COLOMBIA POTENCIA DE LA VIDA

Gobierno del Cambio

# Guía para la anonimización de datos estructurados

DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

SEN 2.0
Sistema Estadístico Nacional-Colombia

# Departamento Administrativo Nacional (DANE)


## Archivo General

### ARCHIVO GENERAL DE LA NACIÓN JORGE PALACIOS PRECIADO

Establecimiento público adscrito al Ministerio de Cultura

**Laura Sánchez Alvarado**
Directora General (E) del Archivo General de la Nación Jorge Palacios Preciado

### Colaboradores:

**Adela Díaz Acuña**
Subdirectora de Transformación Digital e Innovación Archivística

**Carlos Enrique Rojas Núñez**
Profesional Especializado de la Subdirección de Transformación Digital e Innovación Archivística

## DANE

**Julieth Alejandra Solano Villa**
Directora Técnica de la Dirección DIRPEN

**Andrea Milena Roncancio Sánchez**
Coordinadora GIT Prospectiva y Análisis de Datos

### Profesionales:

**Alexander González Coca**

**Johana Catherine Avila Alvarado**

**Gildardo Andrés Vargas Acuña**

**Bryan Felipe Lugo**

**German Leonidas Orjuela**

**Francisco Javier Lara Carrillo**

**Daniel Felipe Ortiz**

**Cristian Alejandro Ruge**

**Mónica Andrea Quiroga Rivera**

# Contenido

# Lista de tablas

# Lista de ilustraciones

# Introducción

Los Sistemas Estadísticos Nacionales (SEN) tienen como propósito proveer información estadística oficial, relevante, comprensiva, confiable y objetiva a los diferentes usuarios, siendo la información un elemento fundamental para la toma de decisiones. Este propósito implica importantes retos para los productores de información estadística, especialmente al observar una mayor demanda de información desagregada, un aumento en el uso de microdatos y una mayor necesidad de explotar los datos disponibles en los sistemas estadísticos.

El SEN colombiano establece dentro de sus objetivos promover entre sus miembros el acceso y el uso de microdatos para la producción y la difusión de estadísticas oficiales a nivel nacional y territorial que requiere el país de manera organizada y sistemática (Art. 5. Ley Estadística 2335 de 2023). De igual forma, el Código Nacional de Buenas Prácticas Estadísticas del SEN, en su principio 10 sobre accesibilidad de la información, incentiva a los miembros del SEN para que implementen prácticas que permitan el acceso de las estadísticas y los microdatos asociados a todo tipo de usuarios con el máximo detalle posible y en diferentes formatos y medios que faciliten la consulta, la visualización y el uso (DANE, 2017: 10). Además, el principio 11 del Código que trata sobre la confidencialidad, pretende incentivar en los productores de información estadística, así como la utilización de técnicas para la anonimización de los microdatos para garantizar la protección de la identificación o la localización geográfica de las fuentes empleadas en el proceso estadístico (DANE; 2017: 11).

Bajo estas premisas y adoptando la Ley Estatutaria 1266 de 2008 que establece el Habeas Data y regula el manejo de la información contenida en bases de datos personales, financieras, crediticias, comerciales, de servicios y provenientes de terceros países, además de establecer disposiciones sobre la recolección, el tratamiento y la circulación de datos personales en el país (Congreso de Colombia, 2008) el Departamento Administrativo Nacional de Estadística (DANE), en su rol de coordinador del SEN, presenta la Guía para la anonimización de datos estructurados, cuyo propósito se centra en orientar a los integrantes del SEN sobre el proceso de anonimización de bases de datos que provienen de registros administrativos, operaciones estadísticas y fuentes secundarias.

Este documento se construye a partir de la experiencia del DANE y el Archivo General de la Nación (AGN), al implementar procesos propios de anonimización en distintas bases de datos y tomando como referencia experiencias nacionales e internacionales de oficinas de estadística, ministerios, secretarías, entre otros. Se espera que los integrantes del SEN puedan identificar en este documento las buenas prácticas, las herramientas y los instrumentos cuando se implementen procesos de anonimización para la producción y la publicación de sus propias estadísticas. La guía también presenta a lo largo de las etapas, la ejemplificación del proceso mediante el uso de una base de datos simulada, siendo este un insumo para que las entidades del SEN interesadas puedan seguir paso a paso el proceso y comparar los resultados que se presentan aquí.
La guía se divide en seis partes: la primera parte se presentan antecedentes nacionales e internacionales sobre los usos de la anonimización; en la segunda parte se relaciona el objetivo y el alcance del documento; en la tercera se presenta el marco conceptual para el proceso de anonimización; en la cuarta parte se relaciona la planeación del proceso de anonimización; en la quinta parte se expone el proceso de anonimización, y finalmente, en la sexta parte se presentan algunas recomendaciones finales que deben ser tenidas en cuenta al momento de aplicar el proceso de anonimización.

# 1. Antecedentes nacionales e internacionales frente al uso de la anonimización

Son varias las experiencias internacionales que permiten observar la implementación de la anonimización, dado que su principal propósito es incrementar la desagregación de la información, así como mantener sus niveles de confidencialidad y generar un mayor aprovechamiento estadístico de la misma. Esta sección presentará algunas experiencias de países como Reino Unido, Estados Unidos y Países Bajos y los avances que se han tenido en el contexto nacional. Asimismo, se exponen experiencias y metodologías aplicadas a la anonimización de censos y herramientas o softwares utilizados en la aplicación de los procesos de anonimización.

## 1.1. Experiencia del DANE en los procesos de anonimización

El DANE ha demostrado un compromiso constante con la transparencia y la promoción del acceso a la información estadística en Colombia. Desde 2011 la entidad ha experimentado un proceso de transformación en la forma en que pone a disposición sus bases de datos, incluyendo encuestas de hogares, calidad de vida y encuestas estructurales como industria, comercio y servicios.

Anteriormente, el DANE operaba bajo la figura de convenios interinstitucionales establecidos con universidades y entidades del orden nacional para el suministro de sus bases de datos. Además, comercializaba algunas de sus operaciones estadísticas, permitiendo a investigadores y personas interesadas acceder a dichos datos previo pago, para ello, la entidad contaba con una dirección de mercadeo encargada de estas actividades. Sin embargo, con el objetivo de fomentar una mayor transparencia y apertura de datos, así como la de asegurar el acceso público a la información generada por la entidad, el DANE ha adoptado una nueva estrategia, en lugar de mantener un enfoque basado en la comercialización de datos, puso a disposición del público en general sus bases de datos a través de su catálogo central, el Archivo Nacional de Datos (ANDA)[1]; esta transformación ha implicado cambios en la estructura organizativa al reemplazar la dirección de mercadeo por una dirección enfocada en potencializar la cultura estadística y promover el acceso abierto a la información.

Esta evolución ha permitido al DANE ampliar el acceso a sus operaciones estadísticas y facilitar su uso por parte de entidades públicas, entidades privadas, investigadores y el público en general. Igualmente, ha fortalecido la transparencia en la publicación de estadísticas del país y ha fomentado un mayor aprovechamiento de la información por parte de diferentes actores, impulsando así el desarrollo de la investigación y el análisis estadístico en Colombia.

[1] Disponible en: https://www.dane.gov.co/index.php/servicios-al-ciudadano/tramites/transparencia-y-acceso-a-la-informacion-publica/informacion-de-interes?highlight=WyJhbmRhIl0=.

De igual manera, la entidad cuenta con experiencia en los procesos de anonimización de bases de datos para garantizar la reserva estadística de las fuentes de datos y, al mismo tiempo, facilitar el máximo aprovechamiento de la información. En la actualidad, posee una amplia gama de operaciones estadísticas certificadas que incluyen la disponibilidad de microdatos y estas operaciones engloban principalmente unidades de observación como personas, empresas y establecimientos comerciales.

Este escenario presenta un desafío pues se debe garantizar la privacidad y la confidencialidad de sus fuentes y la información suministrada. Este hecho se refleja en el esfuerzo continuo de la entidad por promover el uso de técnicas y buenas prácticas en la producción de cada una de las operaciones estadísticas realizadas.

En la última década, el DANE ha realizado diversas actividades en el ámbito de la anonimización de bases de datos, lo que ha impulsado la exploración de nuevas soluciones. Estas actividades incluyen:

- La publicación de microdatos anonimizados que implica el uso de técnicas de anonimización que permitan a los usuarios replicar la mayoría de los indicadores generados con las bases de datos.

- La implementación de procesos de anonimización de bases de datos tradicionalmente no publicadas en forma completa. Sin embargo, el DANE se encuentra en el proceso de publicar microdatos del Censo Nacional Agropecuario (CNA) 2014, del Censo de Edificaciones (CEED) 2017-2023 I Trimestres, los Censos de Habitantes de Calle (CHC) 2019, 2020, 2021 y el Censo de Población y Vivienda 2018.

- La exploración de técnicas avanzadas de perturbación en los procesos de anonimización de datos, abarcando métodos de mayor complejidad, como la generación de datos sintéticos y la aplicación de técnicas como SWAP y PRAM, entre otras.

- La exploración y la implementación de tecnologías de privacidad en el análisis de datos sensibles, como la utilización de enclaves seguros, entornos de ejecución protegidos por hardware que permiten realizar cálculos en datos sensibles sin exponerlos a procesos no autorizados. Además, se ha considerado la aplicación de la privacidad diferencial, una técnica que agrega ruido aleatorio a los datos para preservar la privacidad de los individuos representados en ellos.

- El DANE ha observado la creación del Laboratorio de Datos de las Naciones Unidas, con su programa piloto enfocado en mejorar la privacidad en el intercambio internacional de datos. Este programa, conocido como el UN PET Lab, ha establecido colaboraciones con oficinas nacionales de estadística para abordar los desafíos de privacidad en el análisis de datos sensibles. Estas iniciativas internacionales refuerzan la necesidad de encontrar tecnologías y algoritmos que promuevan la privacidad en el contexto de la preservación de datos sensibles.

- La gestión de la privacidad en un entorno de cambios tecnológicos acelerados por las tecnologías de la información y la comunicación, así como al aprovechamiento del big data. En este contexto, es crucial abordar la vinculación de las fuentes de datos tradicionales con los accesos públicos de diferentes entidades y la información proporcionada por los individuos, con la creciente divulgación de datos a través de redes sociales y entidades privadas, existe la posibilidad de que la información suministrada sea vinculable, lo que plantea un desafío para garantizar la privacidad.

- El desarrollo de algoritmos eficientes para la anonimización de bases de datos de gran volumen. El DANE se enfrenta al desafío de implementar paquetes especializados en la anonimización de datos que sean capaces de manejar eficientemente bases de datos de gran volumen. Actualmente, los programas como SDCmicro, diseñados para tratar bases de

datos de tamaño muestral, presentan limitaciones cuando se supera el límite de 30.000 registros, lo que ocasiona problemas computacionales significativos. Esta limitación restringe su uso en fuentes de datos más grandes, como registros administrativos y censos.

- El aprovechamiento de fuentes secundarias como imágenes satelitales, datos de redes sociales como X y Facebook. Además, se vislumbra la posibilidad de utilizar otras fuentes audiovisuales, como audio, video y fotografías, para las que se ha ampliado recientemente la disponibilidad de algoritmos de análisis. Esto plantea la necesidad de considerar el tratamiento adecuado que se debe dar a estas fuentes de datos no convencionales. Para abordar este reto, el DANE debe establecer marcos éticos sólidos que guíen el uso responsable de estas fuentes de datos, tomando en cuenta los términos y las condiciones que estas redes establecen para su uso. Asimismo, se requiere desarrollar técnicas de anonimización adaptadas a estas nuevas fuentes, teniendo en cuenta su naturaleza y características específicas. Es esencial garantizar la confidencialidad y la privacidad de los individuos representados en estos datos y aplicando rigurosos procesos de anonimización que preserven su identidad.

La experiencia acumulada por el DANE en los procesos de anonimización, ha demostrado su capacidad para proporcionar datos de calidad, preservando la privacidad de los individuos. Este conocimiento adquirido puede ser aprovechado para la construcción de una guía de anonimización de datos más actualizada, que promueva la privacidad, la confidencialidad y el uso ético de la información estadística. La entidad sigue siendo líder en la aplicación de buenas prácticas en la anonimización de datos en Colombia y su experiencia puede contribuir significativamente a mejorar y actualizar las técnicas y los enfoques utilizados en este campo.

## 1.2. Experiencia del Archivo General de la Nación (AGN) en procesos de anonimización

El AGN coordina el Sistema Nacional de Archivos y desarrolla procesos de organización y fortalecimiento de los archivos de la administración pública a nivel nacional y territorial. Además, verifica el cumplimiento de la normativa archivística vigente por medio de la inspección, la vigilancia y el control de los archivos de todas las entidades del Estado y personas naturales y jurídicas de derecho público y privado, de acuerdo con la normatividad vigente.

En este contexto, el AGN en articulación con el Ministerio de Tecnologías de la Información y las Comunicaciones, la Superintendencia de Industria y Comercio, el Departamento Administrativo de la Función Pública, el Departamento Nacional de Planeación y el DANE desarrollaron una *Guía de anonimización de datos estructurado*[2] cuya finalidad es precisar conceptos y proporcionar los lineamientos metodológicos para realizar procesos de anonimización de datos personales e información producida o gestionada por entidades.

El AGN, además, aplica los procesos de anonimización en los siguientes trámites:

**Emisión de Conceptos Técnicos:** en cumplimiento de las funciones dadas al AGN por las Leyes 80 de 1989 y 527 de 2000, este responde las consultas técnicas sobre temas archivísticos que son formuladas por ciudadanos, entidades públicas y empresas privadas.

Dado que dichos conceptos sirven para resolver inquietudes de otros ciudadanos u organizaciones, se decidió publicarlos en la página web de la entidad. Para ello se impartió una directriz interna en la que se solicita a los profesionales encargados de la redacción de estos documentos omitir en

[2] Disponible en https://www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Consulte/Recursos/Publicaciones/2021-06-25_Guia_de_Anonimizacion.pdf

los mismos cualquier información que pueda identificar al solicitante, ya sea persona natural o jurídica; para la remisión de la respuesta se elabora un oficio al cual se le anexa el correspondiente concepto.

**Certificación de tiempo de servicio a Exfuncionarios del DAS:** el AGN recibió la documentación de Historias Laborales y Nóminas del extinto DAS, por lo cual es responsable de atender las solicitudes de Certificaciones de Tiempos de Servicio y de Aportes pagados por pensión, provenientes de los exfuncionarios, de los Fondos de Pensiones y de la Rama Judicial, las cuales deben ir acompañadas de los soportes documentales, en especial de la nómina; para este último proceso, se realiza la digitalización de los folios correspondientes en formato PDF a cuyas imágenes digitales se les realiza un enmascaramiento de la información sensible en color negro para suprimir toda información que corresponda al solicitante, con lo cual se evita que datos personales de otros funcionarios quede expuesta.

## 1.3. Antecedentes en Colombia

En Colombia, el DANE ha sido pionero en el desarrollo de los procesos de anonimización de datos estructurados. En 2014 se publicaron los Lineamientos para la anonimización de microdatos que identificaban tres grandes etapas en el proceso: preanonimización, anonimización de microdatos de uso interno y anonimización (DANE, 2014: 11).

El Ministerio de Salud, basado los lineamientos del DANE, generó sus propios lineamientos para anonimizar microdatos a través del documento Lineamientos para la Anonimización de Datos del Sistema Nacional de Estudios y Encuestas Poblacionales para la Salud (Ministerio de Salud, s.f.). En este ejercicio, el Ministerio incluyó las etapas previstas por el DANE y agregó con mayor detalle, las técnicas de anonimización que se usarían en sus bases de información; específicamente presenta el uso de los métodos de perturbación o los métodos de reducción en la información (Ministerio de Salud, s.f.:12-13).

Adicional a los lineamientos y las orientaciones en materia de anonimización, Colombia ha gestionado la implementación de un marco legal en esta materia, partiendo del derecho a la intimidad personal, familiar y el derecho a conocer, actualizar y rectificar las informaciones que se hayan recogido en las bases de datos y en archivos de entidades públicas y privadas, que aparece consignado en el artículo 15 de la Constitución Política de 1991.

En términos de la confidencialidad, enmarcados en las directrices para el SEN, es importante tener en cuenta la Ley Estatutaria 1266 de 2008 que establece el Habeas Data y regula el manejo de la información contenida en bases de datos personales, financieras, crediticias, comerciales, de servicios y provenientes de terceros países, además de establecer disposiciones sobre la recolección, el tratamiento y la circulación de datos personales en el país (Congreso de Colombia, 2008).

Junto con estas leyes se cuenta con la Ley 1581 de 2012, que trata sobre la protección de datos personales, reglamentada parcialmente por el decreto 1377 de 2013, en el que se señala que el acceso a los datos se debe restringir y la información debe estar sujeta a tratamiento por parte del responsable, como lo indica en su artículo 4, manteniendo los principios de acceso y circulación restringida, de seguridad y de confidencialidad. Para el DANE, particularmente, la Ley 79 de 1993 establece la reserva estadística y la confidencialidad de las fuentes cuando se realizan procesos de recolección de información a través de censos o encuestas.

Respecto a la transparencia y el acceso de la información pública, la Ley 1712 de 2014, regula el derecho de acceso a la información pública, haciendo énfasis en el establecimiento de una política de datos abiertos por parte de las entidades públicas, y Decreto 2573 de 2014 del Ministerio de Tecnología de la información y las Comunicaciones (MinTIC), donde se establecen los lineamientos generales de la Estrategia de Gobierno en línea, indicando los principios y los fundamentos a tener en cuenta en las entidades públicas destacando entre ellos, la excelen-

cia en el servicio ciudadano, la apertura y la reutilización de datos públicos, la estandarización, la innovación, entre otros. Igualmente, se establecen cuatro componentes que facilitarán la masificación de la oferta y la demanda en gobierno en línea, resaltando entre estos, la seguridad y la privacidad de la información, siendo un componente transversal.

Estas normas han tenido su complemento con la reglamentación del SEN, a través de la Ley 1753 de 2015, del Decreto 1743 de 2016 y la Ley Estadística 2335 de 2023. En esta última, se establece el fortalecimiento y el aprovechamiento amplio e intensivo de los registros administrativos, así como el intercambio de información entre los integrantes del SEN, como fuente para la producción de estadísticas oficiales y el mejoramiento de la calidad y la coherencia en las cifras (Art. 11. Ley Estadística Nacional 2335 de 2023).

Finalmente, en 2017 con la actualización del Código Nacional de Buenas Prácticas del SEN, se planteó la implementación de prácticas en materia de acceso y confidencialidad de la información por parte del SEN, incentivando un mayor acceso y uso de la información de los productores de estadísticas en Colombia, así como la difusión de información anonimizada para garantizar la confidencialidad de la misma.

