informes descriptivos de las encuestas y los análisis del productor de los datos. Es importante que los productores de datos oficiales establezcan este tipo de documentos oficiales y que los comuniquen/ publiquen desde el inicio del ciclo de difusión.

La producción de un archivo de microdatos requiere tiempo y recursos especializados, así como un proceso de validación. Además, en ocasiones los INEs deben responder a determinados objetivos analíticos/ científicos, internos o externos, lo que puede aplazar la creación del archivo de microdatos hasta varios meses después de la publicación de los resultados de la encuesta. Sea cual sea el calendario establecido, a los investigadores les gusta que se les comuniquen las fechas de publicación previstas para poder planificar sus propios trabajos. Los plazos deben ser razonables; si se alargan varios años, los resultados serán mucho menos pertinentes.

«Incluso cuando la difusión es bastante rápida, a veces es aconsejable publicar una parte de los microdatos o de los datos agregados antes de que la totalidad de los microdatos esté disponible. [Se recomienda] prever estas cuestiones desde las etapas de planificación, o abordarlas en cuanto se detecte que existe esa necesidad» [14]

Cuadro 16: Política sobre los plazos de publicación de datos del NCHS (Estados Unidos)

«El intercambio abierto de información y de puntos de vista es de interés general y científico. En este sentido, la política del NCHS consiste en difundir los microdatos lo antes posible tras el levantamiento de los datos, teniendo como únicas restricciones aquellas impuestas por los recursos disponibles, las limitaciones técnicas y las exigencias cualitativas. El NCHS no obstaculizará la rápida difusión de los microdatos para reservar a sus empleados, socios o al personal de otras organizaciones los derechos de publicación.

1. Los archivos de datos de uso público se publicarán inmediatamente después de las etapas de preparación, control y validación por parte de los organismos competentes, especialmente el Comité de Control de la Divulgación de Estadísticas del NCHS. Según el grado de interés o especialización de ciertos datos, el NCHS puede solicitar la intervención de determinados colaboradores (entre ellos los donantes) para el proceso de preparación previo a su difusión, incluida su modificación, recodificación y la definición de la estructura definitiva del archivo. Si esto es así, el NCHS podrá comunicar al colaborador en cuestión de datos que todavía no se hayan hecho públicos. Esta accesibilidad a archivos que todavía no están listos para ser difundidos debe reflejarse en la política de confidencialidad del NCHS, y efectuarse en conformidad con las disposiciones legales aplicables al NCHS, con consentimiento informado y de acuerdo con las recomendaciones de un comité de ética sobre investigación. En este contexto, la comunicación de datos suele regirse por un acuerdo que estipula las medidas de protección en materia de confidencialidad que debe tomar el colaborador en cuestión.
2. El NCHS no «bloqueará» los datos listos para su difusión. No dará a sus colaboradores un acceso previo privilegiado a los archivos de datos o tablas listas para su difusión, ni tampoco dará acceso privilegiado a las tablas basadas en archivos de datos que todavía no se hayan hecho públicos.
   Cuando se comunican datos no publicados a un usuario, hay que considerar su posible divulgación a otros solicitantes, dentro de los límites que establezcan las cláusulas de confidencialidad. Los archivos o tablas cuya publicación todavía no se haya aprobado debido al riesgo de divulgación de la información confidencial que contienen pueden, llegado el caso, consultarse en el Centro de Datos del NCHS (o, de acuerdo con la política de confidencialidad del NCHS, mediante contratos de utilización especiales) con objeto de garantizar el máximo acceso a los datos. En casos excepcionales (por ejemplo publicaciones ministeriales con una fecha de aparición tardía), podrá accederse a determinados datos tabulares antes de la difusión general de los datos.

Las excepciones a esta política general serán limitadas y deberán justificarse en cada caso. Las solicitudes deberán presentarse antes de iniciar el levantamiento de datos al responsable de confidencialidad designado por el NCHS, y ser aprobadas por el director general».

Fuente: http://www.cdc.gov/nchs/about/policy/data_release.htm (en inglés), consulta el 7 de junio de 2010.

10. ¿Cuáles son los requisitos en cuanto a infraestructura técnica?

«El acceso [a los microdatos] y la óptima explotación de los datos requieren una infraestructura tecnológica apropiada, un amplio consenso internacional en lo que respecta a la interoperabilidad y mecanismos eficaces para el control de calidad. [...] El mantenimiento a largo plazo de la infraestructura necesaria para el acceso a los datos es de especial importancia. Los centros de investigación y los organismos públicos deberían asumir oficialmente la responsabilidad de garantizar que los datos [...] sean eficazmente protegidos, gestionados y accesibles para que puedan ser utilizados de forma eficiente y adecuada a largo plazo. [...] Es [además] conveniente prestar especial atención a la utilización de técnicas e instrumentos destinados a garantizar la integridad y la seguridad de los datos de investigación. En lo que respecta a la integridad de un conjunto de datos, deberían ponerse todos los medios necesarios para garantizar que los datos sean completos y no haya errores. En cuanto a la seguridad, los datos, al igual que los metadatos y las correspondientes descripciones, deberían protegerse contra la pérdida, la destrucción, la modificación y el acceso no autorizado, intencionados o no, conforme a protocolos de seguridad explícitos». [17]

Se debe poder contar con una infraestructura tecnológica adecuada para cubrir los distintos aspectos del almacenamiento de microdatos (documentación, catalogación y difusión, anonimización y preservación).

Documentación de los microdatos

Se han elaborado normas internacionales sobre metadatos para formalizar la documentación de los microdatos y recursos relacionados. La iniciativa DDI (Data Documentation Initiative) y el estándar Dublin Core descritas en el capítulo 2 son, en este sentido, soluciones prácticas. La elaboración de la documentación según estas normas se facilita gracias a la existencia de editores de metadatos especializados, como por ejemplo el Microdata Management Toolkit de la IHSN (Cuadro 17) y el programa de gestión Nesstar Publisher del NSD noruego.

Catalogación y difusión de los microdatos

Debe informarse a los potenciales usuarios de la existencia y las características de los conjuntos de datos disponibles. A menudo tienen muy poca información sobre esos conjuntos de datos. Así pues, hay que proporcionar metadatos de calidad, preferentemente en forma de catálogo consultable en línea.

Cuadro 17: Microdata Management Toolkit (IHSN)

Con el Microdata Management Toolkit (conjunto de herramientas de gestión de microdatos), desarrollado por el NSD noruego y por el Grupo de Gestión de Datos del Banco Mundial para la IHSN, se pretende potenciar la adopción de las estándares internacionales y buenas prácticas en el ámbito de la documentación, difusión y preservación de microdatos.

El Conjunto de Herramientas incluye dos módulos. **Nesstar Publisher** (editor de metadatos) que se utiliza para documentar los datos de acuerdo con los estándares internacionales de metadatos en vigor (DDI y Dublin Core). Gracias al programa gratuito **Nesstar Explorer** (para explorar y descubrir), los usuarios pueden leer los archivos creados con el Nesstar Publisher. Este programa permite a los usuarios visualizar los metadatos y reexportarlos a distintos formatos comunes (Stata, SPSS, etc.). El Nesstar Publisher y el Explorer se basan en la tecnología Nesstar y han sido desarrollados por el NSD noruego. Finalmente, **CD-ROM Builder** (creador de CD-ROM) se utiliza para generar productos de fácil utilización (CD-ROM, sitios web) para la difusión y el almacenamiento de datos.

Véase http://www.ihsn.org/toolkit.

El objetivo del catálogo de microdatos es que los usuarios puedan acceder a los datos y a la documentación de forma sencilla y práctica. El catálogo de encuestas incluye diversas herramientas para:

- Encontrar el archivo de datos que corresponda a las necesidades del usuario. Esta herramienta es de menor utilidad cuando la cantidad de archivos de microdatos es limitada. Sin embargo, es sumamente práctica cuando el número de archivos aumenta. En ese caso, disponer de una herramienta que permita localizar los archivos por variables resulta realmente útil.

