PROTOCOLO DE ANONIMIZACIÓN
USAID/Colombia Monitoring, Evaluation and Learning Activity

PROTOCOLO DE ANONIMIZACIÓN

USAID/Colombia Monitoring, Evaluation and Learning Activity

I. JUSTIFICACIÓN

Panagora Group es una PYME norteamericana que brinda servicios de monitoreo y evaluación en la industria del desarrollo internacional y actualmente ejecuta la Actividad Monitoreo, Evaluación y Aprendizaje de USAID/Colombia (MEL por sus siglas en inglés). La Actividad MEL, proporciona servicios técnicos y de asesoría a USAID en los siguientes componentes: (1) Monitoreo; (2) Evaluaciones de desempeño e impacto; (3) investigación, evaluación y análisis de datos; (4) servicios de información geográfica (SIG) y (4) Actividades de Colaboración, Aprendizaje y Adaptación (CLA por sus siglas en inglés).

MEL proporciona a la Misión de USAID servicios técnicos y de asesoría, para facilitar la toma decisiones informadas respecto a su gestión en Colombia, con el fin orientar su estrategia de mediano y largo plazo, desde el Componente de Evaluaciones de Desempeño e Impacto. Bajo el liderazgo del experto en evaluaciones y con la asistencia de dos especialistas en el área, se realizan evaluaciones de impacto y de desempeño de los proyectos y actividades seleccionados por USAID/Colombia siguiendo los principios, políticas y guías descritos en el ADS 201.

USAID establece en el ADS 201 los Principios de Evaluación para cada evaluación de desempeño e impacto, que el Componente II de la Actividad MEL realiza. La transparencia es uno de estos principios, como se señala en el ADS 201.3.5.10, "Las conclusiones de las evaluaciones se compartirán lo más ampliamente posible, con el compromiso de una divulgación completa y activa". Como mínimo, esto requiere la publicación de los informes de las evaluaciones en el Centro de Intercambio de Información sobre Experiencias de Desarrollo (DEC) y de los datos de las evaluaciones en la Biblioteca de Desarrollo de Datos (DDL). Esto implica que todos los datos cuantitativos y cualitativos, que consisten en bases de datos, transcripciones de audio y encuestas, deben anexarse al DDL.

Cada entrevistado, informante clave e interesado que proporcione información primaria en una evaluación debe mantener su privacidad protegida antes de que los datos se publiquen en el DDL. Todos los participantes deben dar su consentimiento informado antes de la aplicación de cada instrumento, en el que se establece que se concede el anonimato a todos los informantes que acepten participar. Por esta razón, se requiere la aplicación de este protocolo de anonimización de información para mantener su confidencialidad.

II. OBJETIVO DEL PROTOCOLO DE ANONIMIZACIÓN

El protocolo de anonimización tiene como objetivo definir las directrices para el proceso de anonimización de la información de todos los participantes que proporcionen datos en una evaluación. El protocolo procura garantizar que no sean identificados con un esfuerzo razonable por los usuarios de la información; USAID o el Socio Implementador del programa que se va a evaluar. El protocolo de anonimización se basa en el Archivo de Datos de Ciencias Sociales de Finlandia (FSD) para la des-identificación y anonimización[1].

III. MARCO CONCEPTUAL

1. Información personal:

Según la definición que figura en el Reglamento General de Protección de Datos (RGPD), por "información personal" se entiende toda información relativa a una persona natural identificable. Se considera que una persona natural es identificable si puede ser identificada, directa o indirectamente, en particular haciendo referencia a un identificador como un nombre, un número de identificación, datos de localización o más factores específicos de la identidad física, fisiológica, genética, mental, económica, cultural o social de esa persona natural. (Reglamento general de protección de datos de la UE, artículo 4, párrafo 1).

2. Información identificable

Una persona se considera identificable si puede ser identificada directa o indirectamente. La información que permite la identificación de una persona se conoce como identificador. Existen tres tipos de identificadores:

a) **Identificadores directos**: Información que es suficiente por sí sola para identificar a un individuo. Incluye el nombre completo de una persona, el número de identificación, la dirección de correo electrónico que contiene el nombre personal y los identificadores biométricos (huellas dactilares, **imagen** facial, patrones de voz, escaneo del iris, geometría de la mano o firma manual).

b) **Identificadores indirectos fuertes**: Es la información, que no sea de identificación directa, puede utilizarse para identificar a una persona con bastante facilidad. Incluye la dirección, el número de teléfono, el número de matrícula del vehículo, la cita bibliográfica de una publicación del individuo, la dirección de correo electrónico que no sea en forma de nombre personal, la dirección web de una página que contenga datos personales, un cargo poco común, una enfermedad muy rara, un evento poco común o un cargo que ocupe sólo una persona a la vez (por ejemplo, el presidente de una organización).

