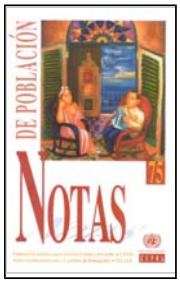

Notas de Población No. 75

LC/G.2186-P
Diciembre de 2002

ISBN: 92-1-322063-4
ISSN
Electrónico: 1681-0333
N.Venta S.03.II.G.77

Resumen
Bajar documento
Solicitar impreso
Enviar por email
Imprimir esta página

Resumen

Resumen

En el número 75 de la serie Notas de población se da a conocer los últimos avances en relación con los censos de población y vivienda, en el momento en que en aproximadamente la mitad de los países de la región se han levantado los censos de la ronda del 2000. Es por ello que en la selección de estudios e informes contenidos en esta publicación se abordan materias referidas a la experiencia aportada por la ronda anterior y a los nuevos desafíos que se encaran. Entre estos últimos destacan los temas emergentes en la agenda de las políticas y programas de desarrollo y las nuevas tecnologías, cuyos propósitos fundamentales son reducir costos y mejorar la calidad y la oportunidad de los resultados censales. Los artículos incluidos se sustentan en actividades llevadas a cabo en años recientes, en el marco de la preparación de los censos de esta década. En ellos se tratan, por una parte, temas conceptuales, vinculados a la finalidad y el contenido de los censos y, por otra, las experiencias en cuanto a la aplicación de tecnologías modernas y el análisis de sus potencialidades.

Los primeros seis capítulos son de interés general y están referidos a la experiencia de la región, en ellos se consideran las diversas etapas y temáticas que debe abarcar un censo. Así, el documento de José Miguel Guzmán y Susana Schkolnik destaca la importancia de los censos como fuente de información para el desarrollo social, plantea los nuevos desafíos que deben enfrentar, el papel del sector privado en los censos, las nuevas formas de relación usuarioproductor y las alternativas a los censos convencionales. El trabajo de Juan Chackiel recoge la experiencia censal reciente en la región y pasa revista a los criterios utilizados en cada fase de su realización, y a los nuevos enfoques conceptuales y tecnológicos. Arij Dekker analiza las formas en que se pueden adaptar las nuevas tecnologías a la operación censal, lo que implica una selección apropiada, el mantenimiento de la integridad de los sistemas censales y estadísticos, la opción de terciarizar actividades y la confidencialidad de la información. Anil Arora presenta la experiencia de Canadá y sugiere cambios que debieran introducirse en el programa censal para 2006. En el artículo de Alejandro Giusti se examina la propuesta de Argentina para satisfacer, mediante encuestas post-censales complementarias, la demanda de información sobre sectores determinados de población, como pueblos indígenas, discapacitados, migrantes, entre otros. A continuación, siguen artículos relacionados con aspectos conceptuales específicos, el trabajo de Elizabeth Solano se refiere a un tema de gran relevancia actual: la investigación sobre la pertenencia étnica. Ralph Hakkert analiza la experiencia latinoamericana en lo que respecta a las preguntas destinadas a investigar la fecundidad y la mortalidad, teniendo en cuenta las deficiencias de las estadísticas vitales en la mayoría de los países de la región. En los artículos que tratan aspectos tecnológicos específicos, se incluye un documento de Werner Haug referido a

PRESERVACIÓN DE ARCHIVOS CON DOCUMENTOS Y MICRODATOS CENSALES Y AUMENTO DE LOS GRUPOS DE GESTIÓN [*]

Wendy L. Thomas[**] **y Robert McCaa**[***]
University of Minnesota Population Center

RESUMEN

Cuando los datos y documentos censales han sido bien preservados permiten una efectiva recopilación, difusión, planificación y un buen uso futuro. En la era electrónica se corre el riesgo de que tanto la documentación como los microdatos no estén bien preservados, es decir, útiles para su uso; de hecho, se pueden perder o quedar ilegibles debido a la obsolescencia tecnológica o a la falta de cuidado. En este trabajo se analizan temas relacionados con la preservación a largo plazo: qué preservar, cómo determinar el valor futuro y de qué manera una buena política en tal sentido puede hacer que crezca el número de grupos de interés. El valor potencial del censo puede aumentar sustancialmente, en particular el estudio de los procesos sociales y demográficos que se produjeron a través del tiempo. Ello se obtiene mediante el uso continuo de los microdatos censales, como sucede con los proyectos IPUMS (estadounidense e internacional), SAR (Reino Unido) y otros.

---

[*] Este documento fue presentado en la Reunión del Grupo de Expertos del Simposio sobre el Examen Mundial de la Ronda 2000 de los Censos de Población y Vivienda, División de Estadística, Secretaría de las Naciones Unidas, Nueva York, 7 al 10 de agosto de 2001.

[**] wlt@pop.umn.edu.
[***] rmccaa@umn.edu.

ABSTRACT

Proper preservation of census data and documents contributes to effective processes of collection dissemination and planning, and the future use of the censuses. In the electronic age there is the risk that both documents and microdata may not be well preserved, that is to say, may no longer be useful; they may be lost or become illegible owing to technological obsolescence or a lack of care. This paper reviews topics related to long-term preservation: what to preserve, how to determine future value and in what way a good policy in that respect can increase the number of interest groups. The potential value of the census may increase substantially, especially for the study of social and demographic processes that have taken place over time. Such studies rely on the continuous use of census microdata, as in the case of IPUMS (United States and international), SARs (United Kingdom) and other projects.

RÉSUMÉ

Des données et des documents censitaires bien préservés permettent une collecte, une diffusion et une planification efficaces, ainsi qu'une bonne utilisation dans l'avenir. L'ère électronique implique certains risques quant à la préservation de la documentation et des microdonnées en termes de leur utilisation : ces données peuvent effectivement se perdre ou devenir illisibles en raison d'une technologie obsolète ou d'un manque d'entretien. Cette étude porte sur l'analyse de différents aspects de la préservation à long terme : quels sont les éléments à préserver, comment déterminer la valeur future et comment des actions judicieuses en ce sens peuvent contribuer à l'émergence de nouveaux groupes d'intérêt. La valeur potentielle du recensement peut augmenter considérablement, notamment l'étude des processus sociaux et démographiques qui ont évolué au fil du temps. , et ce grâce à l'utilisation permanente des microdonnées censitaires, comme cela est le cas dans les projets IPUMS (des Etats-Unis et international), SAR (Royaume-Uni), etc.

I. INTRODUCCIÓN

La preservación de la documentación censal debe ser considerada al comienzo del ciclo de actividades censales, pues contribuye a una efectiva recopilación, difusión, planificación y uso futuro de los censos. La capacidad para aprender de procesos pasados, identificar estrategias para un censo exitoso, conservar y recurrir a actividades y estructuras centrales de censos anteriores, y aplicar eficazmente los datos censales a problemas actuales y futuros depende de la preservación de los datos censales y la documentación relacionada con la recopilación y procesamiento de esos datos.