- asegurar la compatibilidad con las necesidades del investigadores (por ejemplo universo de la encuesta, conceptos y definiciones). Tal es la función de los metadatos, que constituyen la documentación del archivo.

- Acceder a los datos. Requiere el uso de un sistema de extracción y/o carga de datos.

Generalmente, los archivos pueden cargarse a través de un sitio o portal web y un servidor FTP. El INE también puede recurrir a este tipo de herramienta internamente para proporcionar datos en CD/DVD.

- Utilizar los datos. No existe una herramienta única para la realización de los trabajos analíticos de los investigadores. Estos prefieren disponer de los datos en distintos formatos y poder utilizar así las herramientas de su elección. Los formatos más comunes incluyen: SPSS, Stata, SAS y ASCII.

A continuación se indican las características de un buen sistema de catalogación de microdatos.

• ***Desde el punto de vista del usuario***, un buen catálogo:

- Está en concordancia con los estándares internacionales de metadatos.

  Las estándares internacionales de metadatos en XML tales como el DDI o el DC facilitan considerablemente la producción y el mantenimiento de estos catálogos.

- Se basa en la Web para facilitar las búsquedas.

- Es rico en metadatos, incluso a nivel de variables.

  Los catálogos de encuestas son especialmente útiles e importantes cuando los metadatos incluyen una descripción detallada de la encuesta propiamente dicha (información sobre el título, autor principal, muestreo, fecha de creación de los datos, cobertura temática, cobertura geográfica, etc.), pero también de cada una de las variables (nombre y etiqueta de las variables, categorías, formulación de las preguntas, instrucciones para los encuestadores, definiciones).

  Un catálogo con estas características puede elaborarse de forma relativamente sencilla utilizando el estándar de metadatos DDI y las herramientas informáticas de la IHSN, especialmente el Microdata Management Toolkit y la plataforma Archivo Nacional de Datos (ANDA y NADA por sus siglas en inglés) (disponibles en www.ihsn.org).

- Permite realizar búsquedas en todos los campos de la operación estadística. En el marco del estándar DDI, esto significa que el catálogo debe permitir realizar búsquedas a nivel de la operación estadística (título, año, país, organización) pero también a nivel de las variables (nombre y etiqueta de la variable, etiqueta del valor de la variable). Es aconsejable que el catálogo ofrezca funciones de búsqueda en texto completo de fácil utilización para el usuario.

- Incluye información clara sobre la política y los procedimientos de acceso a los datos.

- Incluye una lista de los documentos de referencia, y accesos directos (cuestionarios, manuales, informes).

- Incluye una función de «búsqueda por palabra clave» siguiendo una taxonomía temática.

  Con objeto de facilitar el intercambio de información entre catálogos, la comunidad de repositorios de datos ha elaborado un tesauro para describir los temas cubiertos por los conjuntos de datos registrados en sus respectivos catálogos. Se trata de una lista de términos o conceptos utilizados para describir elementos concretos (conjuntos de datos, variables, libros, etc.). Generalmente, los términos de un tesauro se organizan en forma de árbol o en listas jerarquizadas (de los términos genéricos a las nociones más específicas). El tesauro suele incluir sinónimos y términos relacionados para que los usuarios encuentren lo que buscan aunque no utilicen los términos específicos.

  En muchas ocasiones se utiliza el tesauro para añadir palabras clave en el estudio, o conceptos en las variables. El empleo de un tesauro favorece la coherencia global, ya que de esa forma un mismo objeto se designa siempre con el mismo término. Además, la accesibilidad al tesauro por parte de los usuarios hace que estos puedan emplear los términos

y conceptos que les conducirán a la lista de resultados más adecuada.

Un ejemplo de la utilización del tesauro es el catálogo gestionado por el CESSDA (Consejo de Archivos Europeos de Datos de Ciencias Sociales).

- Permite mostrar los resultados de las búsquedas rápidamente, aunque sea un catálogo de gran volumen, lo que requiere un sistema de indexación eficaz.

- Incluye una herramienta para comparar los productos del catálogo. Es una función útil para comparar las variables entre distintas encuestas estándar o entre distintas versiones de una misma encuesta y que han sido cargadas en el catálogo.

- Proporciona información visible sobre las políticas de acceso aplicables a cada operación estadística (por ejemplo la disponibilidad de los microdatos y, en su caso, una descripción clara de sus modalidades de obtención).

- Incluye buenas herramientas de ayuda en línea.

- Incluye enlaces entre los productos del catálogo y los recursos disponibles en un sitio web externo, y permite añadir informaciones complementarias, como por ejemplo las referencias bibliográficas de las publicaciones utilizadas para el estudio.

- ***Desde el punto de vista del administrador del catálogo,*** un buen sistema de catalogación:

  - Ofrece un entorno seguro para el almacenamiento y el acceso a datos y metadatos.

  - Incluye herramientas para gestionar los sistemas de acceso a los microdatos (desde la aprobación automática para los microdatos accesibles sin restricciones hasta los sistemas de gestión y tratamiento para el acceso sometido a autorización).

  - Integra una solución para compartir los archivos de uso público y los archivos protegidos mediante licencia.

  - Constituye un medio seguro para compartir microdatos y documentación, con lo que se incrementa el acceso de los usuarios finales.

  - Recoge información sobre los usuarios del catálogo, sobre las descargas y, si procede, sobre los fines de la utilización de los datos. Este tipo de información es interesante para los auspiciantes de los operativos estadísticos, ya que les permite evaluar la utilización de los microdatos. También puede ser útil para los usuarios finales, ya que de esta forma pueden informarse sobre la publicación de nuevas versiones de los datos o la revisión de los estudios que han descargado.

Anonimización de los microdatos

La anonimización de los datos debe realizarla personal competente en el ámbito de la estadística, familiarizado con la utilización de software como Stata o SPSS. Existen algunos programas especializados que permiten medir o reducir el riesgo de divulgación. Sin embargo, ninguna de esas aplicaciones ofrece una solución integrada satisfactoria para los archivos complejos de datos jerarquizados. En la práctica, la anonimización sigue siendo en gran parte un proceso *ad hoc*. Actualmente están realizándose una serie de trabajos destinados a elaborar herramientas y directrices para facilitar la anonimización de los microdatos (entre otras organizaciones, la IHSN).

Como ya se ha visto, la anonimización de un archivo de una encuesta se realiza en dos pasos: en primer lugar hay que identificar los elementos que presenten potencialmente un riesgo de divulgación, y luego hay que aplicar técnicas de restricción o de perturbación de datos para reducir ese riesgo. Esta última etapa implica la intervención de alguien con suficiente conocimiento de la materia como para poder recomendar qué limitaciones aplicar a la información, de forma que afecten lo menos posible a los futuros usuarios de los archivos.

Preservación de los microdatos (y metadatos)

Los datos y metadatos digitales son vulnerables a la obsolescencia del software y hardware, a los peligros físicos y al error humano. La preservación de datos y metadatos a largo plazo requiere establecer una serie de procedimientos y una infraestructura adecuada. Los principios y buenas prácticas en materia de conservación de datos se describen con detalle en un documento de trabajo de la IHSN elaborado por el ICPSR (Consorcio Interuniversitario para la Investigación en Ciencias Políticas y Sociales) [8].

11. ¿Cuáles son los requisitos institucionales para la difusión de archivos de microdatos?

Para muchas ONE, la difusión de archivos de microdatos es una actividad muy reciente, así que puede conllevar novedades importantes en lo que respecta a la utilización de sus datos. En un documento titulado *OECD Principles and Guidelines for Access to Research Data from Public Funding* (Principios y directrices de la OCDE para el acceso a los datos de las investigaciones financiadas con fondos públicos), la OCDE define una serie de principios fundamentales destinados a los organismos proveedores de datos:

«Apertura

Por apertura se entiende el acceso en condiciones de igualdad de la comunidad científica internacional, al menor coste posible, que a poder ser no supere el coste marginal de la difusión. [...]