## Antecedentes de anonimización de datos no estructurados en el Centro Nacional de Memoria Histórica en Colombia

El Centro Nacional de Memoria Histórica de Colombia (CNMH), desarrolló la *Guía para la Anonimización de Datos e Información No Estructurada*[3], cuyo objetivo es brindar los lineamientos técnicos y la orientación metodológica para garantizar cualquier información producida, gestionada o recolectada por entidades públicas o privadas que contenga datos personales o información de identificación; se enmarca en las premisas de protección de derechos, transparencia y datos abiertos, acceso e interoperabilidad, eficiencia administrativa y reportes de información.

La CNMH utiliza como definición de dato no estructurado la aportada por el Consejo Nacional de Política Económica y Social en el CONPES 2018: *"información cuya organización y presentación no está guiada por ningún modelo o esquema. En esta categoría se incluye, por ejemplo, textos, audios, contenidos de redes sociales, etc"*.

En este contexto, las técnicas de anonimización varían dependiendo de la información que se utilice, textos, archivos de audio, archivos de video, formatos de imagen u otros formatos multimedia. A continuación, se relacionan las principales técnicas por tipo de información.

### Técnicas de anonimización para documentos físicos

Los documentos en físico deben cumplir con organización archivística y en términos de proceso de digitalización se debe tener en cuenta la conservación física, las condiciones ambientales, las condiciones operacionales, la seguridad de la información y la preservación en el largo plazo. En el caso de documentos originales que posean valor histórico, estos no podrán ser destruidos aun cuando hayan sido reproducidos o almacenados en cualquier medio. Debe controlarse el acceso a los documentos y la autorización de acceso debe ser normalizada de acuerdo con los lineamientos de la entidad.

Los documentos físicos pueden necesitar anonimizarse de forma parcial o total. En este sentido, para retirar los documentos de un expediente debe indicarse el lugar de almacenamiento por medio de un testigo (formato en papel) y en el caso de tratarse

[3] Disponible en https://centrodememoriahistorica.gov.co/wp-content/uploads/2022/08/GUIA-DE-ANONIMIZACION.pdf

de varios documentos, se puede incluir un listado al comienzo del expediente con los documentos que se retiran.

Para el proceso de anonimización se hace una copia del documento original y de la copia se retira la información que se requiera. Posteriormente se hace una copia al documento resultante y es esta la que se dará a conocer en forma de fotocopia o archivo digitalizado. Se sugiere marcar las copias 1 y 2 para identificarlas a partir de la original. El documento original debe conservarse garantizando su integridad.

### Técnicas de anonimización para documentos textuales en formatos digitales

Para su anonimización se recomienda usar una metodología análoga a la descrita anteriormente, es decir, se debe preservar el documento original y paralelamente guardar una copia que contenga el borrado de las partes restringidas que será publicada como copia anonimizada. Para este proceso se pueden utilizar archivos electrónicos como editores y programas de diseño gráfico.

### Técnicas de anonimización para archivos de tipo audiovisual

El lenguaje audiovisual contiene diversos tipos de información sensible de identificación, para ello, se pueden utilizar técnicas de desidentificación en contenido multimedia, cuyo objetivo es ocultar o eliminar identificadores personales, o sustituirlos por identificadores sustitutos en contenido multimedia para que se evite la divulgación y el uso de datos para fines no relacionados con el propósito para el que la información fue obtenida inicialmente.

- **Análisis de audio:** es el proceso de comprimir los datos y empaquetarlos en un formato de audio. Luego Audio Analytics realiza la extracción de significado e información de señales de audio para su análisis. El audio se puede presentar por medio de representación del sonido y archivos de sonido sin formato. Existen tres formatos de audio principales: formato de audio sin comprimir, formato de audio comprimido sin pérdida y formato de audio comprimido Lossy.

  El análisis de audio se puede aplicar a servicios de vigilancia, detección de amenazas, sistema de tele-monitoreo, sistema de red móvil, etc.

- **Análisis de video:** aunque los videos representan el 80% de los datos no estructurados cada día se generan y almacenan millones de pixeles de información, tanto en plataformas, como Youtube, como en cámaras de vigilancia y por personas independientes. Es así como en las dimensiones del análisis, el tamaño del video al ser mayor utiliza la red y el servidor en tiempo de procesamiento y las conexiones de bajo ancho de banda crean mayor tráfico en la red ya que los videos circulan lentamente.

  El análisis de video se puede aplicar en la identificación en accidentes de tránsito, en la policía de tráfico, en negocios, en seguridad, en análisis para inteligencia empresarial, en análisis de objetivos y escenarios, en direction analytics, en eliminar la ecuación humana a través de la automatización, etc.
  Los formatos de imagen son más difíciles de limpiar y la selección de técnicas de desidentificación se aplican según el tipo de información. Por ejemplo, varios archivos multimedia como los formatos de imágenes pueden ser accesibles para ciertas solicitudes en las que el algoritmo de los solicitantes busca metadatos de archivos de imágenes y produce resultados basados en texto para facilitar el borrado de imágenes sensibles.

Debido a las características particulares de la producción audiovisual y al gran número de marcadores de información personal que se puede encontrar en dicho material, el proceso de anonimización implica una actividad de entendimiento del proceso comunicativo que caracteriza el lenguaje natural. El proceso de desidentificación implica reconocer los insumos sonoros, visuales, espaciales, temporales, comportamentales, etc.

Actualmente no se han identificado herramientas informáticas que automáticamente realicen el proceso de anonimización total de los datos. Sin embargo, las nuevas herramientas de inteligencia artificial por medio de reconocimiento facial, el análisis de sonido, el análisis de texto, etc., han permitido un avance en términos de aplicación metodológica para este proceso.

## 1.4. Contexto internacional

La mayor parte de las experiencias internacionales sobre la anonimización de información estadística parten de las indicaciones o recomendaciones propias de los países, establecidas en sus sistemas nacionales de estadística. Este proceso se ha observado también como un posible mecanismo para dar seguimiento a la legislación internacional en materia de protección, privacidad y confidencialidad de la información.

En Reino Unido, por ejemplo, la Oficina del Comisionado de Información desarrolló el Código de Buenas Prácticas para la Anonimización (Oficina del Comisionado de información, 2012), en el cual se presentan los antecedentes jurídicos de la protección de datos de ese país y explica los beneficios de la anonimización y el por qué se debe hacer, todo enmarcado en los principios que deben guiar dicho proceso. También se señalan los riesgos que se pueden presentar al cruzar información en bases que ya se encuentran anonimizadas, generando posibles identificaciones de usuarios y dado que se debe realizar el control de los datos en el sistema estadístico en este país.

Otro esfuerzo internacional destacable en la documentación de procesos de anonimización para orientar a las entidades que producen estadística en los Sistemas Estadísticos Nacionales de los distintos países, es el realizado por la Comisión Económica para Europa de las Naciones Unidas (UNECE, por sus siglas en inglés), quien publicó el *Manual sobre el control de divulgación estadística*, en el que se proveen lineamientos técnicos para el control de la revelación de la información, se describen los métodos aplicables para la protección de la privacidad de la información y se explica detalladamente el programa de anonimización ARGUS, una iniciativa que viene liderando UNECE para implementar mecanismos de anonimización en los productores de estadística. ARGUS es un programa interactivo y libre, cuyas funcionalidades permiten identificar los datos de una base (metadatos), seleccionar y calcular las tablas de frecuencia, establecer la base de los métodos de anonimización y aplicar las diferentes técnicas a las variables relevantes. El programa es compatible con otros programas estadísticos (Morales, 2017:10).

Igualmente, la Red Internacional de Encuestas de Hogares (IHSN por sus siglas en inglés), en 2014 publicó una introducción sobre los controles a tener en cuenta en la divulgación estadística de información para uso de los integrantes del SEN y donde explicaba los métodos para difundir la información de datos confidenciales, teniendo en presentaba los distintos riesgos a los que se enfrentan los productores de información estadística, así como posibles mecanismos para valorar este tipo de riesgos y métodos de anonimización (IHSN, 2014).

Por otra parte, la anonimización permite atender de una mejor manera las demandas y las necesidades de información de distintos usuarios. Dentro de los sectores que se han caracterizado por generar este tipo de requerimientos, respecto a mayores desagregaciones de la información, especialmente en países como Estados Unidos, han sido investigadores, centros de investigación, universidades y el sector público.

En Estados Unidos se ha observado una fuerte relación entre la información y la investigación, siendo este último uno de principales focos para el desarrollo de procesos de anonimización. La Oficina de Censos ha liderado el estudio de técnicas de anonimización e integración de información, así como la opción de acceso a microdatos no anonimizados con propósitos académicos y de investigación. Algunas de las técnicas implementadas se basan, por ejemplo, en eliminar las variables de identificación directa, utilizar técnicas como umbrales geográficos o categóricos de las variables, re-

dondeo, infusión de ruido, recodificación, intercambio de registros basado en rangos o en proximidad (Morales, 2017: 9).

Dentro las bases anonimizadas publicadas por esta institución estadounidense se destacan las de censos demográficos y de encuestas económicas y, para el caso del Censo Agropecuario, se disponen al público algunas tablas agregadas por Estado y Condado. Todo esto bajo el marco de la normatividad nacional e internacional de confidencialidad de la información de los usuarios.

La Oficina de Estadísticas de Países Bajos (CBS, por sus siglas en holandés), ha trabajado conjuntamente con la UNECE (por sus siglas en inglés) para el avance de la investigación sobre procesos de anonimización de la información, destacándose de esta cooperación, la implementación de la iniciativa *Control de Divulgación Estadística* que es un estudio exhaustivo sobre la importancia de la confidencialidad y que busca, mediante la implementación de proyectos, generar mecanismos que permitan garantizarla.

## 1.5. Metodologías aplicadas en la anonimización de los censos económicos en la práctica internacional

### 1.5.1. Eurostat

EuroStat[4] estableció en 2008 una metodología para la anonimización de los datos de la Encuesta sobre la Estructura de los Ingresos, en ella se establecen un conjunto de criterios detallados y reglas generales aplicadas a la anonimización de los microdatos, además de dicha metodología, algunos países establecieron diversos parámetros para ampliar la anonimización específica de algunos datos individuales.

Los institutos nacionales de estadística de cada país donde se aplica la Encuesta sobre la Estructura de los Ingresos se encargan de seleccionar la muestra, preparar los cuestionarios, aplicar la encuesta, sintetizar los resultados y enviarlos a Eurostat, quien posteriormente se encargará de procesar los datos. En tanto la anonimización de los datos, Eurostat en 2008 desarrolló una metodología para la anonimización de los datos de la encuesta donde se estableció un conjunto de criterios detallados y reglas generales aplicadas a la anonimización de los microdatos. Además, de dicha metodología, algunos países adoptaron diversos parámetros para ampliar la anonimización específica de algunos datos individuales.

Las reglas generales aplicadas en la anonimización de los microdatos de la Encuesta sobre la Estructura de los Ingresos son:

- La recodificación de los cuasi-identificadores categóricos NACE, NUTS y SIZE de las empresas. La codificación resulta en mezcla de secciones, subsecciones y divisiones NACE y niveles NUTS 0 o 1 según el Estado miembro.
- Realizar una recodificación global en la variable edad para restringir sus valores a 6 intervalos (14-19, 20-29, 30-39, 40-49, 50-59, 60+).
- Eliminar la ciudadanía y el identificador de la empresa.
- Suprimir el factor de extrapolación para unidades locales.
- Brindar protección adicional a los empledos mediante la microagreación de clasificación individual sin restricciones para las variables métricas (días de ausencia e ingresos), con ello se busca ocultar Información relativa a las personas y estas variables, siempre y cuando sean grupos de al menos tres empleados.

[4] Disponible en https://ec.europa.eu/eurostat/web/microdata/structure-of-earnings-survey

### 1.5.2. Comisión Económica para América Latina y el Caribe (CEPAL)

La CEPAL realizó un estudio denominado "Control de divulgación estadística para las tablas censales del Caribe, una propuesta para ampliar la disponibilidad de datos censales desagregados"[5]. Este estudio revisa el problema del control de divulgación estadística para las tablas censales y las mejores prácticas internacionales, centrándose particularmente en el uso de métodos perturbativos; lleva a cabo análisis comparativos y pruebas del método de perturbación celular y métodos de redondeo aleatorio; recomienda que estos métodos se pongan a disposición de las oficinas de estadística a través del software Recuperación de Datos para Áreas pequeñas por Microcomputador (REDATAM).

#### Práctica actual en el Caribe

Al producir cuadros censales como los publicados en los informes de censos nacionales, las oficinas de estadística del Caribe limitan o evitan la publicación de recuentos pequeños que conduzcan a la divulgación, restringen la publicación de tablas para áreas geográficas y grupos pequeños o se recodifican las variables. Este tipo de supresión o rediseño de tablas se considera una forma de control de divulgación estadística, pero si se usa solo no logra un equilibrio eficiente entre el riesgo de divulgación y la utilidad de los datos. Cuando se utiliza REDATAM para difundir los datos del censo, el riesgo de divulgación se gestiona a través del diseño de la aplicación en línea. Los aspectos relevantes del diseño incluyen: las variables que están disponibles en la aplicación, la forma en que se codifican esas variables, la medida en que los usuarios pueden tabular diferentes variables entre sí (o usar variables para filtrar las consultas de la base de datos), el nivel de desagregación geográfica que se encuentra disponible, y si se le brinda al usuario la funcionalidad de utilizar el lenguaje de programación de REDATAM para especificar consultas a la base de datos.

#### Métodos pre-tabulares

El método de perturbación pre-tabular más utilizado es el intercambio de registros que se ha utilizado en censos en Estados Unidos (censos de 1990, 2000 y 2010), Reino Unido (censos de 2001 y 2011), Suecia (2011), Austria (2011) y Bélgica (2011). La esencia del método de intercambio de registros (o intercambio de datos) es que se intercambia una pequeña proporción de hogares que son similares en términos de algunas características sociales y demográficas básicas, es decir, sus códigos geográficos se intercambian como si los hogares intercambiaran físicamente lugares. La rutina de intercambio se puede refinar para favorecer a los hogares con características raras que corren un mayor riesgo de identificación o divulgación de atributos (esto se conoce como intercambio de registros específicos). La distancia geográfica entre los hogares intercambiados también puede variar según el riesgo de divulgación, dentro de un cierto límite debido a que un hogar no se intercambiaría con otro en una parte completamente diferente del condado.

#### Métodos post-tabulares

La alternativa a la perturbación de los microdatos del censo es aplicar la perturbación después de la tabulación, es decir, perturbar los datos en las tablas finales del censo. De esta forma, el control de la divulgación se convierte efectivamente en un "complemento" del proceso de tabulación. Hay tres enfoques post-tabulares principales que han empleado las oficinas de estadística: ajustes de celdas pequeñas; redondeo aleatorio, y el método de perturbación de celdas de la Oficina Australiana de Estadísticas (ABS).

[5] Disponible en https://repositorio.cepal.org/bitstream/handle/11362/46628/4/S2000895_en.pdf

### 1.5.3. Australia

La Oficina de Estadísticas de Australia (ABS) se compromete a garantizar que se implementen medidas de seguridad para proteger la información sensible recopilada del censo. La seguridad de la información incluye la conversión de nombres a números anónimos y solo se permite que una pequeña cantidad de funcionarios ABS usen estos códigos, mientras se aplican estrictos protocolos de seguridad[6]. La ABS contrató a expertos en criptografía de la Universidad de Melbourne para investigar diferentes métodos de codificación de nombres para su uso en proyectos de integración de datos, haciendo énfasis en el Censo de 2016 para mejorar el valor de los datos del censo, combinándolos con datos de diferentes fuentes (p.ej., encuestas o registros administrativos), obteniendo mejor comprensión del panorama social y económico para ayudar en la toma de decisiones gubernamentales en áreas como salud, educación, infraestructura y economía.

### 1.5.4. México

El Instituto Nacional de Estadística y Geografía (INEGI) realiza desde 1930 censos económicos y el último censo económico se llevó a cabo en 2019 y su objetivo era recopilar información estadística básica (correspondiente a 2018) de todos los establecimientos productores de bienes y servicios, con el fin de producir indicadores económicos del país a gran nivel de detalle geográfico, sectorial y temático. La información recopilada durante los censos económicos realizados en 2019 se sometió al sistema de confidencialidad llamado "supresión parcial con datos complementarios"[7] con el fin de garantizar la confidencialidad de los datos suministrados por los informantes, cumpliendo con lo dispuesto en los artículos 37 y 38 de la Ley del Sistema Nacional de Información Estadística y Geográfica; este sistema consiste en suprimir los datos de las variables económicas de los renglones con problemas de confidencialidad, donde permanecen el número de unidades económicas y para compensar se agregan indicadores relacionados con las variables que se suprimieron, por lo tanto, se muestra la totalidad de renglones existentes de cada cuadro (tengan o no problemas de confidencialidad).

### 1.5.5. Estados Unidos

La Oficina de Censo de Estados Unidos salvaguarda la información con los controles de divulgación estadística, disfrazando los datos originales con métodos de intercambio de datos en el censo económico para establecer estos símbolos válidos y tener un control de información. Como parte de las buenas prácticas esta oficina ha implementado desde 1970 métodos de proyecciones de privacidad del censo, lo que ha mejorado los procesos con el avance de la tecnología.

### 1.5.6. Técnicas de anonimización aplicadas a censos – revisión de literatura

Dick et al (2023) *Confidence-ranked reconstruction of census microdata from published statistics*[8] (Reconstrucción clasificada por confianza de microdatos censales a partir de estadísticas publicadas): consiste en procesos de reconstrucción de datos a partir de información contenida en estadísticas agregadas publicadas para reconstruir filas completas de un conjunto privado de datos por medio de la aplicación de técnicas aleatorias de optimización no convexa. Esta técnica se aplicó a datos del censo de Estados Unidos de 2010 y en conjunto de datos de la Encuesta de la Comunidad Americana.

Fioretto et al (2021) *Differential privacy of hierarchical Census data. An optimization approach* (Privacidad diferencial de los datos censales jerárquicos. Un enfoque de optimización): implica el desarrollo de un mecanismo que optimiza el ruido introducido

[6] Disponible en https://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/844269E83C4B6666CA25823C00178BBB/$File/1351055162_2018.pdf
[7] https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825196530.pdf
[8] Disponible en: https://www.pnas.org/doi/10.1073/pnas.2218605120

para asegurar la privacidad diferencial que garantiza la coherencia entre diferentes niveles jerárquicos (nacional, estatal, condado). El mecanismo se basa en programación dinámica que aprovecha la naturaleza jerárquica del problema y la estructura de la función objetivo.