Es fácil determinar lo que se debe preservar y el modo en que debe ser preservado si se contara con un mundo ideal de recursos ilimitados. Desafortunadamente, éste no es el caso y, aun en los países más ricos, el costo de la preservación y de los temas que giran en ese entorno tienen un profundo impacto en la documentación y en el formato en que deben ser preservados. El objetivo de este trabajo es analizar los tipos de datos y documentación acumulados durante el proceso censal y explorar los beneficios que genera la preservación de este tipo de documentos para futuros censos o usuarios de datos, el aseguramiento de formatos de preservación, adecuación e identificación de grupos de interés que pueden constituir una fuerza eficaz que abogue por la preservación de estos documentos.

La clasificación de la documentación para preservación, en términos de su impacto futuro y de la anticipación de su uso, es útil para identificar las ventajas comparativas de las decisiones que toma cada país.

Al empalmar este tipo de listas de documentación con un inventario de la tecnología disponible, de personal y de conocimientos dentro de un país para procesar la documentación que se desea preservar, los gobiernos contarán con información que les permitirá tomar decisiones bien fundadas. El uso de un cuestionario que genere información sobre infraestructura disponible de preservación dentro de un país también puede brindar opciones de servicios cooperativos o perfilar tecnologías que se adecuan a diversas situaciones. La capacidad para determinar no sólo lo que se preservará sino también lo que no lo será —basada en una comprensión del impacto a largo plazo que tendrá la información incluida en el documento—, es crucial para desarrollar una política de preservación a largo plazo.

II. PRESERVACIÓN A LARGO PLAZO DE LOS DATOS Y LA DOCUMENTACIÓN

a) Definición de preservación a largo plazo

La preservación a largo plazo asume un nuevo significado con los registros electrónicos. "Archivar" es un término utilizado tanto por especialistas en computación/informática como por archivistas; sin embargo, para estos dos grupos el término tiene distintos significados. "Archivar" en el mundo de la computación significa almacenamiento inactivo o memoria indirecta ("off-line"). Para los archivistas, "archivar" significa preservar un registro de información en un formato que es independiente de su medio de producción y protegerlo contra pérdida, modificación o deterioro.

Para los archivistas, un registro electrónico bien preservado tiene las siguientes características. (Dollar, 2000, 47-57):

- **Legible**. Es decir, no está dañado y la secuencia de bits puede ser procesada ya sea por la máquina que la creó, la que la está almacenando, o aquella en la que será almacenada.
- **Inteligible.** Que tiene suficientes metadatos para interpretar los unos y ceros de la imagen en el mapa binario. En otras palabras, información sobre el algoritmo de compresión y el orden de los bytes. Es similar a la extensión de archivo TXT que denota un archivo de texto ASCII de 7 bits. Si carece de este nivel básico de metadatos, el registro es ininteligible para todos los efectos prácticos.
- **Identificable.** Es decir, puede ser localizado mediante una identificación o atributo único.
- **Encapsulado** de modo tal que toda la información del registro (sus metadatos y vínculos) exista como una sola entidad lógica o física.
- **Comprensible** a través del suministro de metadatos completos.
- **Regenerable** en cuanto al contenido lógico, físico e intelectual.
- **Registros auténticos.** "La ciencia de archivar define a los registros auténticos como lo que aparentan ser: registros confiables que con el correr del tiempo no se han modificado, cambiado o corrompido" (Dollar, 2000, 54).

Es importante que este concepto de preservación sea tenido en cuenta al momento de determinar el valor de preservar ciertos registros censales y los costos de distribución, almacenamiento y preservación a largo plazo.

b) El valor de la preservación

Mucho se ha escrito sobre la importancia de organizar y coordinar la realización de censos dentro y entre países (Naciones Unidas, 2000). Este proceso cuenta con el apoyo y asistencia de numerosas agencias intergubernamentales y no gubernamentales. Se ha puesto énfasis en la planificación, recopilación de datos, metodologías, preparación del producto y difusión. El valor de un buen programa censal reside no sólo en preservar datos reales, metadatos y resultados para su uso futuro sino también en servir de apoyo para futuros censos y actividades estadísticas.

Dada la naturaleza periódica de los censos, la conservación de registros sobre la manera en que se llevaron a cabo ciertas actividades específicas puede ser de ayuda en futuros procesos censales de un país y permitir que las agencias adquieran experiencia basada en procesos y estrategias pasadas. Esto reviste particular importancia para aquellos países que no tienen una oficina permanente para el censo. Los registros cuidadosamente seleccionados y preservados brindan información detallada sobre el proceso de planificación y sobre las especificaciones referidas al proceso de recopilación, y explican por qué se tomaron algunas decisiones y cuán efectivas resultaron ciertas actividades. De hecho, estos tipos de procesos y enfoques específicos nacionales son los que podrían ayudar a incorporar y aprovechar actividades y estructuras exitosas.

La preservación y difusión de información sobre calidad de datos y evaluación del proceso es valiosa para futuras actividades censales y crucial para que el uso de datos censales esté bien fundado. Si se entrega información sobre la confiabilidad, limitaciones y validez de los datos finales, los usuarios comprenderán que cualquier cambio de procedimiento puede afectar cualquier análisis que se desee realizar. Éste es el tipo de información que debería ser encapsulada en el proceso de preservación a través de enlaces lógicos o físicos entre los datos censales y los metadatos de procedimiento.

c) Costos de la preservación

El costo de la preservación es un tema importante en todos los países. Los recientes debates sobre conservación del censo 2000 de los Estados Unidos de Norteamérica generaron numerosas respuestas de varios grupos de interés referidas a la preservación de los formularios originales y de los resultados de procesos intermedios. El costo de preservación de formularios con enumeración original en varios formatos y el costo asociado en que se

incurre para hacerlos identificables por futuros usuarios fueron factores clave en la negociación de un plazo final de conservación.

En aquellos países que no poseen oficinas permanentes de censos y/ o instalaciones permanentes para archivos nacionales, el costo de preservación es un gran problema. Si estos costos se tienen en cuenta con anticipación y se los incluye en el análisis de los costos generales del censo, se podrán encontrar opciones adicionales para la asignación de fondos. Por ejemplo, la forma en que los datos censales se obtienen y preparan para su difusión puede reducir el costo que implica crear un registro de preservación de buena calidad. Además, la captura y conservación de información sobre procedimientos a medida que la misma se genera y la creación de enlaces lógicos o físicos con recopilaciones de datos emergentes aumentan la posibilidad de la preservación y, al mismo tiempo, reducen el costo de tener que reconstruir valiosa información de metadatos.

Un análisis previo de los costos y del valor futuro de preservar información permite tomar decisiones bien fundadas y da la oportunidad de discutir en forma oportuna posibilidades de preservación a largo plazo.

III. DETERMINACIÓN DE LO QUE SE DEBE PRESERVAR

a) Preservación de productos

Los elementos esenciales de cualquier censo, en términos de preservación, son los datos y la documentación básica. La forma en que esos datos se identifican y definen varía de país en país. Los temas de confidencialidad y seguridad desempeñan un papel fundamental para determinar no sólo quién debería tener acceso a los microdatos y a los formularios de enumeración sino también si esa información debería ser conservada. Una mayor disponibilidad de microdatos aumenta la posibilidad de que estos datos sean preservados.