Flexibilidad

La flexibilidad supone tomar en cuenta la rápida y a menudo imprevisible evolución de las tecnologías de la información, [...], de los sistemas jurídicos y de las culturas de los distintos países. [...]

Transparencia

La información sobre los datos de investigación y las organizaciones productoras de datos, la documentación sobre los datos y las especificaciones sobre las condiciones que rigen su utilización deberían ser accesibles en el ámbito internacional, con total transparencia, idealmente por Internet. [...]

Conformidad legal

Los mecanismos de acceso a los datos deberían respetar los derechos e intereses legítimos de todos aquellos que intervienen en la investigación pública [...].

Protección de la propiedad intelectual

Los mecanismos de acceso a los datos deberían tener en cuenta la aplicabilidad de los derechos de autor u otras legislaciones sobre la propiedad intelectual que puedan afectar a las bases de datos de la investigación financiada con fondos públicos. [...]

Responsabilidad formal

Los mecanismos de acceso deberían promover prácticas institucionales explícitas y formalizadas, como la elaboración de normas y reglamentaciones, en lo que respecta a la responsabilidad de las distintas partes que intervienen en las actividades relativas a los datos. Estas prácticas deberían estar relacionadas con la autoría de los datos, la mención de los productores de datos, la propiedad, la difusión, las restricciones de utilización, las modalidades financieras, las normas éticas, las condiciones de licencia, la responsabilidad civil y el almacenamiento sostenible. [...]

Profesionalismo

Los mecanismos institucionales para la gestión de datos de investigación deberían basarse en la normativa profesional aplicable, así como en los valores reflejados en los códigos de conducta de los medios científicos correspondientes. (...)

Interoperabilidad

La interoperabilidad tecnológica y semántica es fundamental para facilitar y promover el acceso y la utilización de los datos de investigación en un contexto internacional e interdisciplinario. Los mecanismos de acceso deberían tener debidamente en cuenta los estándares internacionales aplicables en lo relativo a la documentación de los datos [...].

Calidad

El valor y la utilidad de los datos de investigación dependen en gran medida de la calidad de los datos per sé. Los administradores de datos y las organizaciones dedicadas al levantamiento de datos deberían velar especialmente y de manera explícita por el respeto de los estándares de calidad. [...]

Seguridad

Debería prestarse especial atención a promover la utilización de técnicas e instrumentos destinados a garantizar la integridad y seguridad de los datos. [...]

Eficiencia

Uno de los objetivos principales al promover el acceso y la utilización de los datos es mejorar la eficiencia global

del [levantamiento de datos] financiada con fondos públicos, para evitar así la duplicación inútil y costosa de las actividades relacionadas con la recopilación de datos. [...]

Rendición de cuentas

El funcionamiento de los mecanismos de acceso a los datos debería ser evaluado periódicamente por los grupos de usuarios, las instituciones responsables y los organismos financiadores [...]». [17]

El respeto de estos principios requerirá sin duda adoptar nuevos procedimientos y nuevas mentalidades. En sus directrices [17], la OCDE define las grandes cuestiones relacionadas con el acceso a los datos (cuestiones igualmente válidas para los microdatos creados por productores de datos oficiales con fines estadísticos):

- «Cuestiones institucionales y relacionadas con la gestión. Aunque mejorar la accesibilidad es importante para todas las comunidades científicas, la diversidad de las investigaciones sugiere que los enfoques en donde se satisfaga necesidades específicas en un marco en donde coexistan distintos modelos institucionales para la gestión de datos, serán sin duda más eficaces para responder a las necesidades de los investigadores.

- Cuestiones financieras y presupuestarias. La infraestructura de los datos científicos requiere una planificación presupuestaria específica, continua y el apoyo financiero necesario. El uso de estos datos no podrá optimizarse si los costes de acceso, gestión y preservación no son considerados en la planificación o son añadidos en los proyectos de investigación. Sin embargo, hay que destacar también que los costes de almacenamiento y gestión de datos han disminuido notablemente durante los últimos años, y la falta de información sobre tales cambios puede suponer en sí misma un obstáculo para avanzar en este ámbito.

- Cuestiones jurídicas y políticas. Las legislaciones nacionales y los acuerdos internacionales, especialmente en los ámbitos de los derechos de la propiedad intelectual y la protección de la vida privada, influyen directamente en las prácticas de acceso y difusión de los datos y deben tomarse en cuenta sin excepción en el proceso de elaboración de los mecanismos de acceso a los datos.

- Cuestiones culturales y de comportamiento. Para promover las prácticas de acceso e intercambio de datos es necesario establecer mecanismos adecuados de formación y de retribución. Estas consideraciones afectan a todos aquellos que financian, producen, gestionan y utilizan los datos [...]».

[...] «La responsabilidad sobre los distintos aspectos que engloba el acceso a los datos y su gestión debería quedar establecida en una serie de documentos, como por ejemplo la descripción de las actividades oficiales de las instituciones, las aplicaciones para la obtención de subvenciones, los contratos de investigación, los acuerdos de publicación y las licencias». [17]

Los repositorios digitales seguros como alternativa

Para algunos productores de datos, el establecimiento y gestión de un repositorio de datos y de un servicio de difusión es un objetivo poco realista, entre otras cosas por razones presupuestarias y jurídicas. Una de las alternativas consiste en delegar esta tarea a un repositorio ya existente. Como el UKDA, en la Universidad de Essex, que gestiona y difunde los datos de institutos de estadística, centros de investigación y los propios investigadores. O el CPSR (Consorcio Interuniversitario para la Investigación en Ciencias Políticas y Sociales), en la Universidad de Michigan, que tiene una función similar en los Estados Unidos.

Estos archivos de datos no solo garantizan la gestión eficaz de la concesión de licencias, sino que también desempeñan un papel de primer orden en la preservación de los datos y en materia de innovación. Ejemplo de ello son las nuevas páginas web del UKDA, que sirven como guía para la gestión y el intercambio de datos. Se trata de proponer a los creadores, gestores y archivistas de datos las mejores estrategias y métodos de creación, preparación y almacenamiento de los conjuntos de datos compartidos.

12. ¿Cómo promover la utilización de archivos de microdatos?

¿Bastan los archivos de microdatos de uso público y los archivos protegidos mediante licencia para constituir una base sólida de usuarios de dichos productos? Desafortunadamente, parece que no. Hay que persuadir a los usuarios para que se involucren, es decir, deben ser formados en este sentido.

Las encuestas nacionales por muestreo y los archivos de datos de uso público generados a partir de censos de población resultan sin embargo de interés para un amplio conjunto de investigadores y analistas. Con una documentación y una publicidad adecuadas, la utilización de estos conjuntos de datos debería incrementarse. El programa mundial de encuestas demográficas y de salud (*Demographic and Health Survey,* DHS) pone de manifiesto la gran demanda existente de este tipo de conjuntos de datos. El acceso a los conjuntos de datos del programa DHS es gratuito y sencillo. Descargados por un número considerable de usuarios, se han utilizado para una amplia y diversificada serie de estudios y publicaciones.

En los países habituados a crear este tipo de archivos, los debates sobre su adecuada utilización son habituales. Sin embargo, en otros lugares parece que no siempre queda clara la utilidad de los archivos de microdatos. Los centros que empiezan a difundir este tipo de archivos deberán sin duda recurrir a diversos medios para promover su utilización e informar a los potenciales usuarios del valor que suponen, así como de sus limitaciones.

La cultura del intercambio de datos y de la cooperación debería desembocar en un incremento de nuevos conocimientos. Se aconseja encarecidamente a los INEs y a sus colaboradores que promuevan la utilización de archivos de microdatos tanto en el ámbito nacional como en el internacional, especialmente a través de seminarios y eventos de capacitación. No faltan oportunidades para hacerlo.

Los archivos de microdatos tienen gran importancia en los campos de la investigación y la educación. Desempeñan un papel fundamental en la elaboración de políticas y programas, tanto para los gobiernos nacionales como para las organizaciones internacionales, que deben considerarse aliados naturales de los INEs para la promoción de una utilización apropiada de los archivos de microdatos. Por otro lado, la función de las universidades en la formación de los nuevos usuarios es también esencial.