Parra-Arnau, et al (2020) *Differentially private data publishing via cross-moment microaggregation* (Publicación de datos diferencialmente privados mediante microagregación entre momentos): la microagregación de momentos cruzados reemplaza los registros originales en un grupo por estadísticas adicionales lo que reduce la sensibilidad de los datos y aporta mayor privacidad y utilidad que la microagregación normal.

## 1.6. Implementación de instrumentos, software, aplicativos o sistemas que permitan la anonimización de datos

En esta sección se presentan las experiencias de los referentes internacionales frente al uso o la aplicación de herramientas, instrumentos, softwares, aplicativos o sistemas para realizar procesos de anonimización, sus aplicaciones, actividades desarrolladas e implicaciones de uso.

**Tabla 1.** Experiencias internacionales en el uso de software, aplicativos o sistemas para realizar procesos de anonimización

| Referente | Software, aplicativo o herramienta |
|---|---|
| **EuroStat** | Cuenta con un software que permite realizar anonimización, mediante el manual de estadística empresarial europeo, hace énfasis en softwares como T- ARGUS y Sdc-Table que permiten la anonimización de información o el control de divulgación estadística. Son herramientas que realizan sus procedimientos sobre dos tipos de salidas estadísticas, como los datos confidenciales presentados en tablas y datos confidenciales en archivos de microdatos. |
| **CEPAL** | La Cepal desarrolló el software REDATAM que es un sistema computacional de carácter social e interactivo que facilita el procesamiento, el análisis y la diseminación web de la información de censos, encuestas, registros administrativos, indicadores nacionales/regionales y otras fuentes de datos. Asimismo, es una herramienta que permite el procesamiento en línea de la información censal y estadística producida por las instituciones gubernamentales, divulgando dicha información de manera segura y sin ningún costo ya que el usuario interactúa a través de páginas predefinidas seleccionando tabulados e indicadores para luego enviar una solicitud al servidor y posteriormente este devuelve el resultado en forma de tabulado, gráfico o mapa estático; la solicitud del usuario puede tomar formatos como frecuencias, cruces de variables, promedios, conteos o listas por área geográfica. |

| Referente | Software, aplicativo o herramienta |
|---|---|
| **España** | El ayuntamiento de Barcelona, anonimizó los datos personales que tienen a su disposición para proteger la información sensible de los ciudadanos con un software llamado Nymiz. Este es un software que detecta los datos personales en archivos no estructurados y también en datos estructurados, y anonimiza o seudonimiza de forma reversible o irreversible los datos de acuerdo con las necesidades del tratamiento de la información. Es un programa de pago y su costo depende del volumen de datos a procesar. |
| **Canadá** | Statistics Canada desarrolló el software de control de divulgación automatizado G- Confind con el fin de proporcionar el nivel adecuado de protección para celdas confidenciales y minimizar la pérdida de información. G-Confind es un sistema generalizado que evita la divulgación de información confidencial en datos tabulares por medio del método de supresión de celdas, en el que se identifican y suprimen las celdas sensibles y complementarias a proteger, además, puede usarse para auditar patrones de supresión de celdas y encontrar agregados confidenciales y datos tabulares redondos. |
| **Países Bajos** | Para el control de divulgación estadística Estadísticas de Países Bajos fue pionero en el desarrollo del software µ-ARGUS, el cual es un programa interactivo flexible que guía a los usuarios en el proceso de protección de los datos y que está basado en el enfoque de gestión de riesgos. Este software incorpora algunos métodos sofisticados tradicionales para la anonimización de microdatos como p. ej. métodos de enmascaramiento perturbativo y métodos de enmascaramiento no perturbativo y adicionalmente incorpora otros como micro agregación, PRAM, redondeo, codificación superior e inferior, intercambio de rango y adición de ruido. |
| **Estados Unidos** | Desde el Instituto Nacional de Estándares y Tecnología tienen el programa de ingeniería de privacidad y en él se describen 15 herramientas de desidentificación y el software ARX. Estas técnicas son aplicadas a un conjunto de datos para prevenir o limitar los riesgos de la privacidad de los individuos, los grupos protegidos y los establecimientos; estas técnicas pueden ser introducción de ruido como la privacidad diferencial, el enmascaramiento de datos y la creación de conjuntos de datos sintéticos que se basan en modelos que preservan la privacidad. |

Fuente: DANE, a partir de información de referentes internacionales.

# 2. Objetivo y alcance de la guía

## Alcance

La *Guía de anonimización de datos estructurados* tiene como objetivo orientar a las entidades integrantes del SEN sobre la planeación y la ejecución del proceso de anonimización de datos estructurados, que permitan el acceso y el aprovechamiento de los datos teniendo en cuenta el índice de información clasificada y reservada de la Ley de transparencia[9], controlando el riesgo de identificación de la fuente; no obstante, la guía no contiene lineamientos obligatorios para la aplicación del proceso de anonimización.

De esta manera, esta guía puede ser consultada por:

- Instituciones generadoras de datos estructurados.
- Instituciones y personas que usan datos estructurados.
- Entidades del SEN y personas que deseen salvaguardar la información privada de sus unidades de observación.

## Objetivos específicos

- Establecer los conceptos necesarios para comprender los procesos de anonimización de datos estructurados.
- Definir la estrategia de planeación que permita la implementación del proceso de anonimización.
- Presentar los aspectos metodológicos que debe considerar el usuario en la implementación de técnicas de anonimización.
- Orientar a las entidades del SEN sobre las técnicas de anonimización aplicables de acuerdo con las características de los conjuntos de datos que requieran el proceso de anonimización en el marco de la normatividad de protección de datos personales.

Tanto la conceptualización metodológica de la guía, sus lineamientos como su aplicación están orientados a la anonimización de datos estructurados y está dirigida a las entidades del SEN cuya misión es gestionar, almacenar, administrar, procesar, producir, custodiar y publicar información; el proceso consta de seis etapas: revisiones previas; análisis de riesgos; selección de técnicas; análisis de viabilidad; aplicación de técnicas, y evaluación de resultados.

[9] Disponible en: https://www.archivogeneral.gov.co/transparencia/datos-abiertos/indice-informacion-clasificada-reservada#:~:text=El%20Índice%20de%20Información%20Clasificada,calificada%20como%20clasificada%20o%20reservada.

# 3. Marco conceptual para la anonimización

Se define la **anonimización de microdatos** como un proceso técnico que consiste en:

> "(...) transformar los datos individuales de las unidades de observación, de tal modo que no sea posible identificar sujetos o características individuales de la fuente de información, preservando así las propiedades estadísticas en los resultados" (Decreto 1743 de 2016: Art. 2.2.3.1.1).

La finalidad de la anonimización es impedir que, a partir de una información o de una combinación de informaciones, se logren identificar sujetos individuales en un archivo de microdatos (Morales, 2017: 5).

El concepto de **microdato** es fundamental en la concepción del proceso de anonimización, pues corresponde a:

> "(...) los datos sobre las características asociadas a las unidades de observación que se encuentran consolidadas en una base de datos" (Artículo 5. Ley Estadística 2335 de 2023).

A su vez, el Sistema de consulta de conceptos estandarizados del DANE define que una **base de datos** hace referencia a:

> "... un conjunto o colección de datos interrelacionados entre sí, que se utilizan para la obtención de información de acuerdo con el contexto de los mismos y que son almacenados sistemáticamente para su posterior uso" (Sistema de consultas de conceptos estandarizados, 2018).

Según lo anterior, el proceso de anonimización se aplica a aquellos datos que por su naturaleza son sensibles al público debido a la posibilidad de que sea violada su confidencialidad.

**El proceso de anonimización** puede aplicarse tanto a los microdatos obtenidos de **operaciones estadísticas como a los de los registros administrativos** que posee una entidad.

Respecto al primer tipo de datos, el Decreto 1743 de 2016 ha definido operación estadística como la "aplicación de un proceso estadístico sobre un objeto de estudio que conduce a la producción de información estadística" (Decreto 1743 de 2016: Art. 2.2.3.1.1).

El **proceso estadístico** se entiende como un:

> "Conjunto sistemático de actividades encaminadas a la producción de estadísticas, entre las cuales están comprendidas: la detección de necesidades de información, el diseño, la construcción, la recolección; el procesamiento, el análisis, la difusión y la evaluación" (Artículo 5. Ley Estadística 2335 de 2023).

Por tanto, el proceso de anonimización que se aplique a las operaciones estadísticas debe contar con una metodología y una documentación a lo largo de las fases de producción para garantizar la calidad de la información estadística a generar.

La información que se obtiene del proceso estadístico a nivel de los datos de las unidades de observación (hogares, personas, empresas) sirve como insumo para que los usuarios puedan aprovechar estadísticamente la información de dichas operaciones, así como para generar estudios o investigaciones propias que permitan mejorar la toma de decisiones, además de otros usos que consideren pertinentes los responsables de la información a anonimizar.

Es así como la anonimización fomenta en las entidades del SEN la transparencia de la información y priorizar, en este caso, la preservación de la confidencialidad.

Frente a los **registros administrativos,** estos se entienden como un:

> "Conjunto de datos que contiene la información recogida y conservada por entidades y organizaciones en el cumplimiento de sus funciones o competencias misionales u objetos sociales. De igual forma, se consideran registros administrativos las bases de datos con identificadores únicos asociados a números de identificación personal números de identificación tributaria u otros, los datos geográficos que permitan identificar o ubicar espacialmente los datos, así como los listados de unidades y transacciones administrados por los integrantes del SEN" (Artículo 5. Ley Estadística 2335 de 2023).

Para esta fuente en particular, las entidades del SEN pueden contar con procesos de diagnósticos que permiten implementar prácticas para mejorar la captura de la información, mediante la implementación de indicadores de cali-

dad sobre el registro de la entidad[10]. Los resultados de los diagnósticos de los registros administrativos permitirán al mismo tiempo, desarrollar mejores procesos de anonimización y obtener mejoras en la publicación de su información estadística.

Los **datos estructurados** se entienden como:

> "(...) datos que tienen un modelo de datos y formato predefinido y que se ajustan a una forma de tablas, de registros o filas con campos de significados fijos y relaciones o enlaces entre las tablas" (Departamento Administrativo Nacional de Estadística (DANE), Acuerdo equipo de trabajo DIRPEN).

Los datos estructurados siguen una estructura predeterminada, por ejemplo: bases de datos SQL, archivos de Excel, resultados de formularios web, registros administrativos, etc.

Los **datos no estructurados** se entienden como:

> "(...) datos que no tienen un modelo de datos predefinido y no están organizados de manera predefinida o su estructura no se ajusta perfectamente a una tabla de datos relacional" (Departamento Administrativo Nacional de Estadística (DANE).

Los datos no estructurados pueden presentarse en formato de texto, imagen, sonido, video u otros formatos, por ejemplo: correos electrónicos, archivos PDF, publicaciones en redes sociales, imágenes digitales, archivos de audio, archivos de video.

A continuación, se relacionan una serie de conceptos que permitirán al usuario un mayor entendimiento de la información:

**Datos abiertos** se entiende como:

> "(...) aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso

[10] Para mayor información sobre los procesos de diagnóstico de los registros administrativos, se puede consultar la Metodología de diagnóstico de los Registros administrativos para su Aprovechamiento estadístico del DANE. Disponible en: http://www.dane.gov.co/files/sen/registros-administrativos/Metodologia-de-Diagnostico.pdf

y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos" (Ley 1712 de 2014: Ley de Transparencia y del derecho de acceso a la información pública nacional).

**Dato personal** se define como:

"Cualquier información vinculada o que pueda asociarse a una o varias personas naturales determinadas o determinables" (Ley 1581 de 2012, Congreso de Colombia).

**Dato público** se define como:

"(...) aquel que puede ser consultado por cualquier persona de manera directa, sin el consentimiento del titular. Ejemplo: número de identificación, apellidos, lugar y fecha de expedición del documento" (Registraduría Nacional del Estado Civil).

**Dato semiprivado** es:

"(...) aquel que además de ser de interés para el titular, puede ser de interés para cierto sector o grupo de personas. Ejemplo: dirección de residencia y teléfono" (Registraduría Nacional del Estado Civil).

Se entiende por **dato sensible:**

"(...) a aquellos que afectan la intimidad del titular o cuyo uso indebido puede generar su discriminación, tales como aquellos que revelen el origen racial o étnico, la orientación política, las convicciones religiosas o filosóficas, la pertenencia a sindicatos, organizaciones sociales, de derechos humanos o que promueva intereses de cualquier partido político o que garanticen los derechos y garantías de partidos políticos de oposición así como los datos relativos a la salud, a la vida sexual y los datos biométricos" (Ley 1581 de 2012, Congreso de Colombia).

Se entiende **operación estadística** como:

> "(...) la aplicación del conjunto de procesos y actividades que comprende la identificación de necesidades, diseño, construcción, recolección o acopio, procesamiento, análisis, difusión y evaluación, la cual conduce a la producción de información estadística sobre un tema de interés nacional o territorial" (Artículo 5. Ley Estadística 2335 de 2023).

# 4. Planeación del proceso de anonimización

Este capítulo tiene la finalidad de orientar sobre como delimitar el alcance de la anonimización que se pretende realizar, así como, establecer los recursos a utilizar en el proceso. Por ello se estableció la revisión normativa sobre protección de datos e identificación de necesidades de información y que implicó una exhaustiva revisión de restricciones normativas que puedan afectar la publicación de datos a anonimizar, considerando leyes, acuerdos de confidencialidad y normativas pertinentes.

Posteriormente, se realizó una revisión de las demandas de información de los usuarios, analizando solicitudes, derechos de petición y otras necesidades. Esta revisión permite clasificar y cuantificar las demandas e identificar las variables y los periodos más solicitados. Con base en estos resultados, el equipo de trabajo ejecuta el proceso de anonimización, tomando en cuenta requisitos previos como la disponibilidad de la base de datos, un diccionario de datos, una infraestructura tecnológica adecuada y mecanismos de seguridad para proteger la información.

## 4.1. Revisión normativa sobre protección de datos e identificación de necesidades de información

Este subproceso busca realizar una revisión normativa que pueda afectar la publicación de la información sujeta a anonimizar; asimismo, se sugiere identificar las necesidades de información que presentan los usuarios sobre la base de datos. El subproceso se compone de dos pasos:

1. Revisión de restricciones de publicación de la información.
2. Revisión de las necesidades de los usuarios de la información.

### 4.1.1. Revisión de restricciones de publicación de la información

El equipo de trabajo debe realizar una revisión de la normatividad que puede afectar la publicación de la información sujeta a ser anonimizada. En este caso es importante que la entidad verifique las normas y las cláusulas de confidencialidad que tengan

alcance en la información estadística que se espera dejar disponible para el SEN.

Para iniciar esta actividad, el equipo de trabajo tendrá que realizar la revisión de leyes, decretos, resoluciones, convenios institucionales, acuerdos de confidencialidad de la información, estatutos y toda la normatividad que fundamenta el origen del registro administrativo o de la operación estadística de la entidad. El resultado de este análisis determinará si existe alguna norma que impida la publicación de la información de la base de datos a anonimizar.

Un insumo a tener en cuenta en esta actividad para el equipo de trabajo es el alcance de las normas asociadas a la confidencialidad de la información que tiene impacto sobre las entidades del SEN. Por ejemplo, el principio de confidencialidad descrito en la Ley Estatutaria 1266 de 2008 o el artículo 2.2.3.3.5 del Decreto 1743 de 2016, donde se establecen indicaciones sobre la no exposición de la identificación y la ubicación de las unidades de observación al momento de publicar información (Recuadro 1).

**Recuadro 1.** Normas referentes a la confidencialidad de la información

**Ley Estatutaria 1266 de 2008**: el principio de la confidencialidad descrito en la ley indica que "Todas las personas naturales o jurídicas que intervengan en la administración de datos personales que no tengan la naturaleza de públicos están obligadas en todo tiempo a garantizar la reserva de la información".

**El artículo 2.2.3.3.5 del decreto 1743 de 2016** establece que "las entidades que conforman el SEN ..., deberán guardar la confidencialidad de los datos que permitan la identificación y/o localización espacial de las fuentes, cuando estos fueren recolectados exclusivamente para la producción de las estadísticas oficiales y para fines estadísticos".

**Ley Estadística 2335 de 2023, Capítulo V** establece que quienes producen estadísticas oficiales deberán proteger los datos confidenciales de tal forma que, en la publicación de resultados estadísticos, la unidad estadística no pueda ser identificada, directa ni indirectamente, además, deberán tomar todas las medidas regulatorias, administrativas, técnicas y organizativas necesarias para evitar el acceso de personas no autorizadas a los datos individuales.

**Circular 13 del 29 de abril de 2020, Oficina Asesora Jurídica del DANE - FONDANE,** "CRITERIOS JURÍDICOS DE CLASIFICACIÓN DE LA INFORMACIÓN DEL DANE - FONDANE" que consolida de manera clara y precisa cómo se debe clasificar la información teniendo como presupuesto lo dispuesto en la normatividad vigente respecto a información pública, información reservada y clasificada y protección de datos personales, así como los lineamientos de la Guía No. 5 para la Gestión y la Clasificación de Activos de Información de Seguridad y Privacidad de la Información del Ministerio de Tecnologías de la Información y Telecomunicaciones en Colombia.

Después de la revisión normativa y de tener en cuenta los principios de confidencialidad, el equipo de trabajo describirá los hallazgos de la correspondiente revisión. Esta descripción es necesario incluirla en el informe final de anonimización y puede relacionar aspectos como:

- **Normatividad identificada** que impida la publicación de la información, como decretos, leyes, artículos, entre otros.
- **Normas que respalden la publicación de la información** a los diferentes usuarios y demás entidades del SEN.
- **Observaciones y sugerencias** a tener en cuenta al momento de realizar la publicación de la información, para evitar sanciones legales y judiciales a la entidad del SEN.
- **Personal y fecha en la que se hizo la revisión de la normatividad** para presentar una trazabilidad del trabajo sobre la base de datos a anonimizar.

### 4.1.2. Revisión de las necesidades de los usuarios de la información

Este paso tiene como propósito realizar una caracterización de las necesidades de información de los usuarios sobre la base de datos a anonimizar. En este caso, el equipo de trabajo de la entidad del SEN debe tener en cuenta las demandas o los requerimientos realizados por parte de los usuarios.

Para el caso en que la entidad del SEN cuente con demandas de información por distintos usuarios, se recomienda elaborar un listado que contenga los requerimientos solicitados, las fechas de solicitud, las variables requeridas, la frecuencia con la que se hacen las solicitudes y las respuestas dadas a dichos requerimientos.

Se recomienda al equipo de trabajo realizar una revisión del histórico de las solicitudes, mayor a un año y menor a dos años, teniendo en cuenta el volumen de las solicitudes recibidas, la recurrencia y la frecuencia que tiene cada solicitud. Algunas de las solicitudes que podrá revisar el equipo de trabajo se presentan a continuación:

- Solicitudes de información recibidas por distintos usuarios: ciudadanos, academia, empresas, entidades públicas del orden nacional o territorial, organismos internacionales, entre otros.
- Derechos de petición radicados en la entidad durante el último año.
- Necesidades de información propias del área temática de la entidad.