Son cada vez más los países que proporcionan acceso a microdatos a través de: muestras públicas, muestras científicas (restringidas a unos pocos proyectos cuidadosamente seleccionados) y enclaves de datos donde el usuario trabaja en un sitio seguro y el producto es estrechamente controlado. Desde 1985 hasta 1994, de 153 países con un millón o más habitantes, 134 realizaron censos en la ronda de 1990; se contó al 94% de la población mundial. Cincuenta y cuatro países permitieron que los investigadores tuviesen acceso a muestras anónimas de censos de población y vivienda. Otros restringieron el acceso a un solo investigador o centro de

investigación, pero lo que es notable en la ronda de 1990 es no sólo la globalización del censo sino la creciente aceptación de muestras anónimas como instrumentos estadísticos, tendencias que siguen vigentes en la ronda de censos del 2000 (1995-2004).

Por ejemplo, el enfoque utilizado en las Naciones Unidas de suministrar muestras públicas cuyos tamaños variaban entre 1% y 15% en varios tipos de zonas sirve de soporte a una amplia gama de investigaciones tanto a nivel local como nacional. Además, la divulgación de datos (después de haber estado restringidos durante 73 años) ha generado una cantidad de proyectos destinados a que el público tenga acceso a los mismos en formato digital. El más notable es el proyecto Integrated Public Use Microsample (IPUMS). Este proyecto, que se inició en 1992 en la Universidad de Minnesota, incluye 65 millones de registros de microdatos de los Estados Unidos de América. Concebido por Steven Ruggles —director fundador del Minnesota Population Center— y financiado por la National Science Foundation/National Institute of Health, IPUMS es parte de los censos decenales de los Estados Unidos de América, que datan desde 1850 a 1990. La primera versión de la base de datos IPUMS se difundió en cinta magnética en 1993 y por Internet en 1995. Gracias a la expansión de Internet, el problema de distribución de datos se resolvió fácilmente mediante un motor de difusión de datos a través de un sitio web (http// www.ipums.org). La base de datos IPUMS, distribuida en forma gratuita por Internet, se convirtió rápidamente en una de las tres fuentes de datos más frecuentemente citadas en los estudios de investigación demográfica de los Estados Unidos de América.

Con fondos suministrados mayormente por la National Science Foundation, en octubre de 1999 se inauguró un proyecto mundial, denominado IPUMS-International. El consorcio IPUMS-International, que cuenta con la colaboración de equipos nacionales de investigadores, propone integrar microdatos censales en más de una docena de países, por lo menos uno de cada continente. Se incluirán en la base de datos microdatos históricos de censos de Canadá, Noruega, Gran Bretaña, Argentina y Costa Rica y también de los Estados Unidos de América. Los microdatos actuales de Colombia y los Estados Unidos de América se integrarán con los de Francia, Brasil, Méjico, Vietnam, Kenya, Gran Bretaña, Hungría, España y otros. Basándose en un prototipo desarrollado con la cooperación del Departamento Administrativo Nacional de Estadística de Colombia (DANE), se están formando equipos nacionales de usuarios experimentados en el uso de datos censales para que asesoren sobre la manera de unificar los conceptos nacionales de censos utilizando normas internacionales.

Varios países están creando muestras de uso público para incrementar el acceso a los microdatos. Programas del tipo Integrated Microcomputer Processing System (IMPS) y su sucesor CSPro, un sistema de procesamiento de datos censales y de encuestas desarrollado por el International Statistical Programs Center del U.S. Bureau of the Census facilita la difusión de muestras de microdatos suministrando herramientas para realizar tabulaciones cruzadas, producir mapotecas digitales y otros análisis básicos con lo cual se reduce el costo de producción de estos productos en cada país.

Entre los países que distribuyen muestras de microdatos de uso público se encuentra Vietnam, que dio a conocer una muestra (3%) del Censo de Población y Vivienda de 1990 con la intención de producir más adelante una muestra completa del 100%. México difundió un 10% de una muestra diseñada para generar información valiosa a nivel de municipalidades de 100 000 o más habitantes. Francia difundió 5% de las muestras de 1961-1990. Del mismo modo, la Central Bureau Statistics de Kenya ha preparado una mega muestra de la enumeración 1999 (con una densidad máxima del 20%) a fin de completar su impresionante serie de muestras de 1969, 1979 y 1989.

Estas recopilaciones no sólo suministran datos en un formato que se puede preservar sino que incluyen una gama de metadatos. La documentación es sumamente completa e incluye detalles sobre cada aspecto del censo, desde los preparativos iniciales hasta la publicación final de los cuadros. El debate del muestreo es particularmente notable.

Un creciente número de países ofrece datos en formato REDATAM (desarrollado por la División de Población de la CEPAL-Centro Latinoamericano y Caribeño de Demografía (CELADE)) como una forma de almacenar microdatos que pueden utilizar aquellos investigadores y administradores que necesitan estadísticas para áreas pequeñas.

REDATAM "**RE**cuperación de **DAT**os para **Á**reas pequeñas por **M**icrocomputadores" fue originalmente concebido como un programa de computación para la recuperación de datos a bajo costo y se ha convertido en un concepto que comprende un formato patentado de base de datos y un sistema para el desarrollo de software. El formato busca asegurar datos sensibles manteniendo al mismo tiempo la invalorable flexibilidad del acceso a los microdatos. También dispone de un servicio web que beneficia a organizaciones nacionales que se resisten a proporcionar datos pero que están dispuestas a brindar al público acceso a los datos y/o acceso privilegiado a usuarios seleccionados. El programa está disponible en forma gratuita en Internet. REDATAM ha sido desarrollado durante las dos últimas

décadas gracias al apoyo financiero de diversas organizaciones internacionales (CEPAL-Naciones Unidas, FNUAD, Gobierno de Canadá a través de CIDA e IDRC, IDB y otras) (http//www.cepal.cl/celade).

Países con ronda de censos de 1990 en REDATAM:

América Latina: Argentina, Brasil, Chile, Colombia, República Dominicana, Guatemala, Honduras, Nicaragua, Paraguay, Suriname, Uruguay, Venezuela, y el Caribe de habla inglesa.
Asia: Camboya* y Corea del Norte*
África: Benin, Burkina Faso, Burundi, Camerún, Egipto, Gabón, Ghana, Kenya, Madagascar, Malí, Nigeria, Ruanda*, Seychelles, y Zimbabwe*.

* Base de datos con 100% de microdatos de población.