[1] https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html

c) **Identificadores indirectos (o cuasi-identificadores):** Información que por sí sola no es suficiente para identificar a alguien pero que, cuando se vincula con otra información disponible, puede utilizarse para deducir la identidad de una persona. Incluye la edad, el género, la educación, la situación laboral, la actividad económica y ocupación, la situación socioeconómica, la composición del hogar, los ingresos, el estado civil, la lengua materna, el origen étnico, el lugar de trabajo o de estudio y las variables regionales. Los identificadores indirectos relativos a la región de residencia incluyen, por ejemplo, el vecindario, el municipio y la región. La fecha de nacimiento, la fecha de fallecimiento o las fechas de acontecimientos de interés periodístico también pueden ser identificadores indirectos.

3. Información anónima

Existe cuando no se puede volver a identificar a una persona con un esfuerzo razonable sobre la información proporcionada o combinando datos adicionales. En otras palabras, los datos son anónimos si los atributos característicos (por ejemplo, las combinaciones de ciertos identificadores indirectos) pertenecen a más de una persona y no se puede identificar a un sujeto con un esfuerzo razonable. No existen datos completamente anónimos, pero con procedimientos bien ejecutados se puede lograr un resultado en el que las personas individuales no puedan ser identificadas con un esfuerzo razonable. La anonimización se refiere a las diversas técnicas e instrumentos utilizados para lograr anonimidad. Para que los datos cuenten como anónimos, la anonimización debe ser irreversible.

4. Información pseudónima

Existe cuando no se puede volver a identificar a una persona sobre la información pseudonima sin información adicional y separada. La pseudonimización se refiere a la eliminación o sustitución de los identificadores por pseudónimos o códigos, que se mantienen por separado y están protegidos por medidas técnicas y de la organización. Las medidas de organización se refieren a la protección del entorno físico y al control de acceso a los documentos. Las medidas técnicas incluyen, por ejemplo, el almacenamiento seguro de datos y su codificación.

Los datos siguen siendo pseudónimos mientras exista la información de identificación adicional. Los datos pseudónimos se convierten en anónimos cuando se destruye la información de identificación que se mantiene por separado (clave de descifrado, datos personales e información sobre las técnicas utilizadas para hacerlos pseudónimos). Los datos son anónimos si no se pueden vincular a los datos personales originales con un esfuerzo razonable.

IV. PROCEDIMIENTOS PARA ANONIMIZAR INFORMACIÓN CUANTITATIVA[2]

Esta sección del documento presenta el protocolo de anonimización referente a la información de carácter cuantitativo levantada en una evaluación. Para fines de este documento, se entenderá por anonimización de datos cuntitativos como el proceso a través del cual se realizan procedimientos sobre las fuentes de información primarias con el objetivo de impedir la identificación de las unidades de información (ej. Individuos) con base en un conjunto de microdatos.

El resto de esta sección se organiza de la siguiente manera. En el aparte de Requisitos para la anonimización, se describen los insumos necesarios para proceder a realizar la anonimización. El aparte de Identificación y tratamiento potencial señala el paso a paso para identificar las variables que sufrirán algún tipo de transformación con el objetivo de impedir la identificación de las unidades de información. El aparte Anonimización y contraste describe el paso a paso en la aplicación de las trasformaciones necesarias y la realización de ejercicios de comparación entre los datos anonimizados y no anonimizados. Finalmente, el aparte Disposición y entrega describe los archivos que deben ser puestos a disposición de la Actividad MEL una vez se haya concluido la aplicación del protocolo.

1. Requisitos para la anonimización

Para iniciar la aplicación de este protocolo, se debe disponer de los siguientes dos insumos:

- Base de datos incluyendo todas las variables levantadas o anexadas (ej. factores de expansión, datos provenientes de listados de beneficiarios, etc.)
- Diccionario de datos.

Una vez se cuente con los dos insumos anteriores, se procede al siguiente aparte del protocolo.