Con la información recolectada, el equipo de trabajo podrá clasificar y cuantificar las demandas de información, teniendo en cuenta:

- Tipo de usuarios.
- Tipo de solicitudes.
- Variables solicitadas.
- Nivel de desagregación requerido de la información.
- Periodos de la información (anual, semestral, trimestral, etc.).
- Frecuencia de las solicitudes
- Objetivo, finalidad, uso de la información requerida.

El Recuadro 2 presenta un ejemplo de clasificación de este tipo de solicitudes de información.

**Recuadro 2.** Ejemplo de clasificación de solicitudes recibidas.

| Referente | Descripción | N° de solicitudes |
|---|---|---|
| **Tipo de usuario** | Entidades públicas | 4 |
| | Entidades privadas | 3 |
| | Investigadores | 20 |
| | Instituciones académicas | 15 |
| | Otros | 8 |
| **Tipo de solicitudes** | Derechos de petición | 1 |
| | Acceso a información | 30 |
| | Actualización de datos | 19 |
| **Clasificación de variables** | Variables de clasificación | 15 |
| | Variables de ubicación | 10 |
| | Variables temáticas | 25 |
| **Niveles de desagregación** | Nacional | 3 |
| | Departamental | 10 |
| | Municipal | 30 |
| | Temática | 5 |
| | Otros | 2 |
| **Periodos de información solicitados** | Anual | 16 |
| | Mensual | 30 |
| | Trimestral | 20 |
| | Diaria | 23 |

| Referente | Descripción | N° de solicitudes |
|---|---|---|
| **Frecuencia de la solicitud** | Mensual | 120 |
| | Diaria | 26 |
| **Objetivo, finalidad, uso** | Investigaciones o estudios | 36 |
| | Académica | 12 |
| | Política pública | 59 |

Fuente: DANE/DIRPEN.

Basados en la clasificación y la cuantificación de las solicitudes, el equipo de trabajo podrá identificar las variables, los niveles de desagregación, los periodos de la información de la base de datos que tienen mayor demanda lo que le dará información para definir la base de datos y desagregaciones que satisficiera de la mejor forma las necesidades de los usuarios y que igualmente no expongan la identificación de las unidades de observación.

Una ventaja de publicar las bases anonimizadas, para la entidad del SEN, es la disminución de las cargas operativas para responder solicitudes de información que pueden ser repetitivas o recurrentes. Teniendo en cuenta la revisión de los diferentes recursos para encontrar un uso potencial, el equipo de trabajo de la entidad del SEN tomará la decisión del tipo de información que puede suministrar en la base de datos anonimizada para responder a las necesidades de los usuarios.

Después de la revisión de las necesidades de los usuarios de la información, el equipo de trabajo, mediante una bitácora, indicará los resultados encontrados:

- Periodo de revisión del equipo de trabajo de las solicitudes, las peticiones y los requerimientos de información.
- Listado de variables más solicitadas y los niveles de desagregación más demandados.
- Listado de los usuarios que con mayor frecuencia hacen solicitudes de información.
- Los periodos o las bases de datos con determinados cortes que más son solicitados.
- Listado de los diferentes usos identificados para los cuales son radicadas la mayoría de las solicitudes y los requerimientos.

El proceso de anonimización debe ser ejecutado por un equipo de trabajo que tenga acceso y conocimiento de la base de datos a anonimizar. Es importante que este equipo de trabajo cuente con la capacidad de:

- **Conocer temáticamente el contenido de la base de datos a anonimizar:** por ejemplo, si la base es insumo de una operación estadística de un área temática particular es importante que el equipo de trabajo se encuentre conformado por una persona o varias que conozcan el fenómeno registrado en la base de datos. La vinculación de este personal especializado tiene como fin apoyar la toma de decisiones de la anonimización a desarrollar, principalmente en las etapas de riesgos de identificación (Etapa II) y de análisis de viabilidad del proceso (Etapa IV).

- **Manejar herramientas que permitan el análisis exploratorio de datos:** en este caso se necesitará que el equipo cuente con miembros con habilidades para el manejo de paquetes estadísticos como *Python, R, SAS, SPSS, Stata*, entre otros. Igualmente, se recomienda que conozcan técnicas estadísticas que permitan apoyar la etapa de análisis exploratorio de la base de datos (subproceso de la Etapa I) y la etapa de aplicación de las técnicas de anonimización (Etapa V).

Cuando la entidad conforme estratégicamente su equipo de trabajo, este deberá tener en cuenta que existen requerimientos iniciales que permiten la ejecución satisfactoria del proceso de anonimización. Estos son:

- **Disponer de una base de datos:** la base definida por la entidad y que será difundida para el SEN. Esta puede contener una tabla o varias tablas relacionadas entre sí[11]. La entidad debe saber si la base de datos es resultado de una operación estadística o es la resultante de un registro administrativo.

- **Contar con el diccionario de datos de la base a anonimizar[12]:** este documento debe especificar claramente las propiedades básicas de las variables contenidas en la base de datos. Algunas de estas propiedades son: nombre, longitud, obligatoriedad de respuesta, descripción, reglas de validación, entre otros, así como la relación entre ellas.

- **Disponer de una infraestructura tecnológica:** la infraestructura definida por la entidad debe permitir el manejo estadístico de datos; en este caso, debe tener en cuenta paquetes estadísticos, equipos de cómputo que permitan el manejo de datos y en general tecnología que se encuentre acorde con el volumen de información a anonimizar. Cuando las bases de datos son de grandes dimensiones, el equipo de trabajo debe tener en cuenta que algunos paquetes de software no son compatibles, por lo cual requerirá revisiones sobre programas estadísticos y el alcance de estos sobre grandes cantidades de información.

- **Definir mecanismos de seguridad sobre la base de datos a anonimizar:** la entidad debe prever condiciones mínimas para salvaguardar la información, así como definir protocolos de acceso a la información por parte del equipo de trabajo que participará en el proceso de anonimización. Por ejemplo, acuerdos de confidencialidad del personal involucrado en el proceso (equipo de trabajo) debidamente firmados, usos de permisos y contraseñas para el uso de la información, entre otros.

[11] Modelo Entidad-Relación de la base de datos.
[12] Diccionario ejemplo dispuesto, disponible en: https://www.sen.gov.co/files/sen/lineamientos/Guía_Diccionario_de_Datos.xlsx.

**Recuadro 3.** Verificación especial de la base de datos de acuerdo a su origen

La entidad del SEN debe verificar si la base de datos a anonimizar es la resultante de alguna operación estadística o de un registro administrativo.

**Caso 1: Resultado de operaciones estadísticas**

La entidad tendrá que verificar que la base de datos para la anonimización es la resultante de la finalización de la Fase de Ejecución del Proceso Estadístico. Esto significa que la consistencia de los datos recolectados debe haber sido validada, según: i) las técnicas definidas por la entidad responsable; ii) las variables creadas (o calculadas) a partir de la información recolectada, y iii) que se encuentren en la base de datos. En caso de valores faltantes, tendrá que verificar que los métodos de imputación hayan sido aplicados.

**Caso 2: Base de datos de registros administrativos**

La entidad debe verificar que la base de datos sea la versión más actual. En este caso, se recomienda que el periodo de tiempo del registro administrativo que se anonimizará se defina con base en la periodicidad de consolidación de la información. Además, se recomienda revisar la consistencia y la calidad de la base de datos teniendo en cuenta la sección Revisión de la consistencia de la base de datos de la Metodología de diagnóstico de registros administrativos.
Ejemplo:
El Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales (SISBEN) tiene una periodicidad de recolección diaria y de consolidación de la información en el aplicativo SISBENNET mensual. En este caso, se recomendaría que SISBEN realice un corte para anonimizar la base de datos de un periodo, teniendo como fecha de cierre la última consolidación de la información.

Fuente: elaboración propia.

Después de que el equipo de trabajo confirme la disponibilidad de los requerimientos previos, empezará con la primera etapa del proceso de anonimización.

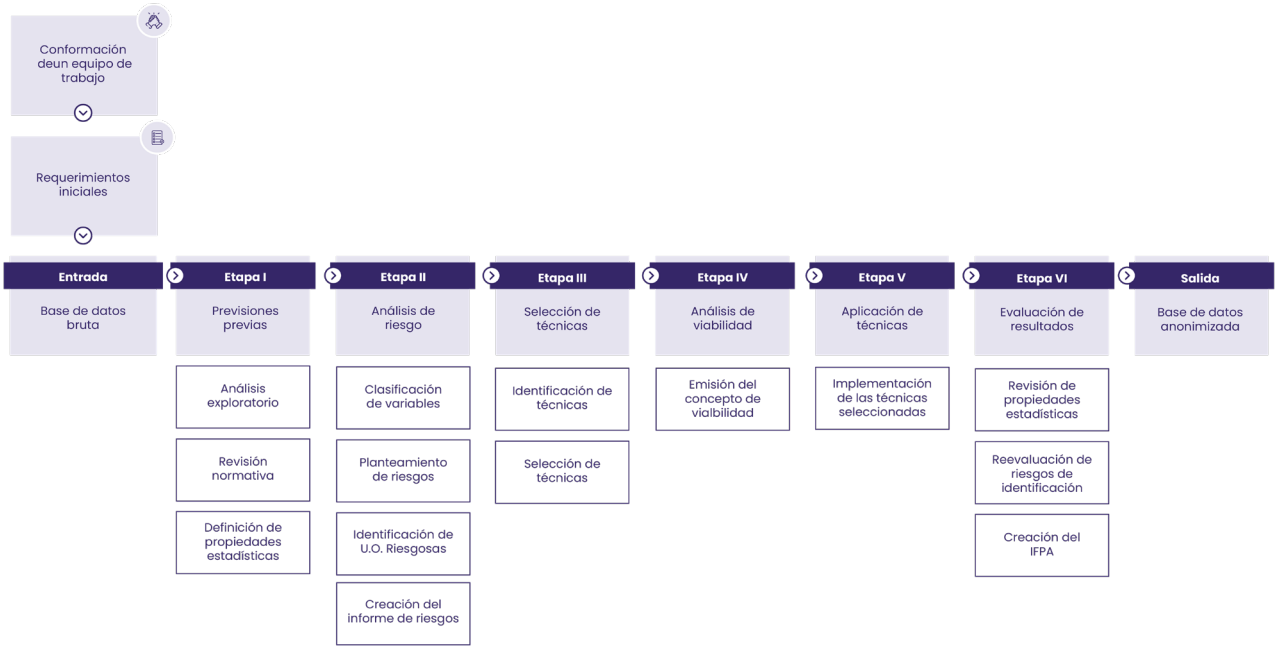
**Insumos:**

- Equipo de trabajo definido.
- Base de datos.
- Dirección de datos completo
- Infraestructura definida

# 5. Ejecución del proceso de anonimización

Este capítulo contiene las etapas en que se desarrollan el proceso de anonimización en datos estructurados. El proceso de anonimización de una base de datos se encuentra compuesto por las siguientes etapas: i) Revisiones previas; ii) Análisis de riesgos de identificación de las fuentes de información; iii) Identificación y selección de técnicas de anonimización; iv) Análisis de viabilidad del proceso; v) Aplicación de técnicas de anonimización, y vi) Evaluación de resultados del proceso. Estas fases se presentan en la Ilustración 1:

**Ilustración 1.** Esquema general del proceso de anonimización




Fuente: DANE- DIRPEN.

Estas etapas presentan distintos subprocesos y actividades a desarrollar para realizar la anonimización de la base de datos deseada.

Debido a que es un proceso de anonimización, los contenidos de las fases se encuentran orientados a presentar entradas y salidas, así como recomendaciones sobre pasos previos que permitirán preparar a la entidad al iniciar el proceso de anonimización.

Igualmente, a lo largo de las etapas se realizan indicaciones y recomendaciones de documentación que permitan conocer la trazabilidad del proceso de anonimización dentro de la entidad. Al finalizar la etapa 6 del proceso, podrá encontrarse un modelo de Informe Final del Proceso de Anonimización (IFPA), el cual recoge todas las recomendaciones de documentación cuando se desarrolla la anonimización de la base de datos.

### Etapa I. Análisis exploratorio de los datos

En esta etapa se deben caracterizar los datos a anonimizar, teniendo en cuenta aspectos procedimentales y temáticos. Está compuesto de cinco pasos:

#### A. Caracterización de la base de datos

Para iniciar con la etapa, el equipo de trabajo caracterizará cada una de las variables contenidas en la base de datos teniendo en cuenta si son cuantitativas (continuas o discretas) o categóricas (4).

**Recuadro 4.** Clasificación de las variables por su tipo

Las variables se pueden clasificar por el tipo de valores que toman:

- **Variable continua:** es una característica medida en las unidades de observación que puede tomar valores dentro de un intervalo específico (infinito) (Sistema de Consulta de Conceptos del DANE, 2018).

Por ejemplo: la altura medida en centímetros, el peso medido en kilogramos, la temperatura medida en grados centígrados, entre otros.

- **Variable discreta:** es una característica medida en las unidades de observación donde su conjunto de posibles valores no corresponde a un intervalo ya que presenta interrupciones (Sistema de Consulta de Conceptos del DANE, 2018).

Por ejemplo: número de hijos, número de intentos en un experimento, número de personas de un hogar, entre otros.

- **Variable categórica:** es una característica medida en las unidades de observación que asigna una de varias categorías cualitativas (Sistema de Consulta de Conceptos del DANE, 2018).

Por ejemplo: una variable dicotómica, que asigna el valor de 1 si el individuo cumple cierta característica y 0 si no la cumple, o estado civil, que asigna a cada individuo alguna de las categorías: soltero, casado, separado, divorciado o unión libre.

Para desarrollar este subproceso, el equipo de trabajo, con ayuda del diccionario de la base de datos (o de toda la documentación disponible), describirá las dimensiones de la base en términos del número de variables, el número de registros y la distribución de todas las variables con respecto a su tipo (cuantitativa o categórica).

Otra clasificación que el equipo de trabajo debe tener en cuenta para las variables de la base de datos es el tipo de información que contiene de la unidad de observación. Estas se clasifican en variables de identificación, de ubicación y temáticas. Por ejemplo: el número de identificación de una persona es una variable de identificación; el municipio; la localidad y el departamento de una entidad, son variables de ubicación, y el ingreso promedio mensual, el nivel de escolaridad y el número de hijos, son variables temáticas.

La caracterización de la base de datos se puede hacer teniendo en cuenta como ejemplo la Tabla 2.

**Tabla 2.** Clasificación de variables de base de datos

| Variables/tipo de variable | Cuantitativas | Categóricas | Total |
|---|---|---|---|
| **Identificación** | Escriba en este espacio el número de variables cuantitativas de identificación | | |
| **Ubicación** | | Escriba en este espacio el número de variables categóricas de ubicación. | |
| **Temáticas** | | | Escriba en este espacio el número de variables temáticas. |
| **Total** | Escriba en este espacio el número de variables cuantitativas. | Escriba en este espacio el número de variables categóricas. | Escriba en este espacio el total de variables de la base de datos. |

Fuente: DANE- DIRPEN.

### B. Cálculo de medidas descriptivas de las variables cuantitativas

Después de que el equipo realice la clasificación de las variables de la base de datos, procederá a calcular las medidas descriptivas estadísticas para cada una de las variables cuantitativas. Un ejemplo de estas se presenta a continuación:

**Tabla 3.** Medidas descriptivas principales para variables cuantitativas

| Variable cuantitativa | Media | Varianza | Cuartil 1 | Cuartil 2 | Cuartil 3 |
|---|---|---|---|---|---|
| **Escriba en este espacio el nombre de la variable cuantitativa.** | | | Escriba en este espacio el primer cuartil de los datos de la variable. | | |

Fuente: DANE- DIRPEN.

Estas medidas servirán como insumo para el análisis de riesgos (Etapa II), el análisis de viabilidad del proceso de anonimización (Etapa IV) y la evaluación de resultados (Etapa VI). Este tipo de medidas se conocen como propiedades globales de las variables y estarán sujetas a verificación por parte del equipo de trabajo para examinar el resultado del proceso de anonimización.

### C. Cálculo de frecuencias de las variables categóricas

Las distribuciones de frecuencias para las variables categóricas hacen parte de las medidas para verificar el proceso de anonimización de la base de datos. Estas distribuciones pueden realizarse teniendo en cuenta la siguiente tabla:

**Tabla 4.** Distribución de frecuencias para una variable con dos categorías

| Variable categórica | Número de unidades de observación que cumplen la categoría | Porcentaje de registros que cumplen la categoría |
|---|---|---|
| **Categoría 1** | Escriba en este espacio el número de unidades de observación que cumplen con la categoría 1. | Escriba en este espacio el porcentaje de unidades de observación que cumplen con la categoría 1. |

| Variable categórica | Número de unidades de observación que cumplen la categoría | Porcentaje de registros que cumplen la categoría |
|---|---|---|
| **Categoría 2** | Escriba en este espacio el número de unidades de observación que cumplen con la categoría 2. | Escriba en este espacio el porcentaje de unidades de observación que cumplen con la categoría 2. |
| **Total** | Escriba en este espacio el número total de unidades de observación que cumplen con las categorías 1 y 2. | Escriba en este espacio el porcentaje de unidades de observación que cumplen con la categoría 2. |

Fuente: DANE- DIRPEN.

### D. Revisión temática del contenido de la base de datos

El equipo de trabajo, después de analizar cada una de las variables, debe realizar una revisión temática de la base de datos, teniendo en cuenta la documentación de la operación estadística o el registro administrativo. Para ello se sugiere responder las siguientes preguntas:

- ¿Cuál es el objetivo de la operación estadística o del registro administrativo?
- ¿Qué cambios metodológicos ha presentado la operación estadística (o registro administrativo) en los últimos tres años?
- ¿Qué tipo de estándares, clasificaciones o nomenclaturas siguen las variables de la base de datos obtenida por la operación estadística o por el registro administrativo?
- ¿Se están usando apropiadamente los estándares, las clasificaciones o las nomenclaturas?
- ¿Qué documentos existen acerca de la operación estadística (o registro administrativo)? ¿Son de fácil acceso?
- ¿Existen otras operaciones estadísticas (o registros administrativos) relacionadas con la temática de la base de datos?

Esta revisión temática servirá como insumo en el planteamiento de los riesgos de identificación de las unidades de observación (Etapa II) y en el análisis de viabilidad del proceso (Etapa IV). Con esto, finaliza el subproceso de Análisis exploratorio de la base de datos.