Si bien estos archivos de microdatos no están en formato de archivo en sentido estricto, han sido capturados de forma tal que el organismo que los generó puede producir un archivo en formato ASCII con metadatos estructurales completos físicamente encapsulados para asegurar su futura comprensión. Es importante que formatos como REDATAM no sean considerados como formatos de archivo a largo plazo. El problema que se presenta cuando no se crea una copia de archivo y los registros se guardan en un formato patentado es el costo de llevar esa información a otro formato. Los formatos patentados pronto se convierten en formatos heredados, cuya edad, dependencia de sistemas, idiomas o hardware los hace difíciles, costosos y a veces imposibles de traspasar.

b) Preservación del proceso

Diversos manuales sobre realización y administración de censos nacionales proporcionan listas detalladas de procedimientos y procesos. Este tipo de información y los detalles sobre ciertos enfoques y metodologías son necesarios para interpretar en forma precisa los datos resultantes. También sería conveniente considerar los tipos de información de proceso, que a veces se pasan por alto, para preservar la memoria institucional. Esto implica registrar y preservar las razones y formas del proceso censal. Es más eficiente en función de los costos utilizar esta información mientras se toman las decisiones que recuperarla más adelante. Se debe prestar especial atención al hecho que dicha información debería ser utilizada en formato no patentado para evitar que la misma se pierda debido a los costos de traspaso.

Todo ciclo censal consiste de cuatro fases (Naciones Unidas, 2000).

- Preparación
- Actividades sobre el terreno
- Procesamiento de datos
- Evaluación

Entonces, para cada fase reviste particular interés la siguiente documentación:

- Informes sobre procedimientos y métodos.
- Comparación de conceptos y procedimientos con censos anteriores y normas internacionales vigentes.
- Informes de evaluación para cada ciclo censal y documentos más importantes en los cuales se basan los informes.
- Libros de registros (desde el del jefe del censo hasta el registro del enumerador), aunque éstos pueden ser demasiado primarios para ser difundidos ampliamente.

La etapa final consiste en documentar los datos difundidos y la respectiva documentación y notas.

IV. CÓMO DETERMINAR EL VALOR FUTURO

a) Quiénes forman los grupos de interés

Según lo ya expresado aún continúan surgiendo normas para permitir el acceso a microdatos censales; el principal problema relacionado con su preservación y acceso no es técnico sino político. A medida que el acceso y uso de los datos censales aumenta, la naturaleza y complejidad de los grupos de interés también aumentan; éstos se encuentran en el sector gubernamental, no gubernamental, académico, comercial y de usuarios nuevos. Si bien siempre habrá intereses que compiten por lo que debería ser conservado, los compartidos por diversos grupos los que ayudarán a identificar la documentación que será preservada a largo plazo. Si se celebran consultas con estos grupos antes de inciar el proceso, aumentará la posibilidad de obtener y conservar financiamiento para la preservación a largo plazo.

b) Impacto futuro y uso anticipado

En términos de patrones de impacto futuro y uso de datos censales, el mayor potencial radica en preservar y mantener el acceso a microdatos.

Los datos censales se usan para analizar problemas específicos de orden social, económico y demográfico, cuya naturaleza cambia constantemente. La capacidad para crear agregados comparables a lo largo del tiempo o para identificar nuevas áreas geográficas o definiciones emergentes depende solamente de la conservación de archivos de microdatos. Éste es el caso de las estadísticas para áreas pequeñas que no se pueden obtener de agregados de áreas más extensas.

Las tabulaciones específicas tienen claras dificultades. El ciclo censal y la producción de agregados estadísticos significan que las preguntas formuladas y los cuadros creados a menudo reflejan inquietudes e intereses que datan de cinco a diez años previos a la publicación. Es probable que las nuevas tabulaciones no sean comparables con las tabulaciones de censos previos debido a cambios en la cohorte o en los agrupamientos de clasificación, en el universo del cuadro o en la definición. Sin acceso a microdatos, sin importar cuán asegurado esté, los investigadores y analistas tienen pocas opciones. Además, los microdatos se prestan para la investigación y el discurso académico y aumentan el valor derivado de un censo individual.

Consideremos el ejemplo de Canadá: en la década de 1970, el National Statistical Services comenzó a difundir grandes cantidades de muestras de microdatos censales. En Canadá, la modificación de la Ley de Estadísticas llevada a cabo en 1971 permitió la difusión pública de microdatos no confidenciales (Tambay y White, 2001). A partir de la década de 1970, Statistics Canada, con su serie de enumeraciones quinquenales, emitió en forma regular muestras de microdatos censales. Hasta el año 1996 los investigadores tenían que solicitar las muestras en forma individual y la distribución estaba muy restringida. En ese año se puso en práctica una propuesta de difusión de datos para que las universidades canadienses pudiesen suministrar muestras de microdatos a los investigadores y alumnos. Como resultado de ello se produjo un desarrollo explosivo de trabajos de investigación. Si bien antes de que se permitiera la difusión sólo cinco o diez especialistas al año podían obtener muestras de microdatos, después de haberse autorizado la difusión se podía tener acceso a una sola muestra cientos de veces al mes en una universidad muy importante. Dada la profusión de proveedores, actualmente es imposible compilar estadísticas de uso, mientras que antes el organismo registraba el nombre de cada usuario. Como el uso de microdatos censales se ha extendido dentro de las aulas universitarias, los especialistas canadienses están enseñando a una generación de ciudadanos más jóvenes la utilidad del censo y están democratizando el

acceso a los datos censales (Lisa Dillon, comunicación privada, 21 de abril de 2001).

En el Reino Unido se construyeron muestras de censos de uso público llamadas SAR (Samples of Anonymized Records) para la enumeración de 1991 con una densidad de muestras del 2.0% de registros individuales y del 0.5% de viviendas. No se identificaron unidades administrativas con menos de 120 000 habitantes. A pesar de la pequeña densidad de las muestras y de la falta de detalles geográficos, se produjo un desarrollo explosivo de trabajos de investigación que utilizaban las SAR. Durante los seis años posteriores a la primer difusión de datos, se publicaron cientos de estudios. Antes de la enumeración de 2001, se realizó una nueva evaluación de los riesgos de la divulgación y en ese momento se tuvo en cuenta la varianza residual y de codificación así como las diferencias en los programas cronológicos y de codificación entre grupos de datos. Con el permiso de la Office of National Statistics se otorgó acceso privilegiado para comparar los registros de encuestas con las SAR (Dale y Elliot, en prensa). El objetivo era determinar los riesgos prácticos, en oposición a los teóricos, de identificar individuos combinando dos fuentes. Los autores consideraron que las evaluaciones anteriores sobre la posibilidad de identificar individuos exageraban los riesgos porque no tuvieron en cuenta la varianza residual, las diferencias de oportunidad y las incompatibilidades de los programas de codificación. A partir de este riguroso ejercicio, Dale y Elliot llegaron a la siguiente conclusión:

> Si un usuario de una base de datos externa no tiene oportunidad de llevar a cabo una verificación, sería inútil que intente realizar este tipo de combinación. En primer lugar, es dable esperar un pequeño grado de superposición lo cual sería un gran factor de disuasión para un intruso.
>
> Sin embargo, si se intenta combinar ambos archivos, la gran cantidad de combinaciones aparentes sería muy confusa ya que el intruso no tendría forma de verificar la identificación correcta.

c) Información sobre censos futuros

La disponibilidad de información de un proceso que es propio de un país y/o estructura organizativa puede servir de insumo en la preparación de censos futuros al brindar un panorama claro de lo que sucedió, cómo sucedió, por qué se llevó a cabo de esa manera y de los logros y dificultades. Este tipo de información reviste particular importancia para aquellos países que no tienen una oficina permanente o que cuentan con mínimo personal

permanente y que deben crear un nuevo sistema para cada censo. El suministro de documentación sobre la razón para qué los procesos y procedimientos se siguieron de cierta forma también puede ser de ayuda para los consultores técnicos internacionales, porque les brinda un panorama claro y completo de las actividades realizadas durante los censos anteriores.