Para garantizar el éxito de la difusión de los microdatos de los INEs es indispensable que las personas interesadas conozcan de la existencia del producto y sus ventajas y, por lo tanto, que se organicen campañas de sensibilización. Para ello hay que identificar y acercarse a las organizaciones y miembros de estas toda vez que son potenciales usuarios. En muchos casos, los INEs son conscientes de esta necesidad y han constatado la demanda en este sentido por parte de los usuarios. Entre otras medidas, pueden introducirse enlaces hacia otros sitios web, realizarse folletos e invitar a los usuarios a cursos o seminarios.

Como ya se ha dicho, es fundamental formar a los potenciales usuarios. La Iniciativa para la Democratización de los Datos canadiense (IDD) organiza sesiones de formación destinadas a promover y sostener la utilización de los archivos de microdatos y otros productos para la investigación y la enseñanza. En Canadá ya había muchos investigadores familiarizados con los archivos de microdatos, pero no ocurría lo mismo con los centros y organismos encargados de su promoción. Así que se puso en marcha un sistema de listas de distribución de correo electrónico para que los investigadores, el personal de los INEs, los intermediarios y otras personas interesadas pudieran hacer preguntas y compartir experiencias. El archivo con las respuestas a esas preguntas constituye una fuente de información inestimable valor y ayuda a los INEs, que de esta forma no necesitan responder a todas las preguntas que reciben.

Una de las claves del éxito de la difusión de archivos de datos entre los investigadores es la interacción del personal de los INEs con la red de investigadores y archivistas. Esto ayuda a ambas partes a identificar sus necesidades y problemáticas y sienta las bases para una comunicación en doble sentido.

Anexo 1

Solicitud de acceso a un conjunto de datos protegidos mediante licencia en el marco de un proyecto de investigación concreto

Este modelo de formulario debe adaptarse a las necesidades específicas de cada caso.

La información que se proporcione en el presente formulario será tratada de forma confidencial, salvo incumplimiento del acuerdo legal concluido, en cuyo caso el INE podrá comunicarla a los institutos de estadística colaboradores de otros países.

Por favor, envíe el formulario debidamente llenado por correo o por fax a la dirección o número siguientes, acompañado de una carta de motivación con el membrete del organismo responsable:

Dirección: [dirección]
Fax: [número de fax]
E-mail (copia escaneada): [dirección de correo electrónico]

Título y número de referencia del o de los conjuntos de datos solicitados (título exacto, año y número de referencia que figuran en nuestro catálogo de encuestas):

_____

_____

Términos del presente acuerdo

En el presente acuerdo,

1. «Investigador principal» se refiere al principal interlocutor designado para toda comunicación relativa al presente acuerdo. El investigador principal asume la responsabilidad de respetar los términos del presente acuerdo, que rigen el acceso a los datos. El investigador principal debe estar habilitado para representar al organismo beneficiario en el marco del presente acuerdo.

2. «Otros investigadores» se refiere a las personas distintas del investigador principal, incluidos los ayudantes, que tendrán acceso a los datos confidenciales.

3. «Organismo beneficiario» se refiere a la organización/universidad/entidad que emplea al investigador principal.

Sección A. Investigador principal

- Nombre _____
- Apellidos _____
- Título _____
- Organización _____
- Función/cargo en la organización_____
- Dirección postal _____
- Teléfono (con código de país) _____
- Fax (con código de país) _____
- E-mail _____

Sección B. Otros investigadores

Indique los nombres, cargos y organismos de adscripción de todos los demás miembros del equipo de investigadores que accederán a los datos confidenciales.

Nombre y apellidos _____
Función/cargo _____
Organismo de adscripción _____

Sección C. Organismo beneficiario

Nombre de la organización _____

Tipo de organización (marque la respuesta correspondiente)

□ Administración pública / ministerios del ramo
□ Universidad
□ Centro de investigación
□ Empresa privada
□ Organización internacional
□ Organización no gubernamental (nacional)
□ Organización no gubernamental (internacional)
□ Otros (especificar) _____

Sitio web de la organización (URL) _____
Dirección postal _____

Sección D. Descripción de la utilización prevista de los datos

Por favor describa el proyecto de investigación (temas, objetivos, métodos, resultados previstos, colaboradores). Si la información proporcionada es insuficiente, la solicitud corre el riesgo de ser rechazada, o quizás recibirá una petición de información adicional. Esta información puede presentarse a modo de anexo junto a la presente solicitud.

Lista del o de los resultados esperados y política de difusión

_____

_____

_____

Sección E. Identificación de los archivos de datos y de las variables que se necesitan

En el sitio web del INE puede accederse a metadatos detallados. Incluyen una descripción de los archivos de datos y de las variables que constituyen cada conjunto de datos. Se solicita a los investigadores que no necesiten consultar el conjunto de datos completo que indiquen el subconjunto de variables o de casos que sean de su interés. De este modo, se reduce el riesgo de divulgación y se aumenta la probabilidad de obtener los datos solicitados.

Con esta solicitud se quiere tener acceso (marque la respuesta correspondiente):

□ Al conjunto de datos completo (todos los archivos y casos).

□ A un subconjunto de variables y/o de casos, según las indicaciones siguientes (nota: las variables como los factores de expansión y los identificadores de los registros son sistemáticamente incluidas en los subconjuntos).

Sección F. Acuerdo de acceso a los datos

El investigador principal y los otros investigadores aceptan someterse a las siguientes condiciones:

1. Los datos confidenciales solo podrán ser consultados por el investigador principal y los otros investigadores designados en el presente acuerdo.

2. Se prohíbe, salvo autorización expresa del INE, toda reproducción o comunicación de los datos confidenciales o de cualquier dato basado en los datos iniciales a personas distintas de las mencionadas en el presente acuerdo.

3. Los datos serán exclusivamente utilizados con fines estadísticos o en el marco de trabajos de investigación. Servirán únicamente para crear datos agregados y no para investigar a a individuos específicos u organizaciones. Los datos no se utilizarán, de ningún modo, con fines administrativos, personales o para hacer cumplir la ley.

4. El investigador principal se compromete a que nadie utilice los datos proporcionados para identificar a un individuo, una familia, una empresa o una organización. No podrá hacerse uso de la identidad de una persona u organización que se haya descubierto fortuitamente. Todo descubrimiento de este tipo deberá señalarse inmediatamente al INE, y no podrá revelarse a nadie que no figure en el presente acuerdo de acceso a los datos.

5. El investigador principal tomará las medidas de seguridad oportunas para evitar el acceso no autorizado a los microdatos protegidos mediante licencia suministrados por el INE. Estos microdatos deberán destruirse obligatoriamente una vez que se hayan finalizado los trabajos de investigación, a menos que se proporcione al INE una garantía suficiente de seguridad de los datos y a condición de que esta autorice su conservación por escrito. El investigador principal deberá confirmar por escrito al INE la destrucción de los microdatos.

6. En todos los libros, artículos, documentos de trabajo, tesis, disertaciones, informes u otras publicaciones basadas en los datos proporcionados por el INE deberá citarse la fuente de esos datos,

de conformidad con la obligación de cita asociada al conjunto de datos proporcionado.

7. Se mandará al INE una copia en formato digital de todas las publicaciones basadas en los datos solicitados.

8. El INE y los organismos financiadores correspondientes se deslindan de toda responsabilidad respecto a la utilización, la interpretación o las conclusiones derivadas de los datos proporcionados.

9. El presente acuerdo entrará en vigor en la fecha en que se concedan los derechos de acceso al conjunto de datos confidencial, y será de aplicación hasta la fecha de finalización del proyecto, o hasta una fecha anterior si el proyecto termina anticipadamente.