### E. Definición de las propiedades estadísticas a conservar en la base de datos

En este subproceso el equipo de trabajo deberá establecer las propiedades estadísticas que se deben mantener en la base de datos anonimizada, en relación con la base de datos sin anonimizar. Algunas de estas propiedades estadísticas son:

- **Mantener tendencias en las variables a través del tiempo:** esta propiedad hace referencia a que las

variables conserven un comportamiento en determinados periodos de tiempo.

Por ejemplo, si la base de datos de una operación estadística de temática económica contiene la variable ingreso de los hogares colombianos y esta variable ha presentado un comportamiento creciente en el primer trimestre de 2016, al publicar la base de datos anonimizada el equipo de trabajo desea garantizar que esta tendencia se conserve.

- **Mantener propiedades globales de las variables:** el equipo de trabajo debe definir cuáles de las medidas estadísticas descritas en el análisis exploratorio de datos (sección 5.1), para las variables categóricas y cuantitativas, se deben mantener sin variación y para qué niveles de desagregación geográfica o temática. Asimismo, debe decidir cuáles de las propiedades globales pueden presentar alguna variación significativa y hasta qué porcentaje de variación es permitido en la base de datos anonimizada.

  Por ejemplo, un equipo decidió que la propiedad global que desea mantener es el promedio de la variable "Ingreso por hogar". Además, aceptará el proceso de anonimización, solamente si el promedio de la variable en la base de datos anonimizada difiere del promedio en la base de datos sin anonimizar en menos del 1%.

- **Mantener cifras por niveles de desagregación geográfica o temática:** el equipo de trabajo debe definir cuáles medidas estadísticas se deben conservar sin variación en los niveles de desagregación geográfica o temática, para garantizar a los usuarios análisis de estadísticas más sectorizados.

  Por ejemplo, un equipo de trabajo decidió mantener para la variable grupos étnicos los totales de cada categoría a nivel departamental; esta propiedad permite caracterizar la población étnica en cada departamento y con esta información los usuarios pueden realizar análisis estadístico por regiones.

- **Mantener correlaciones entre variables:** esta propiedad busca conservar las posibles relaciones lineales o no lineales que se puedan presentar entre las variables. El equipo de trabajo definirá si mantiene los coeficientes de correlación entre dos o más variables (cuantitativas o categóricas) en la base de datos anonimizada, con el fin de no distorsionar los resultados finales.

  Por ejemplo, un equipo de trabajo decidió mantener las correlaciones entre estrato y avalúo comercial de un predio, esto para conservar la integridad de la información ya que son variables importantes para analizar.

Finalmente, en este subproceso el equipo de trabajo definirá las propiedades estadísticas que deberá tener la base de datos anonimizada, dado que son insumo para evaluar el proceso de anonimización. Además, describirá las propiedades estadísticas a conservar en la base de datos, teniendo en cuenta:

- Descripción de las propiedades estadísticas elegidas por el grupo de trabajo para conservar en la base de datos anonimizada.
- Listado de las variables con la respectiva propiedad estadística a conservar en la base de datos anonimizada.
- Nivel de desagregación geográfica o temática en el que se desea conservar las propiedades estadísticas.
- Porcentajes de variación permitidos por variables y niveles de desagregación geográfica o temática para las propiedades globales en la base de datos anonimizada.

Para visualizar estos elementos de los subprocesos de la Etapa I, se presenta a continuación un ejemplo.

## Ejemplo de la Etapa I. Análisis exploratorio de la base de datos

La base de datos anonimizada de la Encuesta Anual de Comercio (EAC) de 2016[13], cuenta con 64 variables y 10.242 unidades de observación, en este caso empresas. Las variables se pueden caracterizar de la siguiente forma:

**Tabla 5.** Clasificación por tipo de variable de la EAC en el 2016

| Variables/tipo de variable | Cuantitativas | Categóricas | Total |
|---|---|---|---|
| **IDENTIFICACIÓN** | 0 | 2 | 2 |
| **UBICACIÓN** | 0 | 0 | 0 |
| **TEMÁTICAS** | 59 | 3 | 62 |
| **TOTAL** | **59** | **5** | **64** |

Fuente: DANE - EAC, Cálculos DANE – DIRPEN.

[13] Disponible en el Archivo Nacional de Datos.

Además, las medidas descriptivas para cinco variables cuantitativas temáticas de la operación estadística son:

**Tabla 6.** Medidas descriptivas de algunas variables cuantitativas de la EAC en el 2016

| Variable | Media | Varianza | Cuartil 1 | Cuartil 2 | Cuartil 3 |
|---|---|---|---|---|---|
| **TOTAL SUELDOS*** | 973.902 | 2.76E+13 | 137.877 | 274.157 | 629.193 |
| **TOTAL PRESTACIONES*** | 491.284 | 9.30E+12 | 58.905 | 118.604 | 280.252 |
| **VALOR AGREGADO*** | 3.220.106 | 2.40E+14 | 360.691 | 834.109 | 2.058.826 |
| **VENTAS CAUSADAS*** | 24.022.327 | 1.89E+16 | 2.789.874 | 6.335.314 | 14.654.517 |
| **PERSONAL REMUNERADO**** | 52.5 | 114575.7999 | 12 | 19 | 38 |

*Cifras en millones de pesos
**Cifra en número de personas
Fuente: DANE- EAC, Cálculos DANE – DIRPEN.

Además, se presenta la distribución de frecuencias de la variable categórica Organización Jurídica:

**Tabla 7.** Distribución de Frecuencias de la variable Organización Jurídica de la EAC

| Organización jurídica | Número de registros por categoría | % de registros por categoría |
|---|---|---|
| **Sociedad en comandita simple** | 111 | 1,08% |
| **Sociedad en comandita por acciones** | 37 | 0,36% |
| **Sociedad limitada** | 1,676 | 16,36% |
| **Sociedad anónima** | 1,567 | 15,3% |
| **Sucursal de sociedad extranjera** | 30 | 0,29% |

| Organización jurídica | Número de registros por categoría | % de registros por categoría |
|---|---|---|
| **Empresa unipersonal** | 63 | 0,62% |
| **Persona natural** | 1,946 | 19% |

| Organización jurídica | Número de registros por categoría | % de registros por categoría |
|---|---|---|
| **Organizaciones de economía solidaria** | 82 | 0,8% |
| **Entidades sin ánimo de lucro** | 50 | 0,49% |
| **Sociedad por acciones simplificada** | 4,652 | 45,42% |
| **Otro** | 28 | 0,27% |
| **Total** | **10,242** | **100%** |

Fuente: DANE- EAC, Cálculos DANE - DIRPEN

Finalmente, la revisión temática del contenido de la base de datos permite concluir que:

- El objetivo de la EAC es conocer la estructura y el comportamiento económico del sector comercio a nivel nacional y por grupo de actividad comercial, de manera que permita el análisis de la evolución del sector y de la conformación de agregados económicos.

- En la página web del DANE[14] se encuentran disponibles todos los metadatos y los microdatos de la EAC. Los documentos contienen información sobre:

  - Recolección de los datos.
  - Procesamiento de la información.
  - Políticas de acceso a los microdatos.
  - Diccionario de datos.
  - Descripción de cada una de las variables (incluidos los estándares utilizados).
  - Referentes internacionales.
  - Cuestionario.
  - Metodología y ficha metodológica.

[14] Disponible en: https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-interno/encuesta-anual-de-comercio-eac

**Revisión de restricciones de publicación de la información**

- Ley 2ª de 1962, que permite el levantamiento de encuestas nacionales y en especial, las de industria, comercio y servicios.

- El Decreto 1633 de 1960 en su artículo 74 establece que "todas las personas naturales o jurídicas, domiciliadas en el territorio nacional y los empleados públicos, en todos sus niveles, están obligados a suministrar información al DANE, dentro de los plazos que al efecto se señalen, los datos que este requiera para el cumplimiento de sus finalidades". Además, establece que los datos suministrados a la entidad tienen un carácter estrictamente reservado y no podrán darse a conocer al público ni a las entidades oficiales, sino únicamente en resúmenes numéricos.

- La Ley 0079 de octubre 20 de 1993, que regula la realización de los censos y las encuestas, decreta que las personas naturales o jurídicas, de cualquier orden o naturaleza, domiciliadas o residentes en el territorio nacional, están obligadas a suministrar información al DANE. Asimismo, especifica que la entidad podrá imponer multas a quienes incumplan esta disposición.

- El Principio de confidencialidad descrito en la Ley estatutaria 1266 de 2008 en su artículo 4 indica: "Todas las personas naturales o jurídicas que intervengan en la administración de datos personales que no tengan la naturaleza de públicos están obligadas en todo tiempo a garantizar la reserva de la información. Inclusive después de finalizada su relación con alguna de las labores que comprende la administración de datos".

- El artículo 2.2.3.3.5 del Decreto 1743 de 2016 indica: "(...) guardar la confidencialidad de los datos que permitan la identificación y/o localización espacial de las fuentes, cuando estos fueren recolectados exclusivamente para la producción de las estadísticas oficiales y para fines estadísticos".

- Ley 79 de 1993 Artículo 5: "Los datos suministrados al DANE, en desarrollo de censos y encuestas, no podrán darse a conocer al público ni a entidades u organismos oficiales, ni a las autoridades públicas, sino únicamente en resúmenes numéricos, que no hagan posible deducir de ellos información alguna de carácter individual que pudiera utilizarse para fines comerciales, de tributación fiscal, de investigación judicial o cualquier otro diferente del propiamente estadístico".

- Ley Estadística 2335 de 2023 establece que quienes producen estadísticas oficiales deberán proteger los datos confidenciales de tal forma que, en la publicación de resultados estadísticos, la unidad estadística no pueda ser identificada, directa ni indirectamente. Además, deberán tomar todas las medidas regulatorias, administrativas, técnicas y organizativas necesarias para evitar el acceso de personas no autorizadas a los datos individuales.

- Es conveniente tener en cuenta además las excepciones descritas en la Ley 1712 de 2014: la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional, la cual menciona en su título III las excepciones de acceso a la información donde se destacan:

Artículo 18. Información exceptuada por daño de derechos a personas naturales o jurídicas: es toda aquella información pública clasificada, cuyo acceso podrá ser rechazado o denegado de manera motivada y por escrito, siempre que el acceso pudiere causar un daño a los siguientes derechos (...).

Artículo 19. Información exceptuada por daño a los intereses públicos: es toda aquella información pública reservada, cuyo acceso podrá ser rechazado o denegado de manera motivada y por escrito en las siguientes circunstancias, siempre que dicho acceso estuviere expresamente prohibido por una norma legal o constitucional (...).

Artículo 20. Índice de Información clasificada y reservada: los sujetos obligados deberán mantener un índice actualizado de los actos, documentos e informaciones calificados como clasificados o reservados, de conformidad a esta ley. El índice incluirá sus denominaciones, la motivación y la individualización del acto en que conste tal calificación.

Artículo 21. Divulgación parcial y otras reglas: en aquellas circunstancias en que la totalidad de la información contenida en un documento no esté protegida por una excepción contenida en la presente ley, debe hacerse una versión pública que mantenga la reserva únicamente de la parte indispensable. (...).

Artículo 22. Excepciones temporales: la reserva de las informaciones amparadas por el artículo 19 no deberá extenderse por un período mayor a quince (15) años

Acorde con la revisión de esta normatividad, no se evidencia alguna norma que impida publicar la información siempre y cuando se proteja la identificación de las unidades de observación.

## Revisión de las necesidades de los usuarios de la información

El grupo de trabajo, después de revisar las solicitudes de información en la entidad durante los últimos dos años, evidenció que:

- La mayoría de las solicitudes de información que llegan al DANE provienen de gremios, asociaciones, investigadores, académicos, centros de investigación, universidades.

- Las variables ingreso por ventas, gastos de personal (sueldos y prestaciones), costos de la mercancía vendida, gastos de operación, personal ocupado, inventarios y movimientos de activos fijos, son las que mayor demanda de información presentan.

- La información correspondiente al 2016 se encuentra de forma recurrente entre los periodos de tiempo solicitado, a través de las diferentes solicitudes de información que ha recibido la entidad.

- La información que los diferentes usuarios manifiestan en sus respectivas solicitudes corresponde a la información desagregada a nivel nacional.

### Definición de las propiedades estadísticas a conservar en la base de datos

En la Encuesta Anual de Comercio (EAC) el equipo de trabajo decide que las propiedades estadísticas que la base de datos anonimizada debe conservar son:

- Para las variables ventas y personal ocupado **mantener la participación % directo de las divisiones** CIIU Rev. 4. A.C. del sector comercio.

- Para las variables número de empresas, ventas, costo de la mercancía, producción bruta, consumo intermedio, valor agregado, remuneración **mantener los totales a nivel nacional.**

- La información correspondiente al subsector vehículos automotores, motocicletas, sus partes, piezas y accesorios debe **mantener los totales a nivel nacional.**

Con la base de datos anonimizada los usuarios puedan replicar los diferentes cuadros de salida que son publicados en el boletín técnico[15] de la EAC para 2016.

### Etapa II: Identificación y medición de los riesgos presentes en datos

Esta etapa tiene como finalidad presentar los pasos a seguir para reconocer los riesgos de identificación de las fuentes en la base de datos que se pretende anonimizar, así como, proporcionar un procedimiento que permita medir los riesgos identificados. (Ejemplos basados en el censo de población + Unidades de Observación y Registros únicos que conlleven a singularizar).

Esta etapa se proyecta a desarrollar en términos de identificación y medición de riesgos en datos estructurados. En esta etapa el equipo de trabajo se planteará todos los posibles escenarios de riesgo de identificación de las unidades de observación de la base de datos.

Un escenario de riesgo de identificación de información confidencial es aquel en el cual existe una posibilidad de que, mediante la combinación de variables de la base de datos, se puedan identificar características de las unidades de observación que deben ser protegidas por el tipo de información que contiene.

La etapa de análisis de riesgos se compone de los siguientes subprocesos:

- Clasificación de variables por su nivel de sensibilidad.

- Planteamiento de riesgos de la base de datos

- Identificación de unidades de observación riesgosas.

- Creación del informe de riesgos.

### Clasificación de variables por su nivel de sensibilidad

La etapa de análisis de riesgos inicia con la clasificación de las variables de la base de datos por su nivel de sensibilidad. El equipo debe realizar la revisión de los riesgos teniendo en cuenta el contexto de la base de datos y las personas encargadas del procesamiento de la base de datos.

[15] https://www.dane.gov.co/files/investigaciones/boletines/eac/bol_eac_2016.pdf

Se debe tener en cuenta que una **variable se considera sensible** cuando es:

> "aquella que puede afectar la intimidad, la honra, la reputación, la tranquilidad personal o la fama de las personas" (Ley 1266 de 2008).

En este punto los expertos temáticos que componen el equipo de trabajo juegan un rol preponderante.

Las variables pueden clasificarse en:

- **Identificadores directos:** las variables denominadas como identificadores directos son todas aquellas variables que contienen información sensible de identificación o ubicación de las unidades de observación. Estas variables pueden coincidir con las variables denominadas en el análisis exploratorio de la base de datos (Etapa I) como variables de identificación o de ubicación. Un ejemplo de este tipo de variables, es el número de cédula de una persona, NIT de una empresa, dirección de una entidad, entre otros.
- **Pseudoidentificadores:** las variables denominadas como pseudoidentificadores son todas aquellas que, combinadas con otras variables, conllevan a la identificación de las unidades de observación. Estas variables pueden coincidir comúnmente con las variables denominadas temáticas o de ubicación en el análisis exploratorio de la base de datos (Etapa I). Un ejemplo de este tipo de variables es la combinación del nivel de escolaridad con el ingreso promedio mensual en cierto municipio. En este caso, son tres variables consideradas como pseudoidentificadores que permiten la identificación de algunas unidades de observación.
- **No confidenciales:** las variables denominadas como no confidenciales son todas aquellas que no permiten la identificación de las unidades de observación de la base de datos, ni siquiera cuando son combinadas con pseudoidentificadores. Algunos ejemplos de este tipo de variables son: habilidades de conducción de un vehículo de una persona, gusto por el deporte, gustos culturales, entre otros.

El equipo de trabajo puede resumir la clasificación de acuerdo con la siguiente tabla:

**Tabla 8.** Tabla resumen de la clasificación de las variables por su tipo de sensibilidad

| Tipo de variable / tipo de sensibilidad | Identificadores directos | Pseudoidentificadores | No confidenciales | Total |
|---|---|---|---|---|
| **CUANTITATIVAS** | Escriba en este espacio el número de variables cuantitativas clasificadas como identificadores directos. | | Escriba en este espacio el número total de variables cuantitativas clasificadas como no confidenciales. | |
| **CATEGÓRICAS** | | Escriba en este espacio el número total de variables categóricas clasificadas como pseudoidentificadores | | |
| **TOTAL** | | | | Escriba en este espacio el número total de variables. |

Fuente: DANE - DIRPEN.

## Planteamiento de riesgos de la base de datos

Después de la clasificación de todas las variables de la base de datos de acuerdo con su nivel de sensibilidad, el equipo de trabajo deberá establecer los riesgos de identificación de la base de datos que será anonimizada. Los riesgos que defina el equipo de trabajo servirán como insumo en la selección de técnicas de anonimización (Sección 5.3.).

En general, los riesgos se pueden entender como todas las posibles combinaciones de las variables (entre identificadores directos y pseudoidentificadores) y sus niveles de desagregación (geográfica o temática), que pueden aumentar la probabilidad de que una o varias unidades de observación sean identificadas por los usuarios de la información.

Es necesario que al realizar combinaciones entre variables, se identifiquen unidades de observación que deben ser protegidas dentro de sus niveles de desagregación geográfica o temática.

Por ejemplo:

- Si en la base de datos de un hospital de Cartagena, se tiene la información de los pacientes atendidos en 2016, al combinar la información de la edad, tipo de enfermedad, medicamentos recibidos y dirección, es posible identificar a las mujeres de más de 70 años que han tenido cáncer de seno para el barrio la Chinita; esta combinación de las variables permite reconocer a la unidad de observación.

- Al tener la información de los ingresos por hogar de los colombianos, al combinar las variables, ingresos anuales, edad, sexo, corregimiento o vereda, y nivel escolar, podemos identificar que hay un hombre de 75 años con nivel educativo profesional en el corregimiento de Puerto Salazar, con ingresos anuales de $ 35.601.804. Con esta combinación podría identificarse que se trata del único Médico del corregimiento.

- Al tener la información de ventas por catálogo de una empresa de arte, al combinar las variables departamento, sexo, ventas reportadas, artículos, se puede identificar que en Yopal hay un hombre y dos mujeres que poseen las ventas más altas de la región, por lo que se puede inferir que se traten de los líderes premium del departamento.