V. INVENTARIO DE TECNOLOGÍA/PERSONAL/ CONOCIMIENTOS DISPONIBLES

Antes de la ronda de censos de 1990, la División de Estadística de las Naciones Unidas distribuyó un cuestionario referido a la cobertura general del censo, estructura organizativa, trabajo cartográfico, listado de casas y/ o viviendas, ensayos (ensayos previos, ensayos piloto, etc.), cuestionarios del censo, enumeradores y supervisores, enumeración, muestreo, procesamiento de datos, evaluación y análisis, difusión de datos, costos y actividades futuras. Además de la información ya solicitada, las siguientes áreas de información relacionadas con la preservación a largo plazo servirían de ayuda para que los países incluyan el tema de la preservación al inicio del proceso. El reconocimiento anticipado de las necesidades y posibilidades de preservación permitirá tomar decisiones de preservación bien fundadas.

- Existencia de un archivo nacional capaz de preservar documentación digital.
- Existencia de un plan de preservación que incluya un plazo de conservación de registros.
- Tipo de documentación (resultados de datos, metadatos, documentos del proceso) que se está preservando y que se espera preservar.
- Disponibilidad de personal capacitado en preservación.
- Uso de depósitos/archivos para datos y documentación.

VI. CONCLUSIÓN

Si los microdatos censales llegan a usarse en forma extensa, es necesario resolver el tema de la confidencialidad estadística de forma tal que satisfaga a los organismos nacionales de estadística, al público y a investigadores.

En la última década Eurostat patrocinó cinco conferencias internacionales sobre el tema. Gracias en parte a estos esfuerzos, y a otros, la regla general ahora consiste en preparar muestras de microdatos para una diversidad de usuarios. De los 52 estados miembro del International Monetary Fund's General Data Dissemination System (Sistema de Difusión de Datos Generales del Fondo Monetario Internacional), unos tres de cuatro difunden muestras de microdatos censales, de un modo u otro. El establecimiento de normas internacionales de microdatos aumentará aún más la disponibilidad de muestras censales con lo que se facilitará la investigación comparativa, tanto en tiempo como en espacio. Siempre que se han adoptado políticas de difusión pública se ha producido un desarrollo explosivo en el campo de la investigación, sin que haya habido una sola instancia en que se violó o siquiera se haya supuesto que se hubiese violado la confidencialidad estadística.

Es importante comprender e incorporar este concepto de preservación ya que de ese modo se asegura que los datos censales estarán protegidos contra pérdida, modificación o deterioro.

"En este sentido, la obligación de los archivistas consiste en explicar a los especialistas en computación, a los especialistas en informática, y a los que no están familiarizados con los archivos la importancia que tiene un espacio físico o lógico, 'independiente de su medio de producción', donde los registros están protegidos contra pérdida, modificación y deterioro para que puedan ser utilizados como evidencia confiable todo el tiempo que sea necesario. De esto se trata archivar" (Dollar, 2000, 26).

BIBLIOGRAFÍA

Dale, Angela y Mark Elliot (s/f), "Proposals for 2001 SARs: An assessment of disclosure risk", *Journal of the Royal Statistical Society, Series A*, Nueva York, en prensa.

Dollar, Charles M. (2000), *Authentic Electronic Records: Strategies for Long-Term Access*, Chicago, Illinois, Cohasset Associates, Inc.

General Statistics Office (2000) *Data and Results from the 3% Sample of The Population and Housing Census*, Hanoi, Centro de Procesamiento de Información Central, agosto.

INEGI (Instituto Nacional de Estadística, Geografía e Informática) (2001), *Contar 2000. Sistema para la consulta de tabulados y base de datos de la muestra: XII Censo General de Población y Vivienda 2000,* Aguascalientes, México.

Joint ECE/Eurostat Secretariat (2001), "Report of the March 2001 Work Session on Statistical Data Confidentiality. Work Session on Statistical Data Confidentiality", Skopje, Comisión Económica para Europa (CEPE)/Oficina de Estadística de las Comunidades Europeas (Eurostat), marzo.

Naciones Unidas (2000), *Handbook on Census Management for Population and Housing Censuses, Studies in Methods,* Series F, N° 83, Nueva York, División de Estadística, Departamento de Asuntos Económicos y Sociales (DESA).

___ (1992a), *Handbook Population and Housing Censuses: Part 1 Planning, Organization and Administration of Population and Housing Censuses, Studies in Methods,* Series F, N° 54, Nueva York, División de Estadística, Departamento de Asuntos Económicos y Sociales (DESA).

___ (1992b), *Handbook Population and Housing Censuses: Part 2 Demographic and Social Characteristics, Studies in Methods,* Series F, N° 54, Nueva York, División de Estadística, Departamento de Asuntos Económicos y Sociales (DESA).

___ (1990), *Manual on Population Census Data Processing using Microcomputers, Studies in Methods,* Series F, N° 53, Nueva York, División de Estadística, Departamento de Asuntos Económicos y Sociales (DESA).

___ (1991), *Emerging Trends and Issues in Population and Housing Censuses, Studies in Methods,* Series F, N° 52, Nueva York, División de Estadística, Departamento de Asuntos Económicos y Sociales (DESA).

Ruggles, Steven, J. David Hacker y Matthew Sobek (1995), "Order out of chaos: General design of the Integrated Public Use Microdata Series", *Historical Methods*, vol. 28.

Ruggles, Steven y otros (2000), "IPUMS-USA: Integrated Public Use Microdata Series for the United States", *Handbook of International Historical*

Microdata for Population Research, Patricia Kelly-Hall, Robert McCaa y Gunnar Thorvaldsen (comps.), Minneapolis, Minnesota.
Tambay, Jean-Louis y Pamela White (2001), "Providing greater accessibility to survey data for analysis. Work Session on Statistical Data Confidentiality", Skopje, Joint ECE/Eurostat, marzo.

ESA/STAT/AC.84/26
1 August 2001

English only

Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade

Assessment and Future Prospects

Statistics Division
Department of Economic and Social Affairs
United Nations Secretariat
New York, 7-10 August 2001

Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders

Wendy L. Thomas and Robert McCaa

Session 4
Maintaining census related activities during intercensal years

UNITED NATIONS SYMPOSIUM
GLOBAL REVIEW OF 2000 ROUND OF POPULATION AND HOUSING CENSUSES:

MID-DECADE ASSESSMENT AND FUTURE PROSPECTS

August 7-10, 2001 / UNSD-NY.