10. El investigador principal deberá solicitar el acuerdo previo del INE en caso de modificación de las especificaciones del proyecto o los dispositivos de seguridad, así como para cualquier cambio relativo al personal o a la organización mencionados en el presente formulario de solicitud de acceso. Cualquier cambio en la organización que emplea al investigador principal implicará la presentación de una nueva solicitud y pondrá fin al proyecto inicial.

11. Ante el incumplimiento de cualquiera de las disposiciones establecidas en el presente acuerdo, el INE emprenderá acciones legales contra las personas consideradas responsables de la infracción, sea esta intencionada o no. El incumplimiento de las consignas del INE se considerará una violación grave del presente acuerdo y podrá contraer acciones judiciales. El INE mantendrá y compartirá con otros repositorios de datos colaboradores una lista de personas que han incumplido las disposiciones del acuerdo que rige el acceso a los datos; estas personas no podrán acceder a los datos en el futuro.

Signatarios

El investigador principal o un representante habilitado del organismo beneficiario manifiesta haber leído y acepta las disposiciones del acuerdo de acceso a los datos, que se especifican en la sección F del presente documento:

Nombre y apellidos _____

Firma _____
Fecha _____

Solicitud examinada por ……, el ……. [fecha]

Decisión del comité:

□ Solicitud aceptada
□ Solicitud rechazada [motivo] _____
□ Solicitud de información complementaria_____

Anexo 2

Modelo de política de acceso a un centro de datos seguro

Esta propuesta de texto habrá que adaptar según el país.

Objetivos

El Centro Nacional de Datos Seguro («Centro») fue creado por el archivo nacional de datos para que los investigadores con determinadas cualificaciones pudieran acceder a los archivos de microdatos estadísticos confidenciales, bajo una estricta vigilancia. El Centro pone a disposición de estos investigadores un mecanismo que les permite consultar archivos de datos detallados con total seguridad, sin poner en peligro el anonimato de los encuestados.

Ubicación

El Centro se encuentra ubicado en [dirección, teléfono, fax, e-mail y sitio web].

Actividades del Centro

Los investigadores pueden acceder a los datos in situ, bajo la supervisión del personal del Centro. Se les proporciona un ordenador, así como el software adecuado y un espacio de trabajo.

Datos

- El personal del Centro prepara los archivos de datos antes de la llegada del investigador y se asegura de que ningún dato confidencial salga del Centro.

- Los investigadores que deseen realizar diversos análisis y que para ello necesiten diversos conjuntos de datos, solo podrán consultar un conjunto de datos a la vez. Los investigadores no podrán en ningún caso fusionar conjuntos de datos por iniciativa propia.

- El Centro autoriza a los investigadores a aportar sus propios datos anonimizados y a asociarlos a los conjuntos de datos del Centro para crear conjuntos de datos fusionados, que se almacenarán en el Centro. Los datos aportados por los investigadores deben haber sido recogidos y ser «propiedad» de los propios investigadores, o bien tratarse de otros datos públicos legalmente obtenidos por dichos investigadores. Los investigadores deberán proporcionar OBLIGATORIAMENTE al Centro documentación completa sobre todos los datos destinados a fusionarse con datos del Centro. Los investigadores que deseen realizar esta operación deberán dirigirse al personal del Centro para asegurarse de que sus datos pueden efectivamente fusionarse con los del Centro. El Centro acepta archivos de datos en formato SAS, SPSS y Stata.

- El Centro realiza regularmente copias de seguridad de todos los archivos. Estas copias de seguridad se guardan en un lugar seguro, cuyo acceso está reservado al personal del Centro. Sin embargo, pueden ponerse a disposición de los investigadores que vuelvan para realizar análisis complementarios. Estos archivos de seguridad contienen los datos proporcionados por los usuarios, así como los archivos fusionados y serán destruidos si el usuario así lo solicita mediante petición escrita.

Equipo y materiales de trabajo

- El Centro dispone de [cantidad] puestos de trabajo para los usuarios, así como de una impresora láser en blanco y negro instalada en una sala de acceso seguro. Las computadoras del Centro no están conectadas a Internet y están configurados de forma que el usuario no pueda utilizar ningún dispositivo extraíble (CD-ROM, DVD, disquete o dispositivo USB).

- Los puestos de trabajo del Centro están equipados con un procesador [Pentium X XXX MHz] y funcionan con [Windows NT / Otro].

Software

- Además del paquete MS-Office, en los puestos de trabajo están instalados CSPro, EPI-Info, SAS, SPS y Stata. Pueden añadirse otros lenguajes analíticos/de programación si así se solicita. Contacte con el Centro para más información sobre las versiones de las aplicaciones disponibles.

- Los investigadores deben tener los conocimientos necesarios para realizar sus propios análisis con las aplicaciones proporcionadas. El Centro no ofrece servicio de asistencia técnica para estos programas.

Espacio de trabajo

- Los investigadores deben trabajar bajo la supervisión del personal del Centro, y solo durante los horarios de oficina habituales (de lunes a viernes, de las 8.30 h a las 17.00 h).

- El acceso al Centro está reservado a los investigadores indicados en el formulario de solicitud. En la entrada se les pedirá un documento de identidad con fotografía.

- Pueden ocupar el mismo puesto de trabajo un máximo de tres investigadores que trabajen en el mismo proyecto.

- El Centro adjudica los puestos de trabajo por orden de llegada.

Supervisión del personal del Centro (con objeto de controlar la divulgación)

- Los investigadores externos no pueden traer consigo documentos, manuales, libros, etc., que puedan permitirles identificar y divulgar informaciones personales disponibles en el Centro. Asimismo, se prohíbe el uso de teléfonos móviles, u otras herramientas de comunicación con el exterior.

- Los investigadores no grabarán ningún resultado, archivo o programa en un dispositivo de almacenamiento extraíble. El personal del Centro se encargará de ello en su caso.

- Los investigadores solo podrán llevarse los resultados de sus análisis tras pasar el control de divulgación efectuado por el personal del Centro. Estos controles de divulgación consisten en buscar las celdas con menos de cinco casos, las tablas con variables geográficas, los modelos con variables geográficas (o variables equivalentes a variables geográficas) o listas de casos.

- Todos los archivos de registro (o logs) deben imprimirse o archivarse en formato digital. Los conservará el Centro, que solo guardará los programas y procedimientos ejecutados por los investigadores externos. Los archivos de registro correspondientes a sus propias investigaciones no se conservarán.

- Todos los resultados generados por los programas estadísticos y todas las notas manuscritas relativas a estos resultados serán sometidos a un control de divulgación por parte del personal del Centro antes de salir de este. Los resultados se limitan a tablas sintéticas. En ningún caso una tabla debe incluir celdas con menos de cinco casos observados. De ser así, esas celdas se suprimirán, generalmente borrándolas. Para que no puedan reconstituirse a partir de otras celdas de la misma línea o columna, el personal suprime los totales de las líneas y columnas correspondientes a esas celdas. Una vez finalizado el control de divulgación, se da a los investigadores una fotocopia de las tablas definitivas. El personal del Centro aplica las mejores prácticas en vigor para determinar si los datos tabulados son identificables y toma las decisiones más prudentes. Las decisiones del Centro son irrevocables y no pueden ser discutidas ni negociadas por parte de los investigadores.

Costes de admisión

Los usuarios del Centro deberán satisfacer una serie de costes correspondientes al alquiler del espacio y el equipamiento, al tiempo de supervisión del personal, al control de divulgación y al mantenimiento de los equipos informáticos (hardware y software), así como a la creación y gestión de los archivos de datos solicitados por el investigador. Las tarifas de acceso al Centro se muestran a continuación:

Organismo de quien depende el investigador principal	Costes de preparación y creación de los archivos (costes fijos)	Utilización del equipamiento (por día y por puesto de trabajo)
Usuarios nacionales		
Personal del organismo miembro del Centro	Gratuito	Gratuito
Otros organismos públicos	[Tarifa/moneda]	[Tarifa/moneda]
Universidad o centro de investigación	[Tarifa/moneda]	[Tarifa/moneda]
ONG	[Tarifa/moneda]	[Tarifa/moneda]
Usuarios internacionales		
Investigación en colaboración con el Centro	[Tarifa/moneda]	[Tarifa/moneda]
Organización internacional	[Tarifa/moneda]	[Tarifa/moneda]
Universidad o centro de investigación	[Tarifa/moneda]	[Tarifa/moneda]
ONG	[Tarifa/moneda]	[Tarifa/moneda]

En caso de operaciones especiales (fusión de datos adicionales, creación de formatos de archivos personalizados o incluso la adquisición e instalación de software específico no estándar) podrá facturarse un importe adicional. Este importe se acordará entre el investigador y el personal del Centro. Los pagos se realizarán por adelantado, antes de utilizar el Centro.