Una vez definidos los riesgos de la base de datos, el equipo de trabajo los organizará en un listado y posteriormente los priorizará teniendo en cuenta la frecuencia en que pueden ser identificadas las unidades de observación. Algunas recomendaciones sobre cómo definir los niveles de riesgos de identificación son:

- Verificar los identificadores que definitivamente deben ser suprimidos de la base de datos por su nivel de sensibilidad.

- Verificar los identificadores directos que podrían ser agrupados con el propósito de minimizar el riesgo de identificación

- Listar todas las posibles combinaciones de las variables que representen un riesgo de identificación de las unidades de observación.

- Analizar los niveles mínimos de desagregación de los identificadores directos o pseudoidentificadores que puedan ser más riesgosos para de-

terminar la identificación de las unidades de observación. Usualmente estos niveles de desagregación hacen referencia a desagregaciones geográficas o temáticas.

- Revisar las medidas descriptivas estadísticas calculadas para las variables cuantitativas en el *Análisis exploratorio de la base de datos* (Etapa I).

- Revisar la distribución de frecuencias de las variables categóricas calculadas en el *Análisis exploratorio de la base de datos* (Etapa I) y utilizarlas en la identificación de categorías con frecuencias considerablemente bajas. Usualmente este tipo de categorías se asocian como riesgosas porque permiten fácilmente identificar a las unidades de observación.

- Considerar la recodificación de las categorías de las variables que presentan frecuencias considerablemente bajas.

- Verificar las bases de datos de entidades externas encontradas en la revisión temática realizada en el *Análisis exploratorio de la base de datos* (subproceso de la Etapa I). El equipo deberá identificar si todas las variables contenidas en esas bases podrían convertirse en pseudoidentificadores respecto a la base de datos original sujeta a anonimizar. Esta revisión permitirá establecer qué variables de las bases de datos externas, combinadas con las variables de la base de datos a anonimizar, aumentan el riesgo de identificación de las unidades de observación.

Cuando el equipo de trabajo obtenga el listado priorizado de los riesgos de identificación de las unidades de observación de acuerdo con la frecuencia en la que ocurran estos riesgos, procederá a evaluar temáticamente la base de datos con el equipo de trabajo. En este caso, el equipo revisará si es necesario considerar todos los riesgos identificados, si se pueden considerar sólo algunos, o si se pueden construir nuevos riesgos a partir de la primera ronda de riesgos planteados.

Después de que el equipo de trabajo haya evaluado todos los posibles riesgos de identificación y haya decidido finalmente cuáles son los definitivos (o más probables), procederá a identificar qué unidades de observación se considerarán riesgosas bajo los escenarios de riesgos definidos en este subproceso.

### Identificación de unidades de observación riesgosas

Con el listado definitivo de riesgos de identificación priorizados, el equipo de trabajo procederá a identificar con mayor nivel de detalle las unidades de observación que son riesgosas bajo todos los riesgos planteados en el subproceso anterior.

Las unidades de observación riesgosas son aquellas que cumplen con al menos una de las condiciones planteadas por el equipo de trabajo para ser susceptibles a identificación. Una unidad de observación puede ser riesgosa por sólo un riesgo o por todos los riesgos planteados por el equipo de trabajo.

La identificación de las unidades riesgosas, por su parte, puede ser una actividad para realizar por parte del equipo del procesamiento de la base de datos dada su experticia y capacidades técnicas en el manejo de este tipo de archivos de información.

A continuación, la Tabla 9 presenta un resumen sobre la identificación de las unidades de observación que resultan riesgosas.

**Tabla 9.** Resumen de unidades de observación riesgosas en la anonimización teniendo en cuenta tres riesgos

| Riesgo | Variables involucradas | ¿Cuándo se considera una unidad de observación riesgosa? | Número de unidades de observación riesgosas | Porcentaje de unidades de observación riesgosas |
|---|---|---|---|---|
| **RIESGO 1** | Escriba en este espacio qué variables están involucradas en este riesgo. | | | Escriba en este espacio el porcentaje de unidades de observación riesgosas por el riesgo 1 con respecto al total de unidades de observación. |
| **RIESGO 2** | | En este espacio explique qué condiciones debe cumplir una unidad de observación para considerarse riesgosa. | | |
| **RIESGO 3** | | | Escriba en este espacio cuántas unidades de observación se consideran riesgosas bajo el riesgo 3. | |

| Riesgo | Variables involucradas | ¿Cuándo se considera una unidad de observación riesgosa? | Número de unidades de observación riesgosas | Porcentaje de unidades de observación riesgosas |
|---|---|---|---|---|
| **TOTAL** | Escriba en este espacio el número de variables involucradas en el análisis de riesgos | | | Escriba en este espacio el porcentaje de unidades de observación riesgosas con respecto al total de unidades de observación. |

Fuente: DANE - DIRPEN.

## Creación del informe de riesgos

Finalmente, el equipo de trabajo creará un informe de riesgos que será utilizado en la identificación y la aplicación de técnicas de anonimización (Etapa III), el cual describirá cómo se clasifican las variables según su tipo de sensibilidad, los criterios utilizados para la definición de riesgos de identificación y las unidades de observación que son riesgosas a la hora de publicar la base de datos.

El planteamiento de los riesgos de identificación de las unidades de observación debe estar debidamente documentado. De manera que en caso de cambios en el equipo de trabajo responsable de la anonimización se pueda asegurar una trazabilidad del proceso.

Se recomienda que el informe contenga al menos la siguiente información:

- Criterios y aspectos considerados en la definición de los riesgos.
- Listado de riesgos definitivos priorizados.
- La tabla resumen de unidades de observación riesgosas obtenida en el tercer subproceso de esta etapa (Tabla 9).
- Fecha de emisión del informe.

### Producto Etapa II:

- Clasificación de variables por su tipo de sensibilidad.
- Planteamiento de riesgos de identificación.
- Identificación de las unidades de observaciones riesgosas.
- Informe de riesgos de identificación.

Para visualizar estos elementos de los subprocesos de la Etapa I, se presenta a continuación un ejemplo.

## Ejemplo de la Etapa II: Análisis de riesgos

Para ejemplificar la etapa de análisis de riesgo, se utilizará una base de datos simulada a la cual se llamará **COL20**[16], la cual contiene información de 32 variables de identificación, ubicación y socioeconómicas, medidas en 496 personas que se encuentran presentes en 345 municipios a nivel nacional. La descripción de cada una de las variables de **COL20** se encuentra disponible en el Anexo C.

Para el análisis de riesgos, con base en la experiencia del DANE, se iniciará con la clasificación de las variables por su tipo de sensibilidad. Las 32 variables de **COL20**, pueden caracterizarse así:

**Tabla 10.** Clasificación de las variables por tipo de sensibilidad de COL20

| Tipo de variable / tipo de sensibilidad | Identificadores directos | Pseudoidentificadores | No confidenciales | Total |
|---|---|---|---|---|
| **CUANTITATIVAS** | 0 | 8 | 0 | 8 |
| **CATEGÓRICAS** | 6 | 14 | 3 | 23 |
| **TOTAL** | **6** | **22** | **3** | **31** |

Fuente: DANE - DIRPEN.

De la anterior tabla, se puede concluir que 28 variables son sensibles, entre identificadores directos y pseudoidentificadores, donde 20 de las variables son categóricas. Algunas variables contenidas en **COL20** que son identificadores directos son nombres, apellidos, dirección, número de identificación, y respecto a los pseudoidentificadores, se tienen variables como RH, grupo étnico, ingresos anuales, número de bienes raíces, entre otras.

Al revisar la distribución de las variables, se observa que la mayor parte de estas son categóricas, recordando que para este tipo de variables las unidades de observación cuentan con una valoración cuando cumplen una característica particular. Es el caso de la variable Grupo étnico que toma valores *afrocolombiano, gitano, indígena o ninguno*. Teniendo en cuenta estas características de las variables categóricas, es posible utilizar como una medida aproximada para la definición de riesgos de identificación, las distribuciones de frecuencias.

[16] https://www.dane.gov.co/files/investigaciones/boletines/eac/bol_eac_2016.pdf

Posteriormente, siguiendo las recomendaciones propuestas en esta guía, se elaboró el siguiente listado de riesgos de la base de datos **COL20**:

- Los identificadores directos, como número de identificación, tipo de identificación, nombre, apellidos, fecha de nacimiento y dirección, deben ser suprimidos definitivamente de la base de datos porque representa una identificación inmediata de las unidades de observación.

- Las siguientes combinaciones entre pseudoidentificadores se consideran riesgosas porque el cruce entre ellas podría, eventualmente, permitir la identificación de una unidad de observación de la base de datos:

- Ingresos anuales (o mensuales) y nivel de escolaridad a nivel municipal y departamental.

- RH y edad a nivel municipal y departamental.

- Ocupación, nivel de escolaridad, ingresos anuales (o mensuales) a nivel municipal.

- Grupo étnico a nivel municipal y departamental.

- Número de habitaciones de la vivienda, materiales de los pisos del hogar e ingresos anuales (o mensuales) a nivel municipal y departamental.

- Número de bienes raíces con la variable "tiene vehículo" (marca y modelo) a nivel municipal y departamental.

- Número de viajes fuera del país, ocupación e ingresos anuales (o mensuales).

- Valores frecuentes de la variable asistencia a eventos culturales (o deportivos), edad e ingresos a nivel municipal y departamental.

- La desagregación a nivel municipal es altamente riesgosa para la identificación de las unidades de observación con respecto a algunas variables temáticas como es el caso de la variable *Ingresos anuales*, porque se podría tener con certeza que en el municipio El Castillo en el departamento del Meta se encuentran 2 personas con ingresos anuales superiores a 10 millones de pesos.

- A partir de las medidas descriptivas estadísticas calculadas en la Etapa I, se identificaron las variables cuantitativas más sensibles. A continuación, se presentan los resultados de este ejercicio para **COL20**:

**Tabla 11.** Medidas descriptivas de las variables cuantitativas de COL20

| Variable | Media | Varianza | Cuartil 1 | Cuartil 2 | Cuartil 3 |
|---|---|---|---|---|---|
| **Ingresos anuales*** | $ 52.174.008 | 2.15679E+15 | $ 13.188.150 | $ 40.048.606 | $ 89.000.899 |
| **Ingresos mensuales*** | $ 4,347,834 | 1.49777E+13 | $ 1.099.012 | $ 3.337.384 | $ 7.416.742 |
| **Número de hijos nacidos** vivos** | 1,97 | 2,84 | 0 | 2 | 3 |
| **Número de bienes raíces**** | 2,38 | 3,25 | 1 | 2 | 4 |
| **Número de viajes fuera del país**** | 3,47 | 6,74 | 1 | 3 | 6 |

*Unidades: millones de pesos
**Números enteros
Fuente: DANE- DIRPEN.

- Con el propósito de identificar las variables categóricas más sensibles, se buscan aquellas categorías con menor participación a nivel nacional. En este ejemplo, se utilizaron las distribuciones de frecuencia calculadas en la sección 5.2 (Tabla 5), de las variables categóricas que el equipo consideró eran las más sensibles, como por ejemplo: grupo étnico, nivel de escolaridad y RH.

  Es importante recordar que las distribuciones considerablemente bajas presentan altos niveles de riesgo. Por ejemplo, se identificaron las variables de RH; nivel de escolaridad y grupo étnico como variables que permiten fácilmente la identificación de las unidades de observación.

- Estas categorías deben revisarse cuidadosamente, ya que las unidades de observación que pertenezcan a esta categoría pueden ser riesgosas a nivel municipal.

A continuación, se presentan las frecuencias para las variables que tienen un mayor nivel de sensibilidad en la base de dato **COL20**:

**Tabla 12.** Distribución de frecuencias del RH en COL20

| RH | Número de registros | Porcentaje de registros |
|---|---|---|
| **O+** | 244 | 49.2% |
| **A+** | 212 | 42.7% |
| **B+** | 5 | 1.0% |
| **AB+** | 3 | 0.6% |
| **O-** | 23 | 4.6% |
| **A-** | 6 | 1.2% |
| **B-** | 1 | 0.2% |
| **AB-** | 2 | 0.4% |
| **Total** | 496 | 100% |

Fuente: DANE - DIRPEN.

En este caso, las categorías de RH B-, AB-, AB+ y A-, se consideran como las más sensibles, dada su baja frecuencia de aparición en las unidades de observación.

**Tabla 13.** Distribución de frecuencias del nivel de escolaridad en COL20

| RH | Número de registros | Porcentaje de registros |
|---|---|---|
| **Primaria** | 54 | 10,9% |
| **Secundaria** | 26 | 5,2% |
| **Básica Media** | 70 | 14,1% |
| **Técnico** | 30 | 6% |
| **Tecnólogo** | 88 | 17,7% |
| **Profesional** | 130 | 26,2% |
| **Posgrado** | 98 | 19,8% |
| **Total** | 496 | 100% |

Fuente: DANE - DIRPEN.

Para este segundo caso, se tendrían las categorías de secundaria y técnico como los niveles de escolaridad más sensibles dada su baja frecuencia en las unidades de observación.

**Tabla 14.** Distribución de frecuencias del grupo étnico en COL20

| Grupo étnico | Número de registros | Porcentaje de registros |
|---|---|---|
| **Afrocolombiano** | 20 | 4.0% |
| **Indígena** | 12 | 2.4% |
| **Rrom** | 4 | 0.8% |
| **Ninguno** | 460 | 92.7% |
| **Total** | 496 | 100% |

Fuente: DANE - DIRPEN.

En este último caso, las categorías Rrom, indígena y afrocolombiano son los grupos étnicos más sensibles dada su baja frecuencia en las unidades de observación.

Además, se presenta la tabla de identificación de unidades de observación riesgosas, la cual sirve como insumo para la identificación de las técnicas de anonimización a utilizar y definir la viabilidad del proceso de anonimización:

**Tabla 15.** Unidades de observación riesgosas para los cinco riesgos más frecuentes para la anonimización de COL20

| Riesgo | Variables involucradas | ¿Cuándo se considera una unidad de observación riesgosa? | Número de unidades de observación riesgosas | Porcentaje de unidades de observación riesgosas |
|---|---|---|---|---|
| **1** | Ingresos mensuales y departamento | Las tres personas con el ingreso anual más alto en cada departamento. | 99 | 20% |
| **2** | Grupo étnico, departamento | Las personas pertenecientes a un grupo étnico particular a nivel departamental. | 36 | 7,3% |
| **3** | Número de habitaciones de la vivienda, departamento | Todas las personas que vivan en una vivienda con un número de habitaciones por encima del promedio departamental. | 235 | 47,4% |
| **4** | Número de viajes fuera del país y departamento | Todas las personas que hayan viajado fuera del país más veces que el promedio departamental. | 225 | 45,4% |

| Riesgo | Variables involucradas | ¿Cuándo se considera una unidad de observación riesgosa? | Número de unidades de observación riesgosas | Porcentaje de unidades de observación riesgosas |
|---|---|---|---|---|
| **5** | Nivel de escolaridad y departamento | Todas las personas con posgrado en aquellos departamentos con menos de cuatro personas a ese nivel de escolaridad. | 31 | 6,3% |

Fuente: DANE - DIRPEN.

La anterior tabla evidencia que, de los cinco riesgos más probables seleccionados, más del 40% de las unidades de observación de la base de datos tienen riesgo de ser identificadas por las variables *número de habitaciones de la vivienda (47,4%) y número de viajes fuera del país (45,4%)* (Riesgos 3 y 4). De la misma forma, 99 unidades de observación tienen riesgo de ser identificadas por sus ingresos mensuales (Riesgo 1).

Teniendo en cuenta que estas tres variables son cuantitativas, es posible identificar las técnicas de anonimización más idóneas para minimizar el riesgo de identificación de aquellas unidades de observación.

**Recuadro 5. Ejemplo Comisión de la Verdad**

La Comisión de la Verdad de Colombia, establecida en el marco del Acuerdo de Paz firmado en 2016, tiene como objetivo principal esclarecer los hechos y responsabilidades en las violaciones de derechos humanos durante el conflicto armado interno. En línea con su mandato, la Comisión emprendió la tarea de recopilar y analizar información estadística sobre diversos tipos de victimización.

Este organismo llevó a cabo un proceso estadístico integral, utilizando la unión de 112 bases de datos relacionadas con diferentes hechos victimizantes en el conflicto armado. Posteriormente, tras la consolidación de la información, se generaron 100 réplicas de imputación múltiple específicamente para los componentes de homicidios, desapariciones forzadas, secuestros y reclutamiento ilícito. Estas bases de datos contenían registros con variables potencialmente identificadoras como edad, sexo, etnia y ubicación. Su publicación permitiría estudios sobre patrones de violencia, pero también conllevaría riesgos de reidentificación de las víctimas.

Por esta razón, el DANE realizó un riguroso análisis del proceso de anonimización aplicado por el equipo investigador sobre las 100 réplicas imputadas.

**Análisis de riesgo**

La metodología consistió en detectar registros únicos en la base de datos según distintas combinaciones de variables como llaves pseudoidenficadoras. Como resultado, entre más inhabitual la combinación (ej. una mujer indígena de 92 años secuestrada en un municipio de 500 habitantes) mayor es la probabilidad de reidentificación. Se probaron dos escenarios: (1) departamento y fecha del evento y (2) departamento, fecha del evento y municipio.

En el escenario 1 se encontraron 2.661 registros únicos de alto riesgo (1,1% del total), mientras que en el escenario 2 fueron 35.003 registros (4,4%).

Para evaluar la posibilidad real de reidentificación se realizaron ejercicios de rompimiento cruzando las bases de datos con registros externos como el Registro de Población y el Registro de Víctimas. Usando las mismas variables en ambas bases se buscaron coincidencias exactas, encontrando tasas de entre 1,1% y 11,1%.

** Los ejercicios de rompimiento (cruces de la información con bases externas), pueden realizarse para determinar si existen más unidades riesgosas, adicionales a las halladas en la base de datos a anonimizar.

## Etapa III. Selección y aplicación de las técnicas de anonimización

Esta etapa presenta las posibles técnicas de anonimización que se pueden utilizar, en función a las particularidades de las variables contenidas en la base de datos y el uso que se pretenda dar a la base anonimizada.

## Técnicas de anonimización en datos estructurados

En esta etapa el equipo de trabajo conocerá e identificará las técnicas de anonimización más comunes para variables cuantitativas y categóricas. Además, seleccionará una o más técnicas para aplicar a cada uno de los riesgos planteados en la etapa anterior.

La etapa de identificación y selección de técnicas de anonimización se compone de dos subprocesos así:

- Identificación de técnicas de anonimización más comunes y adecuadas.
- Selección de técnicas de anonimización.

## Identificación de técnicas de anonimización más comunes

En este subproceso se presentan de manera general las técnicas de anonimización más comunes por tipo de variable, que permiten minimizar el riesgo de identificación de las unidades de observación.