2. Identificación y tratamiento potencial

El segundo paso del protocolo se enfoca en identificar dentro de la base de datos cuales son las variables que permiten la identificación de las unidades de información con base en el conjunto de los microdatos. Es importante indicar que una vez se haya completado este aparte del protocolo, se debe enviar el resultado de este (diccionario de datos con las variables agregadas durante este aparte) al Líder del componente de evaluación de la Actividad MEL o a quién este designe, quién validará el ejercicio y dará

[2] El presente protocolo utiliza como documentos de referencia *Guía para la anonimización de bases de datos DANE* https://www.dane.gov.co/files/sen/registros-administrativos/guia-metadatos.pdf. *Lineamientos para la anonimización de datos del sistema nacional de estudios y encuestas poblacionales para la salud, Ministerio de Salud y Protección Social* https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/lineamientos-anonimizacion-sistema-encuestas.pdf

la aprobación para continuar con los siguientes apartes del protocolo. En este aparte del protocolo se realizan en los dos siguientes pasos.

a) Identificación

En este paso, se toma el diccionario de datos y se agrega una columna al mismo asignando a cada una de las variables una de las etiquetas descritas a continuación:

i. **Identificación geográfica:** Variables que contienen información ya sea codificada o no de ubicación a nivel departamento, municipio, región, área o zona (urbano, rural), vereda, barrio, vivienda, hogar, sector, sección, manzana, UPM, USM, segmento y cualquier otra variable de este tipo, contenida en el formulario de la encuesta o sido heredadas de la muestra de forma directa o generadas por los cálculos de factores de expansión.

ii. **Identificación o contacto directo de unidades de información**: Variables que contienen datos relacionados con los encuestados y personal del equipo de levantamiento como son nombres, apellidos, número de documento de identificación, número de teléfono, número de orden en con el que fue listado en el hogar, correos electrónicos, nombres en redes sociales, dirección.[3]

iii. **Otros:** Variables que contienen datos relacionados con preguntas numéricas o categóricas que no permiten una identificación directa de las unidades de información. Edad, nivel educativo, área del predio, identificador de unidades de la misma vivienda, identificador de unidades del mismo hogar, percepción sobre algún elemento particular, etc.

b) Tratamiento potencial

En este paso, se toma el diccionario de datos y se agrega una columna adicional, asignando a cada una de las variables una de las etiquetas descritas a continuación:

i. **Permanece en base anonimizada:** Variables que no permiten la identificación directa o contacto de las unidades de información. Todas las variables bajo la etiqueta Otras son asignadas a este grupo. Respecto a las variables de identificación geográfica, todas las variables correspondientes a agregaciones espaciales con más de 50 unidades de información en cada una de las categorías de la variable respectiva deben ser marcadas bajo esta etiqueta a menos que se indique lo contrario (ej. si todos los municipios incluidos en el levantamiento tienen más de 50 unidades de

[3] Aunque la dirección es una variable de carácter geográfico, se incluye dentro de este grupo ya que permite una ubicación precisa de la unidad de información

información, entonces esta variable se marca bajo la categoría permanece en base anonimizada).[4]

ii. **Se elimina de la base anonimizada:** son variables que permiten la identificación de las unidades de información con base en el conjunto de microdatos. Todas las variables bajo la etiqueta Identificación o contacto directo de unidades de información son asignadas a este grupo. Respecto a las variables de identificación geográfica, todas las variables correspondientes a agregaciones espaciales con 50 o menos unidades de información en cada una de las categorías de la variable respectiva deben ser marcadas bajo esta etiqueta a menos que se indique lo contrario.

3. Anonimización y contraste

Una vez se ha recibido validación por parte del Líder del componente de evaluación o quién este designe respecto a los resultados del aparte de identificación y tratamiento potencial, se procede a ejecutar la anonimización propuesta. Una vez se haya realizado este paso, se deben realizar las siguientes validaciones:

i. **Revisión:** Examinar a través de tablas de frecuencia que el número de las observaciones de ambas bases de datos (base completa y anonimizada) tengan el mismo número de observaciones.

ii. **Estadísticas descriptivas:** Seleccionar al menos 4 variables correspondientes a la categoría *Otras* e *Identificación geográfica* que permanecieron en la base anonimizada y realizar estadísticas descriptivas con las mismas. Para el caso de variables continuas contrastar la igualdad en la media, varianza, mínimo, máximo y la mediana (p50) entre las dos bases de datos, para el caso de variables categóricas contrastar la igualdad a través de tablas de frecuencia.

4. Disposición y entrega

Una vez se hayan completado los pasos anteriores, y en el marco de los productos acordados contractualmente, se le hará entrega a la Actividad MEL del contenido en la siguiente lista:

- Base de datos incluyendo todas las variables levantadas o anexadas (ej. factores de expansión, datos provenientes de listados de beneficiarios, etc.)
- Diccionario de la base de datos.
- Base de datos producto de la aplicación del protocolo de anonimización
- Diccionario de la base de datos producto de la aplicación del protocolo de anonimización.

[4] Para la aplicación del criterio de unidades de información se debe tomar la unidad mínima de información. Es decir, si se están encuestado individuos correspondientes a hogares, el criterio debe ser aplicado con respecto al número de individuos.