Archiving Census Documentation and Microdata: Preserving Memory, Increasing Stakeholders

Wendy L. Thomas and Robert McCaa
University of Minnesota Population Center
wlt@pop.umn.edu; rmccaa@umn.edu

1.0 Introduction

The preservation of various types of census materials must be raised early in the cycle of census activities. Well-preserved data and documentation contribute to effective data collection, dissemination, planning, and future use of the population census. The ability to learn from past processes, identify strategies that contribute to a successful census, retain and build on core activities and structures from previous censuses and effectively apply census data to current and future issues is all dependent upon the preservation of census data and the materials related to the collection and processing of that data.

In an ideal world with unlimited resources the questions of what to preserve and how to preserve it would be easier to address. Unfortunately this is not the case and even in the wealthiest of countries the cost of preservation, and questions surrounding the means of preservation, have a profound impact on what materials are preserved and in what format. The purpose of this paper is look at the types of data and documentation accumulated during the census process and explore the benefits of preserving these types of documents in terms of informing future censuses and data users, ensuring appropriate preservation formats, and identifying stakeholders who may be an effective force in lobbying for the preservation of various classes of documents.

Classifying materials for preservation in terms of their future impact and anticipated use is useful for identifying the trade-offs in preservation decisions for individual countries. By coupling this type of materials lists with an inventory of the available technology, personnel and knowledge within a country to process materials for preservation, governments will have the information necessary to enable them to make informed preservation decisions. The use of a questionnaire to elicit information on the available infrastructure for preservation within a country may also bring to light options for cooperative services or a profile of appropriate technologies for a variety of situations. The ability to not only determine what will be preserved, but also what will not be preserved, based on an understanding of the long-term impact of the information contained in the document is instrumental in developing a long-term census preservation policy that will meet the needs of future generations.

2.0 Long-term preservation of data and documentation

2.1 Definition of long-term preservation

Long-term preservation takes on a new meaning with electronic records. “Archiving” is a term used both by computer/information technology specialist and archivist, yet conveys different meanings to these two groups. “Archiving” in the world of computing refers to inactive or off-line storage. To archivists “archiving” means to preserve an information record in a format that is independent of its production environment and to protect that record from loss, alteration or deterioration.

For archivists, well-preserved electronic records have the following characteristics. (Dollar, 47-57) They are:

- **Readable**. In short, they are undamaged and the bit-stream can be processed either the machine that created it, the machine that is storing it, or the machine on which it will be stored.
- **Intelligible**, having sufficient metadata to interpret the 1s and 0s of the bitmap image. In other words information regarding the compression algorithm and the byte order. This is similar to the file extension TXT denoting a 7-bit ASCII text file. Without this basic level of metadata the record is for practical purposes unintelligible.
- **Identifiable** in that a unique ID or attribute can locate them.
- **Encapsulated** so that all the information associated with a record (its metadata and linkages) exist as a single logical or physical entity
- **Understandable** through the provision of full metadata.
- **Reconstructable** in terms of the logical, physical and intellectual content.
- **Authentic** records. “Archival science defines authentic records as being what they purport to be – reliable records that over time have not been altered, changed, or otherwise corrupted.” (Dollar, 54)

It is important to keep this concept of preservation in mind when assessing the value of preserving particular census records and in determining the costs of distribution, storage and long-term preservation.

2.2 The value of preservation

Much has been written on the importance of organizing and coordinating the process of census taking within and between countries (United Nations, Department of Economic and Social Affairs, Statistical Division. *Handbook on Census Management,* 2000). Numerous intergovernmental and non-governmental agencies provide support and assistance for this process. Emphasis has been placed on planning, data collection, methodologies, product preparation and dissemination. The value of a strong archival program lies not only in preserving the actual data, metadata and data products for future use, but also in its ability to contribute to future census and statistical activities.

Given the periodic nature of census taking, maintaining records on how specific activities were performed can inform future census processes within a country, allowing agencies

to learn from past processes and strategies. This is particularly important in countries that do not have and cannot afford to have a permanent office for the census. Carefully selected and preserved records can provide detailed information on the planning process, specifications of collection, and insight into why certain decisions were made and how effective particular activities were. In particular, it is these types of country specific processes and approaches that can assist in retaining and building on successful core activities and structures.

Preservation and communication of information on data quality and process evaluation is of value for informing future census activities and is essential for the informed use of census data. Communicating information on the reliability, limitations and strengths of the final data allows users to understand the impact of any procedural changes on any analysis they may wish to perform. This is the type of information that should be encapsulated through logical or physical links between the census data and the procedural metadata in the preservation process.

2.3 Costs of Preservation

The cost of preservation is an issue for all countries. Recent discussions of retention schedules for the 2000 U.S. Census elicited numerous responses from various stakeholder groups concerning both the preservation of original forms and intermediary process output. The cost of preserving original enumeration forms in various formats and the associated cost of making these identifiable for future users was one of the key factors in negotiating a final retention schedule.

In countries without permanent census offices and/or permanent national archives structures, the cost of preservation becomes a major issue. By looking at these costs early and including them in the discussion of the overall costs in undertaking a census, additional options for allocating funds may be found. For example, the way in which census data is captured and prepared for dissemination can reduce the cost of creating a preservation quality record. In addition, capturing and retaining procedural information as it is produced and creating the logical or physical links to emerging data collections, increases the likelihood of preservation while reducing the cost of reconstructing valuable metadata information.

Early discussions of the costs and future value of information preservation allow for both informed decisions and the opportunity to discuss cooperative long-term preservation possibilities in a timely fashion.

3.0 Determining What to Preserve

3.1 Preserving the products

The essential elements of any census in terms of preservation are the resulting data and basic documentation. How that data is identified and defined varies by country. Issues of confidentiality and security play a major role in determining not only who should have access to the microdata and enumeration forms, but also whether that information should

be retained at all. Increasing the availability of microdata contributes to the likelihood that these data will be preserved.

Access to microdata is being made available by an increasing number of countries in a variety of forms: Public samples, scientific samples (restricted to a few carefully screened projects), and through data enclaves where the user works in a secure site and output is tightly controlled. From 1985 through 1994, of 153 countries with populations of one million or more, 134 conducted enumerations in the 1990 round of censuses. 94% of the world's population was counted. 54 countries provided researchers access to anonymized census samples of individuals and households. Some countries restricted access to a single investigator or research facility, but what is remarkable about the 1990s is not only the globalization of the census, but the growing acceptance of anonymized samples as statistical instruments. These trends are continuing in the 2000 round of censuses (1995-2004).