Tenga en cuenta que las técnicas de anonimización se dividen en:

- **Técnicas basadas en la no perturbación de datos:** estas técnicas utilizan supresiones parciales, reducción o recodificación de la información para minimizar el riesgo de identificación de las unidades de observación. Este tipo de técnicas son comúnmente utilizadas para evitar que los datos atípicos sean de fácil identificación.

- **Técnicas basadas en la perturbación de datos:** estas técnicas se refieren a procedimientos que implican la modificación sistemática de datos (a veces en pequeñas cantidades aleatorias), de manera tal que las cifras no sean lo suficientemente precisas como para revelar información sobre casos individuales. Pueden incluirse nuevos datos, suprimir y modificar los existentes beneficiando la confidencialidad estadística[17].

A continuación, se describen brevemente las técnicas y su implementación teniendo en cuenta el tipo de variables que pueden estar contenidas en las bases de datos. En este caso para aquellas técnicas basadas en no perturbación.

[17] González M. [2017], p. [29].

**Tabla 15.** Unidades de observación riesgosas para los cinco riesgos más frecuentes para la anonimización de COL20

| Técnicas | Descripción | Tipo de variable | Ejemplos variables (base col20) | Referencia bibliográfica |
|---|---|---|---|---|
| **ELIMINACIÓN DE VARIABLES** | Esta técnica suprime toda la información de una variable.<br><br>Se usa cuando la variable contiene información de identificación directa de la unidad de observación. | En variables categóricas | CÉDULA;TIPO DE IDENTIFICACION; NOMBRE; APELLIDOS;DIRECCIÓN;BARRIO; MUNICIPIO;FECHA DE NACIMIENTO;RH<br><br>Estas variables son eliminadas porque debido a la información contenida es posible identificar directamente a las unidades de observación. Además, algunas de ellas contienen información sensible, por lo tanto, permitirían reconocer las unidades de observación. | Hundepool et al. (2010). |

| Técnicas | Descripción | Tipo de variable | Ejemplos variables (base col20) | Referencia bibliográfica |
|---|---|---|---|---|
| **RECODIFICACIÓN GLOBAL** | Esta técnica consiste en combinar diversas categorías de las variables categóricas en una más general que tenga mayor frecuencia y menor información.<br><br>En el caso de las variables continuas, consiste en agrupar por medio de intervalos para mantener la utilidad de los datos.<br><br>Esta técnica es recomendable cuando se desean proteger unidades de observación con riesgo de identificación a partir de las variables pseudoidentificadoras. | En variables continuas o categóricas | GRUPO ÉTNICO;NÚMERO DE HABITACIONES DE LA CASA-;NÚMERO DE VIAJES REALIZADOS FUERA DEL PAÍS; NIVEL DE ESCOLARIDAD<br><br>Estas variables son recodificadas para combinar las categorías de las variables y poder tener en cada nueva categoría una mayor frecuencia de unidades de observaciones. | Hundepool et al., (2012).<br><br>Templ et al., IHSN Working Paper No. 007 (2014). |

| Técnicas | Descripción | Tipo de variable | Ejemplos variables (base col20) | Referencia bibliográfica |
|---|---|---|---|---|
| **CODIFICACIÓN SUPERIOR E INFERIOR** | Esta técnica consiste en proteger la identificación de las unidades de observación que presentan los valores más altos o más bajos de cada variable.<br><br>Se utiliza cuando se presentan valores máximos y mínimos en el nivel de desagregación geográfico o temático que son de fácil identificación. | En variables continuas o categóricas. | TÉCNICA NO UTILIZADA EN LA ANONIMIZACIÓN DE LA BASE DE DATOS EJEMPLO COL20 | Hundepool et al. (2012). |
| **SUPRESIÓN LOCAL** | Esta técnica consiste en reemplazar los valores de una o más variables de las unidades de observación identificadas como riesgosas por valores faltantes.<br><br>Esta técnica se usa cuando la combinación entre las variables pseudoidentificadoras permita la identificación de las unidades de observación. | En variables categóricas. | TÉCNICA NO UTILIZADA EN LA ANONIMIZACIÓN DE LA BASE DE DATOS EJEMPLO COL20 | Hundepool y De Wolf (2012).<br><br>Templ et al., IHSN Working Paper No. 007 (2014). |

Fuente: DANE - DIRPEN.

Respecto a las técnicas basadas en perturbación, se tienen las siguientes recomendaciones:

**Tabla 17.** Técnicas basadas en la perturbación de datos según el tipo de variable

| Técnicas | Descripción | Tipo de variable | Ejemplos variables (base col20) | Referencia bibliográfica |
|---|---|---|---|---|
| **MICROAGREGACIÓN** | Esta técnica consiste en reemplazar los valores de algunas unidades de observación, por el valor promedio calculado sobre ellas. | En variables continuas. | EDAD;INGRESOS ANUALES;INGRESOS MENSUALES;NÚMERO DE HIJOS; NACIDOS VIVOS;CUANTAS PERSONAS COMPONEN EL HOGAR; NUMERO DE HABITACIONES DE LA CASA;NUMERO DE BIENES RAICES;NUMERO DE VIAJES FUERA DEL PAÍS<br><br>Estas variables son microagregadas porque se busca proteger la identificación de las unidades de observaciones con los ingresos anuales más altos por departamento. | Hundepool et al., (2012) Templ et al., IHSN Working Paper No. 007 (2014). |

Fuente: DANE - DIRPEN.

| Técnicas | Descripción | Tipo de variable | Ejemplos variables (base col20) | Referencia bibliográfica |
|---|---|---|---|---|
| **REDONDEO** | Esta técnica consiste en sustituir los valores de las unidades de observación en aquellas variables que tienen decimales por valores redondeados (cero decimales). Comúnmente, se usa después de aplicar microagregación y cuando la información de las variables técnicamente se debe expresar en unidades enteras y no decimales. Por ejemplo, número de hijos. | En variables continuas. | EDAD;NÚMERO DE HIJOS NACIDOS VIVOS;NÚMERO DE HABITACIONES DE LA CASA;NÚMERO DE BIENES RAÍCES;NÚMERO DE VIAJES FUERA DEL PAÍS. Estas variables son redondeadas para mantener la información de las variables en unidades enteras después de su microagregación. | Hundepool et al. (2012). |

| Técnicas | Descripción | Tipo de variable | Ejemplos variables (base col20) | Referencia bibliográfica |
|---|---|---|---|---|
| **INTERCAMBIO DE DATOS** | Esta técnica consiste en intercambiar la información de las unidades de observación identificadas con riesgo, con la información de las unidades de observación que no tienen riesgo de identificación. Este intercambio de datos se realiza de manera aleatoria entre pares de observaciones (con riesgo de identificación y sin riesgo). | En variables continuas o categóri- | TÉCNICA NO UTILIZADA EN LA ANONIMIZACIÓN DE LA BASE DE DATOS EJEMPLO COL20 | Hundepool et al., (2010), p. 58. |

| Técnicas | Descripción | Tipo de variable | Ejemplos variables (base col20) | Referencia bibliográfica |
|---|---|---|---|---|
| **AGREGAR RUIDO** | Esta técnica consiste en añadir una cantidad aleatoria definida por el equipo de trabajo sobre los valores de las unidades de observación (ruido aleatorio). Comúnmente, es utilizada cuando se desea proteger las unidades de observación y se ha identificado que por medio de cruces de información con bases de datos externas se expone la información confidencial. | En variables continuas | TÉCNICA NO UTILIZADA EN LA ANONIMIZACIÓN DE LA BASE DE DATOS EJEMPLO COL20 | Hundepool A. 2012, p. 54 Templ et al., IHSN Working Paper No. 007 (2014), p. 9. |

Fuente: DANE - DIRPEN.

Otra de las técnicas que ha ganado un espacio en la literatura y en los procesos de anonimización de bases de datos, especialmente en Estados Unidos, es el uso de datos sintéticos. Está técnica consiste en el uso de datos simulados, siendo una alternativa a los métodos previamente explicados. En este caso, se produce una nueva base de datos mediante el uso de algoritmos de simulación, que conserve las propiedades estadísticas de la base de datos no anonimizada.

Para la generación de los datos simulados se puede hacer uso de métodos como regresión cuantílica, imputación adicional y datos combinados (Hundepool, et al., 2010: 58). Comúnmente, cuando las propiedades estadísticas sobre la base de datos sin anonimizar es no perturbar la información, o hacer lo menos posible, los métodos de datos sintéticos no son los más adecuados, ya que proporcionan las mismas tendencias, propiedades globales o correlaciones; sin embargo, modifican todos los campos a nivel de microdato. Cuando se

modifican todos los campos, el nivel de utilidad de la información puede disminuir considerablemente.

Algunos ejemplos y desarrollos teóricos de este tipo de técnicas se pueden consultar en la siguiente bibliografía:


- Domingo-Ferrer J., Drechsler J. and Polettini S (2009) Report on synthetic data files. Technical report, Deliverable of Project ESSNET-SDC.

- Drechsler J., Bender S. and R¨assler S. (2008a) "Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel" Transactions on Data Privacy 1(3): 105–130.

- Fienberg S.E. (1994) "A radical proposal for the provision of micro-data samples and the preservation of confidentiality" Technical Report 611, Carnegie Mellon

- University Department of Statistics.

- Fienberg S.E. and Makov U.E. (1998) "Confidentiality, uniqueness and disclosure limitation for categorical data" Journal of Official Statistics 14(4): 385–397.

- Liew C.K., Choi U.J. and Liew C.J. (1985) "A data distortion by probability distribution". ACM Transactions on Database Systems 10: 395–411.

- Reiter J.P. (2005a) "Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study" Journal of the Royal Statistical Society, Series A 168:185–205.

- Woodcock S.D. and Benedetto G. 2007 Distribution-preserving statistical disclosure limitation. Disponible en SSRN: http://ssrn.com/abstract=931535.


## Selección de técnicas de anonimización con base a los riesgos identificados

Cuando el equipo de trabajo identifique las técnicas de anonimización más comunes por tipo de variable, seleccionará una o más técnicas que permitan minimizar la ocurrencia de cada uno de los riesgos identificados en la Etapa II.

Con base en la tabla Clasificación de las variables por tipo de sensibilidad de la base de datos simulada construida en la Etapa II (Tabla 10), el equipo puede agregar la columna: Técnica(s) de anonimización a utilizar.

**Tabla 18.** Planteamiento de técnicas de anonimización para cada uno de los riesgos identificados

| Riesgo | Variables involucradas | ¿Cuándo se considera una unidad de observación riesgosa? | Número de unidades de observación riesgosas | Porcentaje de unidades de observación riesgosas | Técnica(s) de anonimización a utilizar |
|---|---|---|---|---|---|
| **RIESGO 1** | Escriba en este espacio qué variables están involucradas en este riesgo. | | | Escriba en este espacio el porcentaje de unidades de observación riesgosas por el riesgo 1 con respecto al total de unidades de observación. | Escriba en este espacio la(s) técnica(s) de anonimización que minimizarán la ocurrencia del riesgo 1. |
| **RIESGO 2** | | En este espacio explique qué condiciones debe cumplir una unidad de observación para considerarse riesgosa. | | | |
| **RIESGO 3** | | | Escriba en este espacio cuántas unidades de observación se consideran riesgosas bajo el riesgo 3. | | |

| Riesgo | Variables involucradas | ¿Cuándo se considera una unidad de observación riesgosa? | Número de unidades de observación riesgosas | Porcentaje de unidades de observación riesgosas | Técnica(s) de anonimización a utilizar |
|---|---|---|---|---|---|
| **TOTAL** | Escriba en este espacio el número de variables involucradas en el análisis de riesgos. | | | Escriba en este espacio el porcentaje de unidades de observación riesgosas con respecto al total de unidades de observación. | Escriba en este espacio todas las técnicas de anonimización que utilizará para minimizar la ocurrencia de todos los riesgos. |

Fuente: DANE- DIRPEN.

Es importante destacar que el equipo de trabajo deberá realizar la selección de las técnicas de anonimización, teniendo en cuenta el tipo de variables involucradas en cada uno de los riesgos y de las propiedades estadísticas que se desean conservar en la base de datos.

Las técnicas de anonimización más comunes fueron presentadas en el anterior subproceso, con una referencia bibliográfica que se puede consultar para profundizar en su aplicación.

La Tabla 18 (Planteamiento de técnicas de anonimización para cada uno de los riesgos identificados) sirve como insumo para la toma de decisiones en el análisis de viabilidad que se realizará en la siguiente etapa.

**Productos de la Etapa III:**

- Características de las técnicas de anonimización.
- Técnicas de anonimización para cada uno de los riesgos identificados.

### Recuadro 6. Experiencia Proyecto Demografía Empresarial

Los Indicadores de Demografía y Dinámica Empresarial brindan información estadística sobre los principales eventos demográficos de las empresas activas formales ubicadas en el territorio nacional. Calcula la población total, los nacimientos y la trayectoria de vida de empresas activas empleadoras y económicas formales en el territorio nacional, considerando desagregaciones por tamaño según el número de empleados, actividad económica según la Clasificación Industrial Internacional Uniforme Revisión 4 Adaptada para Colombia (CIIU Rev. 4 AC) y organización jurídica de las empresas.

Con la finalidad de lograr mayores desagregaciones y precisión en la información, Demografía empresarial utiliza el Directorio Estadístico de Empresas (DEE) y el Registro Estadístico de Relaciones Laborales (RELAB), operaciones que se alimentan de registros admimistrativos y que poseen un periodo de actualización mensual.

En este sentido se realiza la integración de las fuentes de información en una única base que posea la información relevante para calcular los indicadores. Dentro de las variables se encuentran: Naturaleza jurídica, Representante legal, Tamaño de la empresa, Fecha de constitución de la empresa, Sector económico, Si es una empresa económica o no, entre otras.

En el proceso de anonimización se seleccionaron las siguientes técnicas.

```
# Adición de ruido

database <- addNoise(
                obj = database,
                variables = c('ocu_depind', 'ocu_depind_h', 'ocu_depind_m'),
                noise = 0.5,
                method = "addition"
            )
```

Con el método de Adición de Ruido se busca modificar los datos (numéricos) de las variables ‘ocu_depind’, ‘ocu_depind_h’ y ‘ocu_depind_m’ para que sus frecuencias se homogenicen y que no haya picos tan pronunciados en algunos valores.

```
# Supresión local

database <- localSupression(
                obj = database,
                keyVars = c(
                    'razonsocial',
                    'aportante_tipo_documento',
                    'aportante_numero_documento',
                    'sexo_rl'
                ),
                k = 3
            )
```

Con el método de Supresión Local, asignándole un valor de k=3, se busca una 3-anonimización, es decir, que no haya menos de 3 registros diferentes que compartan mismos valores en las variables seleccionadas. La forma de lograr esta corrección es reemplazando el valor de una variable por Null/missing y así lograr una anonimización efectiva.

```
# Aleatorización de variables

database <- shuffle(
                obj = database,
                method = 'ds',
                form = pila_16 + pila_17 + pila_18 + pila_19 + pila_20 + pila_21 + pila_22
            )
```

El método de Aleatorización de Variables busca aleatorizar la base de datos usando una combinación lineal de los valores de las variables usadas. Esto ayuda a que la base quede anonimizada, pero que las correlaciones entre variables se mantengan gracias a la combinación lineal antes propuesta.

## Etapa IV. Análisis de viabilidad

En esta etapa se describen algunos criterios para tener en cuenta por el equipo de trabajo al analizar la utilidad del proceso de anonimización de la base de datos.

En términos generales, esta etapa busca establecer el beneficio que puede proveer el equipo de trabajo al generar mayores desagregaciones de la información, frente a los riesgos de identificación de las unidades de observación que se encuentran en la base de datos. Frente a esto, el equipo de trabajo deberá analizar las necesidades de los usuarios, las limitaciones normativas, las políticas de la entidad y los aspectos temáticos de la base de datos.

Para adelantar este proceso, se emite el concepto de viabilidad del proceso de anonimización considerando los siguientes criterios:

1. **Revisión temática y normativa:** después de la revisión normativa, planteada en la sección 4.1. y la revisión de la documentación temática prevista en la sección 5.3.3., el equipo analizará si se encontró alguna norma, ley o una directriz temática de la entidad productora de la información, que impida la publicación de la mayoría de las variables incluidas en la base de datos, o de las variables más útiles para los usuarios. A partir de esta revisión, el equipo podrá considerar que el proceso de anonimización de la base de datos no es viable.

2. **Técnicas de anonimización:** después de analizar la selección de las técnicas previstas en la sección 5.3.3., el equipo podría identificar que ninguna de estas permite que la base de datos anonimizada conserve las propiedades estadísticas definidas en la sección 5.1.5., por lo que podría considerar que no es viable realizar el proceso de anonimización.

3. **Nivel de utilidad de la información:** cuando el equipo revise todas las limitaciones a nivel normativo y las técnicas disponibles para las variables contenidas en la base de datos, deberá analizar el nivel de utilidad de la información que podrá ser

publicada a los usuarios. Si esta utilidad es baja y no permite la réplica de las cifras publicadas por la entidad, se considerará que el proceso de anonimización no es viable.

Cuando el equipo identifique que el proceso de anonimización se ve afectado por el primer criterio, definitivamente no es viable. Sin embargo, cuando se presentan los criterios 2 y 3, existe una forma alternativa de anonimizar la base datos:

- El equipo definirá propiedades estadísticas más flexibles para que la base de datos anonimizada pueda conservar y garantizar la utilidad de la información para los usuarios. La flexibilidad de las propiedades estadísticas puede darse al cambiar el nivel de desagregación geográfica o temática, modificándolas a categorías más generales, al aumentar la posible variación en las propiedades globales o al eliminar variables sensibles que son de alto riesgo de identificación.

**Por ejemplo:** un equipo deseaba publicar la variable "Ingreso promedio por hogar" a nivel departamental. Sin embargo, con las técnicas existentes, evidenció que aún se podían identificar algunas unidades de observación. Por lo tanto, decidió modificar la desagregación de nivel departamental a nivel regional, que, aunque no conserva la utilidad que se esperaba, sigue siendo información relevante para el usuario y además permite la réplica de las cifras publicadas por la entidad.

Finalmente, si el equipo logra proponer nuevas propiedades estadísticas que permiten que la base de datos anonimizada siga siendo útil para los usuarios, el proceso se considera viable.

A modo de resumen frente a los tres criterios, el equipo de trabajo podría diseñar un marco para establecer el riesgo máximo tolerable para anonimizar la base de datos, teniendo en cuenta el siguiente esquema.