Los pagos se efectuarán a: [información relativa a la modalidad de pago]

Envío de las propuestas de proyectos de investigación

Los investigadores utilizarán el formulario de la página siguiente para someter sus proyectos a la aceptación del Centro. Se aconseja a los investigadores candidatos que, antes de redactar el proyecto, se dirijan al personal del Centro para comprobar que los datos que les interesan estén efectivamente disponibles. El proyecto debe describirse de forma que ayude al personal del Centro a crear los archivos analíticos necesarios. Deberán precisarse claramente las variables necesarias y, si corresponde, la selección de casos requerida. El archivo de datos analítico solo contendrá los elementos indispensables para la realización de los análisis solicitados. El solicitante deberá indicar para qué requiere los datos solicitados. Los proyectos muy grandes y complejos o, inversamente, los proyectos pequeños pueden requerir un mayor intercambio de información entre el personal del Centro y los solicitantes, lo que puede demorar el proceso. El trabajo de preparación de los archivos de datos podrá realizarse en un plazo razonable si los grandes proyectos se dividen en varias partes y si los datos necesarios se definen claramente.

Los investigadores que deseen fusionar datos del Centro con datos externos deberán proporcionar estos últimos previamente al personal del Centro.

El proyecto de investigación será evaluado desde su recepción por un comité de control reunido para tal efecto.

Para evaluar el proyecto, se tendrán en cuenta los siguientes criterios:

- Factibilidad técnica y científica del proyecto.
- Disponibilidad de los recursos del Centro.
- Riesgo de divulgación de la información confidencial.

Téngase en cuenta que, al aceptar la solicitud, el Centro no está aprobando la pertinencia general del proyecto, su metodología ni sus teorías subyacentes, así como tampoco está reconociendo ningún mérito ni valor en particular.

La aprobación del Centro constituye una simple apreciación de la legalidad de la utilización del archivo de datos en el marco de las investigaciones descritas e indica que el proyecto podrá probablemente completarse en el Centro.

Anexo 3

Solicitud de acceso a un centro de datos seguro

Este modelo de formulario debe adaptarse a las necesidades específicas de cada caso.

La información que se proporcione en el presente formulario será tratada de forma confidencial, salvo incumplimiento del acuerdo legal concluido, en cuyo caso el Centro podrá comunicarla a los institutos de estadística colaboradores de otros países.

Por favor envíe el formulario debidamente llenado por correo o por fax a la dirección o número siguientes, acompañado de una carta de motivación con el membrete del organismo responsable:

Dirección: [dirección]
Fax: [número de fax]
E-mail (copia escaneada): [dirección de correo electrónico]

Título y número de referencia del o de los conjuntos de datos solicitados (título exacto, año y número de referencia que figuran en nuestro catálogo de encuestas):

_____

_____

Términos del acuerdo

En el presente acuerdo,

1. «Investigador principal» se refiere al principal interlocutor designado para toda comunicación relativa al presente acuerdo. El investigador principal asume la responsabilidad de respetar las disposiciones del presente acuerdo, que rigen el acceso a los datos. El investigador principal debe estar habilitado para representar al organismo beneficiario en el marco del presente acuerdo.

2. «Otros investigadores» se refiere a las personas distintas del investigador principal, incluidos los ayudantes, que tendrán acceso a los datos confidenciales.

3. «Organismo beneficiario» se refiere a la organización/universidad/entidad que emplea al investigador principal.

4. «Representante del organismo beneficiario» se refiere a una persona habilitada para representar al organismo beneficiario en el marco del presente acuerdo.

Sección A. Investigador principal

- Nombre _____
- Apellidos _____
- Título _____
- Organización _____
- Función/cargo en la organización _____
- Dirección postal _____
- Teléfono (con código de país) _____
- Fax (con código de país) _____
- E-mail _____

Sección B. Otros investigadores

Indique los nombres, cargos y organismos de adscripción de todos los demás miembros del equipo de investigadores que accederán a los datos confidenciales.

- Nombre y apellidos _____
- Función/Cargo _____
- Organismo de adscripción _____

Adjuntar a la presente solicitud un resumen de la trayectoria profesional o el CV de cada una de las personas que participen en las investigaciones, indicando su nacionalidad.

Sección C. Organismo beneficiario

Nombre de la organización _____

Tipo de organización (marque la respuesta correspondiente)

□ Administración pública / ministerios del ramo. . Universidad
□ Centro de investigación
□ Empresa privada
□ Organización internacional
□ Organización no gubernamental (nacional)
□ Organización no gubernamental (internacional)

□ Otros (especificar) _____

Sitio web de la organización (URL)_____
Dirección postal_____

Sección D. Representante del organismo beneficiario

- Nombre _____
- Apellidos _____
- Título _____
- Prof./Dr./Srta./Sra./Sr. _____
- Organización _____
- Función/cargo en la organización _____
- Dirección postal _____
- Teléfono (con código de país) _____
- Fax (con código de país) _____
- E-mail _____

Sección E. Descripción de la utilización prevista de los datos

Por favor, describa el proyecto de investigación (temas, objetivos, métodos, resultados previstos, colaboradores). Indique por qué los conjuntos de datos públicos no responden plenamente a sus necesidades. Si la información proporcionada es insuficiente, la solicitud corre el riesgo de ser rechazada, o quizás recibirá una petición de información adicional. Esta información puede presentarse a modo de anexo junto a la presente solicitud.

Lista del o de los resultados esperados y política de difusión.

¿Tiene previsto fusionar el conjunto de datos con otros datos? □ SÍ □ NO

En caso afirmativo, indique la referencia de los demás conjuntos de datos a ser fusionados.

_____
_____
_____
_____
_____

Sección F. Identificación de los archivos de datos y de las variables que se necesitan

En el sitio web del INE puede accederse a metadatos detallados ncluyendo una descripción de los archivos de datos y de las variables que constituyen cada conjunto de datos. Se solicita a los investigadores que no necesiten consultar el conjunto de datos completo que indiquen el subconjunto de variables o de casos que les interesen para que el Centro pueda preparar los archivos de datos.

Con esta solicitud se quiere tener acceso (marque la respuesta correspondiente):

□ Al conjunto de datos completo (todos los ficheros y casos).
□ A un subconjunto de variables y/o de casos, según las indicaciones siguientes (nota: las variables como los factores de expansión y los identificadores de los registros son sistemáticamente incluidas en los subconjuntos).

Sección G. Software necesario

Los investigadores utilizarán el siguiente software:

☐ CSPro ☐ SAS ☐ SPSS ☐ Stata

Otros (precisar): _____

Observaciones:

- El Centro actualiza regularmente su software. Por favor, contáctenos para más información sobre la versión disponible de cada aplicación.
- Los investigadores que requieran utilizar software que no figure en la lista del software estándar proporcionado por el Centro deberán suministrar la correspondiente licencia de uso, en vigor. El personal del Centro instalará el software para su utilización durante los trabajos de investigación (la licencia de uso continuará siendo propiedad del investigador). Recomendamos contactar con el Centro antes de finalizar la solicitud para comprobar la factibilidad técnica del proyecto.

Sección H. Acuerdo de acceso a los datos

Con la condición de que sea aprobado por ambas partes, se firmará el siguiente acuerdo:

El investigador principal, los otros investigadores y el representante del organismo beneficiario aceptan las siguientes condiciones:

1. 1. Los datos confidenciales solo podrán ser consultados por el investigador principal y los otros investigadores designados en el formulario de solicitud, que firmarán un acuerdo de confidencialidad.