The approach used in the United States of providing public samples of sizes ranging from 1-15% for various area types supports a wide range of research at both the local and national level. In addition, the release of the data from restricted status after 73 years has resulted in a number of projects to make this data accessible to the public in digital format. The most noted of these is the Integrated Public Use Microsample (IPUMS) project. This project, begun in 1992 at the University of Minnesota, integrates sixty-five million microdata records for the United States. Conceived by Steven Ruggles, founding director of the Minnesota Population Center, and funded by the National Science Foundation and the National Institutes of Health, IPUMS integrates the decennial censuses of the United States, dating from 1850 to 1990. The first version of the IPUMS database was released on tape in 1993 and by 1995 via the Internet. Thanks to the expansion of the Internet, the data distribution problem was easily solved by means of a web site driven data dissemination engine (http://www.ipums.org). The IPUMS database, distributed free of charge via the Internet, quickly established itself as one of the three most frequently cited data sources in population research about the United States.

In October 1999, with major funding secured from the National Science Foundation, a global effort was inaugurated, dubbed, IPUMS-International. With the cooperation of national teams of investigators, the IPUMS-International consortium proposes to integrate census microdata for more than a dozen additional countries, with at least one from each continent. Historical census microdata for Canada, Norway, Great Britain, Argentina, and Costa Rica will be included in the database as well as those for the United States. Contemporary microdata for Colombia and the United States will be integrated along with those for France, Brazil, Mexico, Vietnam, Kenya, Great Britain, Hungary, Spain, and others. Based on a prototype developed with the cooperation of the Colombian National Statistical Office (Departamento Administrativo Nacional de Estadística, or DANE), country teams of experienced census data-users are being formed to advise on how to harmonize the national census concepts using international norms.

The creation of public use samples is being used by a variety of countries to increase access to microdata. Software such as the Integrated Microcomputer Processing System (IMPS and its successor CSPro), a system for data processing of censuses and surveys, developed by the International Statistical Programs Center of the U.S. Bureau of the Census, facilitates the dissemination of microdata samples by providing tools for cross tabulation, electronic map production and other basic analysis, thereby reducing the cost of producing these products for individual countries.

Examples of distributed microdata public use samples include Vietnam, which has released a 3% sample of the 1999 Population and Housing Census, with the intention of producing a full 100% sample at a later date. Mexico has released a 10% sample designed to yield valuable information at the level of municipalities of 100,000 or more in size. France has released 5% samples for 1962-1990. Likewise the Central Bureau of Statistics of Kenya has prepared a mega-sample of the 1999 enumeration (with a maximum density of twenty percent) to complete its impressive series of samples for 1969, 1979, and 1989.

These collections not only provide data in a preservable format, they include a range of metadata. The documentation is extraordinarily complete, and includes details on every aspect of the census from earliest preparations to the final publication of tables. The discussion of sampling is particularly noteworthy.

A growing list of countries is offering data in the REDATAM format (developed by the United Nations Demographic Center for Latin America and the Caribbean, CELADE), as a way of storing microdata and making them useful to researchers and administrators who need small area statistics.

REDATAM "**RE**trieval of **DAT**a for small **A**reas by **M**icrocomputer" was originally conceived of as a low cost data retrieval computer program and has grown into a concept that involves a proprietary database format, as well as a software development system. The proprietary format is to secure sensitive data while keeping the invaluable flexibility of microdata access. A web service is also available and benefits national organizations reluctant to give away data but is ready to provide public access to data and/or provide privileged access to selected users. The program is freely available via the Internet. REDATAM has been developed over the last two decades thanks to the financial support from several international organizations (ECLAC-United Nations, UNFPA, the Canadian Government through CIDA and IDRC agencies, IDB and others).
(http://www.cepal.cl/celade)

Countries with 1990 round censuses in REDATAM:

Latin America: Argentina, Brazil, Chile, Colombia, Dominican Republic, Guatemala, Honduras, Nicaragua, Paraguay, Suriname, Uruguay, Venezuela, and English-speaking Caribbean.
Asia: Cambodia* and North Korea*
Africa: Benin, Burkina Faso, Burundi, Cameron, Egypt, Gabon, Ghana, Kenya, Madagascar, Mali, Nigeria, Rwanda*, Seychelles, and Zimbabwe*.

* = database with 100% of the microdata for the population

While these microdata files are not in an archival format in the strict sense, they have been captured in a way that allows for the authoring agency to output a formatted ASCII file with complete structural metadata physically encapsulated to ensure future understandability. It is important that formats such as REDATAM not be viewed as long-term archival formats. The problem with not creating an archival copy and maintaining records in a proprietary format is the cost of eventually having to migrate that information to another format. Proprietary formats soon become legacy formats that due to age, dependency on legacy languages, systems, or hardware becomes difficult, costly and sometimes impossible to migrate.

3.2 Preserving the process

Several manuals and handbooks on performing and managing a national census give detailed lists of procedures and processes. This type of information and details of particular approaches and methodologies are needed for accurately interpreting the resulting data. In addition to this information, consideration of the types of process information that will be of value in preserving institutional memory is useful and often missed. This involves recording and preserving the why as well as the how of the census process. Capturing this information as decisions are made is more cost-effective than reconstructing it at a later date. Attention should be paid to capturing it in a non-proprietary format to reduce the likelihood that the information will be lost due to migration costs.

The complete census cycle consists of four phases (United Nations, Department of Economic and Social Affairs, Statistical Division. *Handbook on Census Management)*:

- Preparation
- Field Operations
- Data Processing
- Evaluation

For each phase, the following documentation is of particular interest:

- reports on procedures and methods.
- comparison of concepts and procedures with the preceding census and current international standards
- evaluation reports for each cycle of the census and the most important documents on which the reports are based
- record books (from manager of the census to the enumerator's logbook) although these may be too raw for general dissemination

The final step is to document the data disseminated and the associated documentation and notes.

4.0 How to assess future value

4.1 Who are the stakeholders

As is clear from the above discussion, standards for census microdata and microdata access are still emerging. The major question of preserving and allowing access to microdata is no longer a technical one but a question of policy. As access to and use of census data expands, the character and complexity of stakeholder groups also expands. These stakeholder groups will be found among governmental, non-governmental, academic, commercial and new user groups. While there will always be competing interests for what should be retained, common interests among several groups will help to identify materials for long-term preservation. Consulting these stakeholder groups early in the process increases the likelihood of obtaining and maintaining funding for long-term preservation.

4.2 Future impact and anticipated use

In terms of the future impact and use patterns for census data, preserving and retaining access to microdata holds the greatest potential. Census data is used to address specific social, economic and demographic issues that change in character over time. The ability to create comparable aggregations over time or for new emerging geographic areas or definitions rest solely on the retention of microdata files. This is particularly true for small area statistics that cannot be derived from larger area aggregations.

For topical tabulations the difficulties are clear. The timing cycle of the census and the production of statistical aggregations mean that the questions asked and the tables created often reflect the concerns and interests from five to ten years prior to publication. New tabulations may no longer be comparable to tabulations from previous censuses due to change in cohort or classification groupings, universe for the table or other changes in definition. Without access to microdata, however it is secured, researchers and analysts are left with few options. In addition, microdata lends itself to research and scholarly discourse and increases the value derived from an individual census.