1. 2. Los datos serán exclusivamente utilizados con fines estadísticos. Servirán únicamente para crear datos agregados; queda prohibida toda investigación sobre particulares u organizaciones individuales. Los datos no se utilizarán, de ningún modo, con fines administrativos, personales o para hacer cumplir la ley.

1. 3. El investigador principal se compromete a que nadie utilice los datos proporcionados para identificar a un individuo, una familia, una empresa o una organización. No podrá hacerse uso de la identidad de una persona u organización que se haya descubierto fortuitamente. Todo descubrimiento de este tipo deberá señalarse inmediatamente al INE y no podrá revelarse a nadie que no figure en el presente acuerdo de acceso a los datos.

1. 4. En todos los libros, artículos, documentos de trabajo, tesis, disertaciones, informes u otras publicaciones basadas en los datos proporcionados por el INE deberá citarse la fuente de estos datos, de conformidad con la obligación de cita asociada al conjunto de datos proporcionado.

1. 5. Se mandará al Centro una copia en formato digital de todas las publicaciones basadas en los datos solicitados.

1. 6. El recopilador inicial de los datos, el Centro y los organismos financiadores se deslindan de toda responsabilidad respecto a la utilización, la interpretación o las conclusiones derivadas de los datos proporcionados.

1. 7. Ante el incumplimiento de cualquiera de las disposiciones establecidas en el presente acuerdo, el Centro emprenderá acciones legales contra las personas consideradas responsables de la infracción, sea esta intencionada o no. El incumplimiento de las consignas del Centro se considerará una violación grave del presente acuerdo y podrá contraer acciones judiciales. El Centro mantendrá y compartirá con otros repositorios de datos colaboradores una lista de personas que han incumplido las disposiciones del acuerdo que rige el acceso a los datos; estas personas no podrán acceder a los datos en el futuro.

1. 8. El Centro se reserva el derecho de poner fin a un proyecto en cualquier momento si considera que las actividades de un investigador pueden comprometer la confidencialidad o las normas deontológicas que se han de respetar en un contexto de investigación.

1. 9. Ningún documento impreso, archivo electrónico u otro documento o soporte saldrá del Centro sin haber sido previamente supervisado por el personal del Centro, con objeto de eliminar todo riesgo de divulgación.

1. 10. Podrá prohibirse definitivamente la entrada al Centro al investigador principal y los otros investigadores si el director del Centro lo considera necesario para proteger la integridad o la confidencialidad del mismo.

Signatarios

Los abajo firmantes manifiestan haber leído y aceptan las disposiciones del acuerdo de acceso a los datos, que se especifican en la sección H del presente documento:

El investigador principal

Nombre y apellidos _____
Firma _____Fecha _____

El representante del organismo beneficiario

Nombre y apellidos _____
Firma _____Fecha _____

El Centro espera que todos los investigadores apliquen las normas y los principios relativos a la investigación estadística durante la realización de sus trabajos. Solo se podrán realizar los análisis que hayan sido aprobados. El incumplimiento de estas consignas resultará en la anulación del proyecto de investigación y una posible prohibición de acceso al centro en el futuro.

Bibliografía

[1] Altman, M. y King, G. 2006. *A Proposed Standard for the Scholarly Citation of Quantitative Data*. http://gking.harvard.edu/files/cite.pdf

[2] Boyko, E. y Watkins, W. 2003. *Sécurité des données, sécurité des environnements : les deux sont indispensables*. En Eurostat. XIX Seminario del CEIES (Soluciones innovadoras para el acceso a los microdatos), Lisboa, 26 y 27 de septiembre de 2002, págs. 109-118.
www.cnis.fr/Agenda/CR/CR_0118.PDF

[3] CENEX-SDC. 2007. *Handbook on Statistical Disclosure Control*. http://neon.vb.cbs.nl/cenex/
CENEX-SDC_Handbook.pdf

[4] Dupriez, O. y Greenwell, G. 2007. *Quick Reference Guide for Data Archivists*. Documento de la IHSN. http://www.ihsn.org/home/download.php?file=DDI_IHSN_Checklist_OD_06152007.pdf

[5] Eurostat. 2009. *Work Session on Statistical Data Confidentiality*, Manchester, 17-19 de diciembre de 2007. En «Methodologies and Working Papers». http://www.unece.org/stats/publications/ Proceedings_statistical_data_confidentiality.pdf

[6] Hamilton, E. y Humphrey, C. 2000. *Measuring the Impact of DLI: Use of the NPHS Public Use Microdata File in Academic Outcomes*.

[7] Hamilton, E. y Humphrey, C. 2002. *L'IDD et l'ENSP :* É*tude de compatibilité*. Otoño de 2002. http://www.statcan.gc.ca/dli-ild/doc/update-bulletin-v52-fra.pdf

[8] ICPSR (Consorcio Interuniversitario para la Investigación en Ciencias Políticas y Sociales). 2009. *Principes et bonnes pratiques en matière de conservation de données*. Documento de trabajo n.°003 de la IHSN (Red Internacional de Encuestas de Hogares). Diciembre de 2009.
http://www.ihsn.org/home/index.php?q=focus/principles-and-good-practicepreserving-data

[9] ISO/IEC. 1999. *ISO/IEC 11179-1 – Information technology– Specification and standardization of data element – Part 1 : Framework for the specification and standardization of data elements*. http://metadata stds.org/11179-1/ISO-IEC_11179-1_1999_IS_E.pdf

[10] King, Gary. 1995. «Replication, Replication», *PS: Political Science and Politics*. Con los comentarios de 19 autores.
http://gking.harvard.edu/files/replication.pdf

[11] King, Gary. 1995. «A Revised Proposal, Proposal», *PS: Political Science and Politics,* vol. 28, n.º 3, septiembre de 1995, págs. 443-499. http://gking.harvard.edu/files/abs/replicationabs.shtml

[12] Lambert, D. 1993. «Measures of Disclosure Risk and Harm», *Journal of Official Statistics*, vol. 9, págs. 407-426.

[13] Madsen, P. 2003. *The Ethics of Confidentiality: The Tension Between Confidentiality and the Integrity of Data Analysis in Social Science Research*. Mimeo. Carnegie Mellon University (Estados Unidos), junio de 2003.

[14] NCHS (Centro Nacional de Estadísticas de Salud de EE. UU.). 2002. *Policy on Micro-data Dissemination*.
http://www.cdc.gov/nchs/data/NCHS%20Micro-Data%20Release%20Policy%204-02A.pdf

[15] NCHS (Centro Nacional de Estadísticas de Salud de EE. UU.). 2004. *NCHS Staff Manual on Confidentiality*. http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

[16] NCHS (Centro Nacional de Estadísticas de Salud de EE. UU.) - Research Data Center. 2008. *Guidelines for Proposal Submission*.
http://www.cdc.gov/nchs/data/r&d/guidelines_10_14_08c.pdf

[17] OCDE (Organización para la Cooperación y el Desarrollo Económicos). 2007. *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*.
http://www.oecd.org/dataoecd/9/60/38500823.pdf

[18] Statistique Canada. *Comment citer les produits de Statistique Canada*.
http://www.statcan.gc.ca/pub/12-591-x/12-591-x2006001-fra.htm (sitio consultado el 22 de junio de 2010).

[19] Tambay, J. L., Goldmann, G. y White, P. 2001. *Providing Greater Access to Survey Data for Analyses at Statistics Canada*. Ponencias del congreso anual de la American Statistical Association.