Consider the example of Canada: In the 1970s National Statistical Services began to disseminate census microdata samples in growing numbers. In Canada, the 1971 revision to the Statistics Act made possible the public release of non-confidential microdata (Tambay and White 2001). Since the 1970s Statistics Canada, with its series of quinquennial enumerations, has regularly issued census microdata samples. Until 1996, researchers had to request samples individually and distribution was highly restricted. In that year a data liberation initiative was instituted to permit Canadian universities to disseminate microdata samples to researchers and their students. The result was an explosion of research. While before liberation five or ten scholars might acquire microdata samples per year, afterward, a single sample at a major university might be accessed hundreds of times per month. The profusion of suppliers means that usage statistics are now impossible to compile, where before the agency recorded every user by name. Given the widespread use of census microdata in the university classroom, Canadian scholars are educating a younger generation of citizens about the utility of the census and democratizing access to census data (Lisa Dillon, private communication, April 21, 2001).

In the United Kingdom, public use census samples called SARs (Samples of Anonymized Records) were first constructed for the 1991 enumeration, with a sample density of 2.0% for individual records, and 0.5% for households. Administrative units with fewer than 120,000 inhabitants were not identified. Notwithstanding the small density of the samples and the absence of geographical detail there was an explosion of research using the SARs. Hundreds of studies were published within six years of the initial release of the data. In anticipation of the 2001 enumeration, disclosure risks were re-assessed, now taking into account error and coding variability as well as differences in timing and coding schemes between datasets. With the permission of the Office of National Statistics privileged access was granted to attempt to match survey records against the SARs (Dale and Elliot, in press). The purpose was to test the practical, as opposed to the theoretical, risks of identifying individuals by matching two sources. The authors reasoned that prior assessments of the likelihood of identifying individuals exaggerated the risks because they neglected to take into account error, differences in timing and incompatibilities of coding schemes. From this rigorous exercise in sleuthing Dale and Elliot conclude:

> For a user of an outside database, attempting this sort of match with no opportunity for verification would prove fruitless. In the first place, the small degree of expected overlap would be a considerable deterrent to an intruder. However, if a match between the two files was attempted the large number of apparent matches would be highly confusing as an intruder would have no way of checking correct identification.

4.3 Informing future censuses

The availability of process information specific to a country and/or organizational structure can inform preparations for future censuses by providing a clear picture of what took place, how it took place, why it was handled in a particular manner, and the successes and difficulties that occurred. This type of information is particularly important for countries with no permanent office or with minimal permanent staff that must essentially create a new system with each census. Providing documentation of why processes and procedures were followed in a specific way is also helpful to international technical consultants, providing them with a clear and well-rounded picture of the previous census activities.

5.0 Inventory of available technology/personnel/knowledge

Prior to the 1990 census round, the United Nations Statistical Division distributed a questionnaire concerning general coverage of the census, organizational structure, cartographic work, house and/or household listing, testing (pre-tests, pilots, etc), the census questionnaires, enumerators and supervisors, enumeration, sampling, data processing, evaluation and analysis, data dissemination, costs and future activities. In addition to the information already requested, the following areas of information related to long-term preservation would be useful in helping countries integrate the discussion of preservation early in the process. Early recognition of preservation needs and possibilities will help in making informed preservation decisions.

- Presence of a national archive capable of preserving digital materials
- Existence of a preservation plan including a record retention schedule
- Types of materials (data products, metadata, process documents) currently preserved and planned for preservation
- Availability of trained preservation staff
- Use of external depositories/archives for data and documentation

6.0 Conclusion

If census microdata are to become widely used, issues of statistical confidentiality must be resolved to the satisfaction of the national statistical agencies and the public as well as researchers. Eurostat sponsored five international conferences on the subject over the past decade. Thanks in part to these efforts and others, the standard practice is now to prepare microdata samples for a variety of users. Among the 52 member-states in the International Monetary Fund's General Data Dissemination System, almost three of every four disseminate census microdata samples, in one guise or another. The development of international microdata standards will increase further the availability of census samples, thereby facilitating comparative research, both in time and space. Everywhere that public dissemination policies have been adopted, an explosion in research has resulted, without a single instance of a breach, or even the allegation of a breach, in statistical confidentiality.

Understanding and incorporating this concept of preservation is important in that it ensures that census data will be protected from loss, alteration, and deterioration. "In this regard , the obligation of archivists is to explain to computer specialists, information technology specialists, and others who are unfamiliar with archives the importance of a physical or logical space, 'independent of the production environment,' where records are protected from loss, alteration, and deterioration so that they may be used as trustworthy evidence as far into the future as is necessary. This is what archiving should be about." (Dollar, 26)

References

Dale, Angela and Mark Elliot. In press. "Proposals for 2001 SARS: An assessment of disclosure risk," *Journal of the Royal Statistical Society, Series A.*

Dollar, Charles M. *Authentic Electronic Records: Strategies for Long-Term Access*. (Chicago, IL:Cohasset Associates, Inc., 2000)

Mexico. Instituto Nacional de Estadística, Geografía e Informática. *Contar 2000. Sistema para la consulta de tabulados y base de datos de la muestra: XII Censo General de Población y Vivienda 2000*. (Aguascalientes, Mexico: 2001).

Ruggles, Steven, Catherine A. Fitch, Patricia Kelly Hall, Matthew Sobek. 2000. "IPUMS-USA: Integrated Public Use Microdata Series for the United States," in Patricia

Kelly-Hall, Robert McCaa and Gunnar Thorvaldsen, eds., *Handbook of International Historical Microdata for Population Research*, Minneapolis MN, 259-284.

Ruggles, Steven, J. David Hacker, and Matthew Sobek. 1995. “Order out of chaos: General design of the Integrated Public Use Microdata Series.” *Historical Methods* 28: 33-39.

Secretariat. 2001. "Report of the March 2001 Work Session on Statistical Data Confidentiality," Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje.

Tambay, Jean-Louis and Pamela White. 2001. "Providing greater accessibility to survey data for analysis," Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje March.

United Nations, Department of Economic and Social Affairs, Statistics Division. *Handbook on Census Management for Population and Housing Censuses, Studies in Methods, Series F No. 83*, (New York: United Nations, 2000)

United Nations, Department of Economic and Social Affairs, Statistics Division. *Handbook Population and Housing Censuses: Part 1 Planning, Organization and Administration of Population and Housing Censuses, Studies in Methods, Series F No. 54*, (New York: United Nations, 1992)

United Nations, Department of Economic and Social Affairs, Statistics Division. *Handbook Population and Housing Censuses: Part 2 Demographic and Social Characteristics, Studies in Methods, Series F No. 54*, (New York: United Nations, 1992)

United Nations, Department of International Economic and Social Affairs, Statistics Division. *Manual on Population Census Data Processing using Microcomputers, Studies in Methods, Series F No. 53*, (New York: United Nations, 1990)

United Nations, Department of International Economic and Social Affairs, Statistics Division. *Emerging Trends and Issues in Population and Housing Censuses, Studies in Methods, Series F No. 52*, (New York: United Nations, 1991)

Vietnam. General Statistics Office. *Data and results from the 3% sample of The Population and Housing Census*. (Hanoi: Central Data Processing Centre, August 2000).