[20] UKDA (Centro de Almacenamiento de Datos del Reino Unido), Universidad de Essex. 2002. *Good Practices in Data Documentation*. Versión revisada.
http://www.esds.ac.uk/news/goodPractice.pdf

[21] UKDA (Centro de Almacenamiento de Datos del Reino Unido), Universidad de Essex. 2009. *Managing and Sharing Data. A Best Practice Guide to Researchers*. Segunda edición.

http://www.data archive.ac.uk/news/ publications/ managingsharing.pdf

[22] UK Statistics Authority (Autoridad Estadística del Reino Unido). 2009. *Code of Practice for Official Statistics*. Edición 1.0. Enero de 2009. http://www.statisticsauthority.gov.uk/ assessment/ code-of-practice/code-of-practicefor-official-statistics.pdf

[23] CEPE-ONU (Comisión Económica para Europa de las Naciones Unidas). 2000. *Terminology on Statistical Metadata* (Conference of European Statisticians Statistical Standards and Studies N° 53). Ginebra. http://www.unece.org/stats/publications/53metadaterminology.pdf

[24] CEPE-ONU (Comisión Económica para Europa de las Naciones Unidas). Conference of European Statisticians. 2007. *Managing Statistical Confidentiality and Microdata Access: Principles and Guidelines of Good Practice*.

http://www.unece.org/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

[25] CEPE-ONU (Comisión Económica para Europa de las Naciones Unidas). 2009. *Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes*. http://www.unece.org/stats/publications/Confidentiality_aspects_data_integration.pdf

[26] CEPE-ONU (Comisión Económica para Europa de las Naciones Unidas) y Statistics Sweden (Oficina de Estadística de Suecia). 2003. *Statistical Confidentiality and Access to Microdata*. Ponencias presentadas en la Conferencia de Estadísticos Europeos de 2003.

http://www.unece.org/stats/publications/statistical.confidentiality.pdf

[27] UNSD (División de Estadística de las Naciones Unidas). 1994. Sexto principio de *Principes fondamentaux de la statistique officielle*. http://unstats.un.org/unsd/methods/statorg/FP-French.htm

[28] US Bureau of the Census (Oficina del Censo de EE. UU.) - Software and Standards Management Branch - Systems Support Division. 2008. *Survey Design and Statistical Methodology Metadata*. Washington D. C. Sección 3.4.4.

[29] US Federal Committee on Statistical Methodology. 2005. *Statistical Policy Working Paper 22 (Revised 2005) – Report on Statistical Disclosure Limitation Methodology*.

http://www.fcsm.gov/working-papers/spwp22.html

[30] Watkins, W. y Boyko, E. 1996. «Data Liberation and Academic Freedom», *Government Information in Canada/Information gouvernementale au Canada*, vol. 3, n°2.

http://www.usask.ca/library/gic/v3n2/watkins2/watkins2.html

[31] Watkins, W. *The Data Liberation Initiative: A New Cooperative Model*. Documento inédito escrito para la Canadian Social Science Federation.

http://library2.usask.ca/gic/v1n2/watkins/watkins.html

Sitios web

African Association of Statistical Data Archivists (AASDA) / Asociación Africana de Repositorios de Datos Estadísticos
http://www.aasda.net/home_dev/index.php

American Statistical Society, privacidad, confidencialidad y seguridad de los datos
http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main

Archivo de documentos de formación de la IDD (alojado en el *Ontario Universities Scholars Portal Economic and Social Data Service*)
https://ospace.scholarsportal.info/handle/1873/69

Australian Bureau of Statistics / Oficina Estadística de Australia (ABS)
http://www.abs.gov.au/

Central Statistics Office / Oficina Central de Estadísticas de Irlanda (CSO)
http://www.cso.ie/

Centro Latinoamericano y Caribeño de Demografía, CEPAL – Comisión de Estadística de las Naciones Unidas
http://www.cepal.org.ar/software/icepa8c.html

Council of European Social Science Data Archives / Consejo de Archivos Europeos de Datos de Ciencias Sociales (CESSDA)
http://www.cessda.org/

Data Documentation Initiative Alliance (DDI)
http://www.ddialliance.org

Data.Gov (Reino unido)
http://data.gov.uk

Data.Gov (Estados Unidos)
http://www.data.gov/

Department of Census and Statistics - Sri Lanka / Departamento de Estadísticas y Censos de Sri Lanka
http://statistics.sltidc.lk

División de Estadística de las Naciones Unidas
http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp

Dublin Core Metadata Initiative / Iniciativa de Metadatos Dublin Core (DCMI)
http://dublincore.org/

Economic and Social Data Service (ESDS)
http://www.esds.ac.uk/aandp/create/research.asp

European Social Survey / Encuesta Social Europea
http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=78&Itemid=190

Institute for Social and Economic Research, The British Household Panel Survey
http://www.iser.essex.ac.uk/ulsc/bhps/

International Household Survey Network / Red Internacional de Encuestas de Hogares (IHSN)
http://www.ihsn.org

Inter-university Consortium for Political and Social Research / Consorcio Interuniversitario para la Investigación en Ciencias Políticas y Sociales (ICPSR)
http://www.icpsr.umich.edu

Luxembourg Income Study (LIS)
http://www.lisproject.org/

Michigan Census Research Data Center / Centro de Investigaciones sobre el Censo de Michigan (MCRDC)
www.isr.umich.edu/src/mcrdc/

National Center for Health Statistics / Centro Nacional de Estadísticas de Salud de EE. UU. (NCHS)
http://www.cdc.gov/nchs

National Opinion Research Center (NORC) de la Universidad de Chicago
www.norc.org/DataEnclave

National Science Foundation (Estados Unidos)
http://www.nsf.gov/index.jsp

Oficina Estadística de la Unión Europea (Eurostat)
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

Programa de Centros de Datos de Investigación (CDR) de Statistique Canada
www.statcan.gc.ca/rdc-cdr/index-fra.htm

Programa DHS
http://www.measuredhs.com

Statistique Canada, Iniciativa para la Democratización de los Datos (IDD)
http://www.statcan.gc.ca/dli-ild/dli-idd-fra.htm

UK Data Archive / Centro de Almacenamiento de Datos del Reino Unido (UKDA)
http://www.dataarchive.ac.uk/sharing/metadata.asp
http://securedata.ukda.ac.uk/about/about.asp

UK Statistics Authority / Autoridad Estadística del Reino Unido
http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html

US Bureau of the Census / Oficina del Censo de EE. UU.
http://www.census.gov/population/www/cen2000/pums/index.html
www.census.gov/srd/sdc/

Wikipedia
http://en.wikipedia.org

Acerca de la IHSN

En febrero de 2004, representantes de numerosos países y organismos de desarrollo se reunieron en Marrakech (Marruecos) para celebrar la segunda Mesa Redonda Internacional sobre la Gestión basada en los Resultados en el ámbito del desarrollo. Su objetivo era analizar cómo mejorar la coordinación de la ayuda de los donantes para reforzar los sistemas estadísticos y la capacidad de seguimiento y evaluación, de forma que los distintos países pudieran hacerse cargo de sus procesos de desarrollo. Entre otras cosas, en esa mesa redonda se decidió poner en marcha el Plan de Acción de Marrakech para la Estadística (MAPS).

Una de las principales recomendaciones del MAPS era la creación de una Red Internacional de Encuestas de Hogares (IHSN). Al hacerla realidad, la comunidad internacional reconocía el papel determinante que desempeñan los modelos de encuesta en la planificación, aplicación y seguimiento de las políticas y los programas de desarrollo. La IHSN proporciona a las organizaciones nacionales e internacionales una plataforma para una mejor coordinación y gestión de la recogida y el análisis de los datos socioeconómicos, así como para movilizar los medios necesarios para mejorar la eficacia y la eficiencia de los métodos de realización de encuestas en los países en desarrollo.

Los documentos de trabajo de la IHSN potencian el debate e intercambio de ideas sobre la concepción y realización de las encuestas domiciliarias, así como sobre el análisis, la difusión y la utilización de los datos recogidos mediante esas encuestas. Para proponer la publicación de un texto en la serie de documentos de trabajo de la IHSN, póngase en contacto con su secretaría mandando un mensaje de correo electrónico a la siguiente dirección: info@ihsn.org.

www.ihsn.org
E-mail: info@ihsn.org