**Tabla 19.** Criterios para analizar la viabilidad de la anonimización de la base de datos

| Criterio | Nivel de afectación | Decisión |
|---|---|---|
| **1. Revisión temática y normativa** | **Alta** | No es viable la anonimización de la base de datos. |
| | Todas las variables presentan riesgos. | |
| **2. Técnicas de anonimización** | **Medio** | Flexibilizar las propiedades estadísticas de la base de datos (Etapa III. Selección de las técnicas de anonimización). |
| | Se podrían identificar algunas unidades de observación. | |
| **3. Nivel de utilidad de la información** | **Medio** | |
| | Se podrían identificar algunas unidades de observación. | |

Fuente: DANE - DIRPEN.

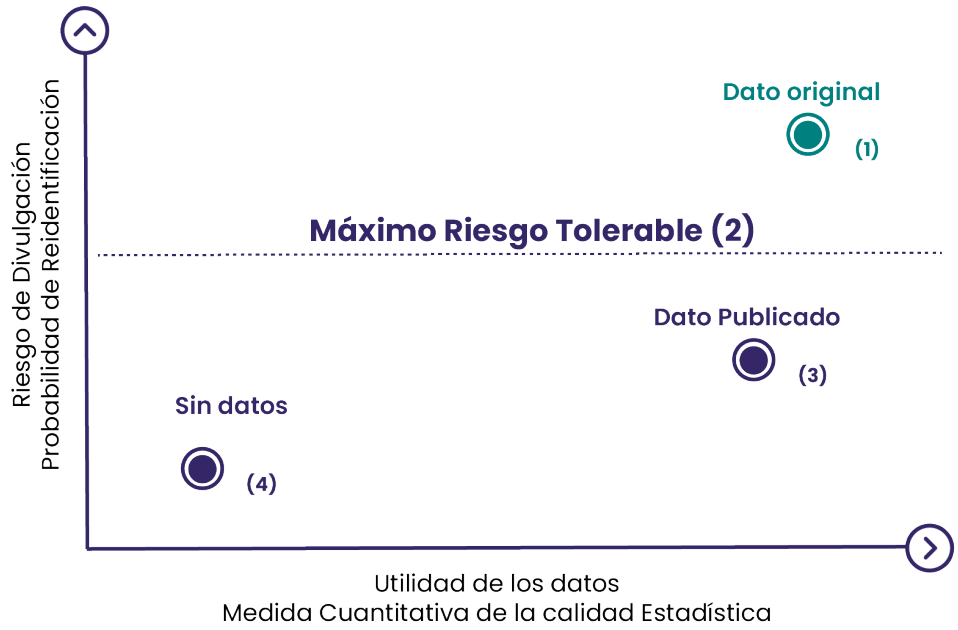
**Producto Etapa IV:**

- Informe y concepto de viabilidad del proceso de anonimización.

En términos conceptuales, el análisis de viabilidad que realiza el equipo de trabajo en esta etapa, puede verse de forma detallada en el Recuadro 7.

### Recuadro 7. Análisis de viabilidad en términos conceptuales

En en la literatura, al buscar identificar el nivel máximo de riesgo tolerable para anonimizar la información, esto es el balance entre los beneficios y los riesgos de anonimizar la información, se puede utilizar el siguiente esquema de revisión.




En esta gráfica se muestra la relación entre la utilidad de los datos o alguna medida cuantitativa de la calidad estadística de la base de datos y el riesgo de divulgación de la información o la probabilidad de reidentificación de la información.

Al hacer el balance entre la base de datos original y la base de datos anonimizada, se pueden tener varios escenarios: por ejemplo, en el punto (1) se observa que la base de datos original presenta un alto nivel de utilidad para el usuario de la información, y a su vez, tiene una alta probabilidad de que las unidades de observación sean identificadas. De hecho, la base de datos original se encuentra por encima del *máximo riesgo tolerable*, punto (2), riesgo que el equipo encargado del proceso de anonimización consideró no podría sobrepasarse para así conservar la confidencialidad de todas las unidades de observación.

El segundo escenario, puede darse cuando el equipo decide no publicar la base de datos, punto (4), dado que la utilidad de la información para el usuario es nula, así como el riesgo de identificar las unidades de observación. Por lo que no tendría beneficios explícitos la decisión de anonimizar o mantener la base de datos en su estado original.

Finalmente, en el último escenario, después de la aplicación de las técnicas de anonimización, el dato publicado, punto (3), es el que se considera que conserva el equilibrio adecuado entre la utilidad de la información para el usuario y el riesgo de identificación de las unidades de observación, además, se encuentra por debajo del *máximo riesgo tolerable.*

## Etapa V. Aplicación de técnicas de anonimización

En esta etapa el grupo de trabajo implementará las técnicas de anonimización asociadas a los riesgos de identificación seleccionadas en la Etapa II, obteniendo así una primera versión de la base de datos anonimizada que será examinada cuidadosamente en la siguiente etapa. Para la aplicación de las técnicas el equipo debe tener en cuenta los siguientes pasos:

- Clasificación de las técnicas asociadas a cada riesgo identificado, tal y como se evidencia en la Tabla 18 (Planteamiento de técnicas de anonimización para cada uno de los riesgos identificados).

- Seleccionar el software que permita implementar las técnicas de anonimización elegidas con el fin de planear de manera eficiente el proceso de anonimización. Dentro de los paquetes de software a tener en cuenta están µ-Argus (es un software gratuito y de código abierto que proporciona técnicas de sustitución, supresión y perturbación de datos), desarrollado por el Instituto de Estadística de Países bajos, T-Argus (es un software gratuito de código abierto que proporciona un técnica de anonimización basada en la teoría de la información) cuando se tienen datos agregados o tablas, SAS, también se ha publicado un paquete de anonimización para R, entre otros.

- Es importante que el equipo de trabajo crea rutinas (algoritmos) que ayuden a implementar las técnicas para que disminuya el riesgo de identificación de las unidades de observación. Se recomienda que en las rutinas el equipo de trabajo siga la siguiente estructura:

- Cargue de base de datos a anonimizar.

- Tipos de riesgos identificados y la explicación de estos.

- Consolidación de riesgos identificados.

- Técnica de anonimización a aplicar.

- Verificación de riesgos.

- Exportación de base de datos anonimizada.

A continuación, se presentan una parte de la rutina utilizada por el DANE en el ejercicio simulado para la anonimización de la Encuesta Anual de Comercio (EAC), utilizando el paquete estadístico SAS:

**Recuadro 8. Ejercicio simulado para la anonimización de la EAC**

Se presenta la forma en que fue cargada la base de datos a anonimizar:

```
libname EAC "\\BASES_ANONIMIZADAS\EAC"; run;
data base11;
set work.'2011_if_0000'n;
run;
```

Posteriormente, se presenta la forma en que identificó los valores máximos de la variable Ventas, que corresponde al riesgo 1 previsto en la base de la EAC:

```
PROC SQL;/*RIESGO 1 MAXIMOS A NIVEL NACIONAL*/
CREATE TABLE MAXIMOS
AS SELECT *,  MAX(VENTA) AS MAX_VENTA
FROM BASE11;
QUIT;
DATA MAXIMOS_VENTAS;
SET MAXIMOS;
IF VENTA = MAX_VENTA THEN ID_RIESGO1 = 1;
RUN;
```

Se relaciona la forma en que se considera el número de unidades de observación riesgosas, de acuerdo con los diferentes riesgos planteados en la EAC:

```
PROC FREQ DATA=EAC.EAC_RIESGO2016 ORDER=INTERNAL;
    TABLES Riesgo1 /  SCORES=TABLE;
    TABLES Riesgo2 /  SCORES=TABLE;
    TABLES Riesgo3 /  SCORES=TABLE;
    TABLES Riesgo4 /  SCORES=TABLE;
    TABLES RiesgoTotal /  SCORES=TABLE;
RUN;
```

Finalmente, se presenta la programación para aplicar la técnica de microagregación cuyo objetivo es eliminar, uno de los riesgos de identificación planteado en la base de datos de la EAC.

```
/*GENERAR UN IDENTIFICADOR*/
data ranqueo (where =(ranqueo in (1,2,3)));
set orden;
retain ranqueo 0;
if first.llave then do ranqueo=0;
end;
ranqueo=ranqueo+1;
by llave;
run;

/*APLICA TÉCNICA: PROMEDIAR LOS DATOS MAS ALTOS
DE TODAS LAS VARIABLES POR LLAVE*/
proc sql;
create table Metodo_1 as select
CIIU4,
intio, mean (intio) as promedio_intio,
/*VARIABLES, MEAN(VARIABLE) AS PROMEDIO_VARIABLE,*/
from ranqueo
group by llave; quit;
```

## Producto Etapa V:

- Base de datos anonimizada.
- Rutina del proceso de anonimización.

## Ejemplo de la Etapa V: Aplicación de técnicas de anonimización

Retomando la base **COL20**, esta cuenta con seis variables que son identificadores directos, 22 variables que son pseudoidentificadores y tres variables no confidenciales.

Antes de identificar y aplicar las técnicas de anonimización, se define que, en este ejercicio simulado, el objetivo del proceso de anonimización es mantener los promedios de las variables cuantitativas (edad, ingresos anuales, ingresos mensuales, número de hijos nacidos vivos, cuantas personas componen el hogar, número de bienes raíces) con una variación inferior al 5% a nivel departamental.

El equipo de trabajo lista los riesgos y acorde con el tipo de variable, indica la técnica de anonimización a utilizar, como se evidencia en la Tabla 20.

**Tabla 20.** Riesgos identificados y técnica de datos a utilizar

| Riesgo | Descripción | Variables involucradas | Tipo de variable | Número de unidades de observación riesgosas | Técnica de anonimización |
|---|---|---|---|---|---|
| **1** | Las tres personas con los ingresos anuales más altos por departamento. | - DEPARTAMENTO<br>- INGRESOS ANUALES | - CATEGÓRICA<br>- CUANTITATIVA | 99 | Microagregación |
| **2** | Las personas pertenecientes a un grupo étnico a nivel departamental. | - DEPARTAMENTO<br>- GRUPO ÉTNICO | - CATEGÓRICA<br>- CATEGÓRICA | 36 | Recodificación |
| **3** | Todas las personas que vivan en una vivienda con un número de habitaciones por encima del promedio departamental. | - DEPARTAMENTO<br>- NÚMERO DE HABITACIONES DE LA CASA | - CATEGÓRICA<br>- CUANTITATIVA | 235 | Redondeo; recodificación |
| **4** | Todas las personas que hayan viajado fuera del país más veces que el promedio departamental. | - DEPARTAMENTO<br>- NÚMERO DE VIAJES REALIZADOS FUERA DEL PAÍS | - CATEGÓRICA<br>- CUANTITATIVA | 225 | Redondeo; recodificación |
| **5** | Todas las personas con posgrado en aquellos departamentos con menos de 4 personas a ese nivel de escolaridad. | - DEPARTAMENTO<br>- NIVEL DE ESCOLARIDAD | - CATEGÓRICA<br>- CATEGÓRICA | 31 | Recodificación |

Fuente: elaboración propia.

Para la base de datos anonimizada a publicar se han eliminado las siguientes variables: **CÉDULA, TIPO DE IDENTIFICACIÓN, NOMBRE, APELLIDOS, DIRECCIÓN, BARRIO, MUNICIPIO, FECHA DE NACIMIENTO, RH.** Esta acción se realiza dada la naturaleza de la información registrada en estas nueve variables, consideradas como identificadores directos, y por tanto sensibles, puesto que permitirían reconocer a las unidades de observación en la base de datos.

Posteriormente, se identifica el porcentaje de unidades riesgosas teniendo en cuenta el listado de riesgos previstos en la Tabla 20 y priorizándolos de acuerdo con el número de riesgo que pueden afectar a las unidades de observación, tal y como se detalla en la Tabla 21.

**Tabla 21.** Porcentajes de unidades de observación por números de riesgos

| Riesgo total | N° de casos | % de caos |
|---|---|---|
| **Ningún riesgo** | 109 | 22% |
| **Un solo riesgo** | 204 | 41% |
| **Dos riesgos** | 137 | 27% |
| **Tres riesgos o más** | 46 | 10% |
| **TOTAL** | **496** | **100%** |

Fuente: elaboración propia.

A partir de la información de la Tabla 20, se encuentra que el 77% de las unidades de observación presentan algún tipo de riesgo, por lo tanto, se define que para este ejercicio las unidades que presenten uno o más riesgos, son las que deben ser anonimizadas.

Adicional al establecer el porcentaje de unidades que presentan algún riesgo, es importante detallar el porcentaje de unidades de observación que pertenecen a cada uno de los riesgos priorizados, como se muestra en la Tabla 22.

**Tabla 22.** Porcentajes de unidades de observación para cada riesgo priorizado

| Tipo de riesgo | | N° de casos | % de caos |
|---|---|---|---|
| **Riesgo 1** | Sin riesgo | 397 | 80% |
| | En riesgo | 99 | 20% |
| **Riesgo 2** | Sin riesgo | 460 | 92% |
| | En riesgo | 36 | 7% |
| **Riesgo 3** | Sin riesgo | 261 | 52% |
| | En riesgo | 235 | 47% |
| **Riesgo 4** | Sin riesgo | 271 | 54% |
| | En riesgo | 225 | 45% |
| **Riesgo 5** | Sin riesgo | 465 | 93% |
| | En riesgo | 31 | 6% |

Fuente: elaboración propia del equipo de trabajo.

El procedimiento de anonimización para **COL20**, se describe a continuación:

1. Para lograr minimizar el riesgo de identificación que se genera por la variable ingresos anuales (Riesgo 1), se analizan cada una de las técnicas expuestas para variables continuas (Tabla 17).

   El riesgo 1 se presenta cuando los ingresos anuales son desagregados geográficamente ya que presentan tres unidades de observación riesgosas por departamento. Para este caso, la técnica de redondeo no minimizaría el riesgo de identificación dado que no perturbaría la información lo suficiente.

   La técnica de intercambio de datos únicamente cambiaría los ingresos anuales de persona y mantendría los mismos valores altos en los ingresos, lo cual permitiría la identificación de las unidades de observación. La adición de ruido perturbaría la información en un nivel muy bajo y los ingresos altos aún seguirían siendo riesgosos.

   Por estas razones, se decidió que

la técnica de microagregación es la que mejor perturba la información, dado que reemplaza los tres valores más altos por un mismo valor y disminuye considerablemente la probabilidad de identificar alguna de las unidades de observación.

Para iniciar la aplicación de la técnica, se deben:

- Identificar las unidades de observación asociadas al riesgo 1, es decir las tres personas con ingresos altos para cada departamento.
- Al identificarlas, se verifica que cumplan con características similares; por ejemplo, se revisó el grado de escolaridad y la frecuencia con que las personas viajan por fuera del país. En este caso, se observa que se tienen tres unidades de observación que cuentan con posgrado y que han viajado uno o más veces por fuera del país.

2. Para aplicar la microagregación, se recodificó la variable nivel de escolaridad así:

- **Bachilleres** para las unidades de observación que habían reportado en su nivel de escolaridad primaria (1), secundaria (2), educación básica media (3).
- **Técnicos** para las unidades de observación que habían reportado en su nivel de escolaridad técnico (4), tecnólogos (5).
- **Profesionales** para las unidades de observación que habían reportado en su nivel de escolaridad profesional (6), posgrado (7), maestría (8), doctorado (9).
- **No reporta** para las unidades de observación que no reportaron información referente a su nivel de escolaridad.

3. Con esta recodificación, se logra obtener en la base de datos anonimizada el mismo grado de escolaridad entre las tres observaciones riesgosas por la variable ingresos mensuales.

4. Respecto a los departamentos, se verifica que las tres unidades de observación con ingresos más altos en la base de datos anonimizada coincidan con la frecuencia asociada a los viajes que realizan fuera del país.

5. Posteriormente, se calcula el promedio de las variables edad, ingresos anuales, ingresos mensuales, número de hijos nacidos vivos, cuantas personas componen el hogar, número de habitaciones de la casa, número de bienes raíces, número de viajes fuera del país.

6. Los promedios encontrados en el punto anterior se reemplazan en los valores originales de las tres unidades de observación para lograr así minimizar el riesgo de identificación. Este proceso es repetido para cada departamento con las tres unidades de observación con ingresos anuales más altos.

Para ejemplificar esta actividad de reemplazamiento de los valores (punto 6) se presenta a continuación para el departamento del Amazonas, la aplicación de la técnica de microagregación de los ingresos anuales para las unidades de observación 71, 127 (unidad de observación riesgosa) y 417 que se encuentran en la base de datos COL20.

Las siguientes tablas presentan la información que tienen dichas unidades de observación en la base de datos, antes de la aplicación de la técnica (Tabla 23), y después la información que tendrán dichas unidades de observación en la base de datos anonimizada (Tabla 24).

**Tabla 23.** Unidades de observación riesgosas en Amazonas. Datos originales.

| ID persona | Departamento | Edad | Ingresos anuales |
|---|---|---|---|
| **71** | AMAZONAS | 52 | $ 87.595.509 |
| **127** | AMAZONAS | 49 | $ 127.500.652 |
| **417** | AMAZONAS | 86 | $ 114.654.116 |

Fuente: DANE - DIRPEN.

**Tabla 24.** Unidades de observación riesgosas en Amazonas. Datos anonimizados.

| ID persona | Departamento | Edad | Ingresos anuales |
|---|---|---|---|
| **71** | AMAZONAS | 62 | $ 109.916.759 |
| **127** | AMAZONAS | 62 | $ 109.916.759 |
| **417** | AMAZONAS | 62 | $ 109.916.759 |

Fuente: DANE - DIRPEN.

7. Al momento de asignarle el valor promedio a las tres unidades de observación con los ingresos más altos por departamento, las variables *edad, número de hijos nacidos vivos, número de habitaciones de la casa, número de bienes raíces, número de viajes fuera del país,* se observa que estos valores registran decimales.

   A partir de ese resultado, se procede a aplicar la Técnica de Redondeo que implica dejar los registros sin unidades decimales. Este procedimiento se realiza para cada una de las variables para mantener los valores exactos y establecer como condición que los valores sean números enteros.

   Por ejemplo, una mujer que tiene en la variable número de hijos nacidos vivos y que tenga reportado 3,2, en la base anonimizada este valor será redondeado a 3.

   Con la aplicación de estas técnicas se logra minimizar el riesgo de identificación de las unidades de observación para el riesgo 1 (*Las tres personas con los ingresos anuales más altos por departamento*), y el riesgo 5 (*Todas las personas con posgrado en aquellos departamentos con menos de cuatro personas a ese nivel de escolaridad*) (Tabla 20).

8. Para el Riesgo 2, es decir para *Las personas pertenecientes a un grupo étnico a nivel departamental*, se aplica la Técnica de Recodificación a la variable *grupo étnico*.

   En este caso, se define que las observaciones que reportaron ser afrocolombiano (1), indígena (2) y Rrom (3), se les asignaría la categoría "*Pertenece a etnia*" y las que no habían reportado pertenecer a una etnia, se les asignaría la categoría "*No Pertenece a una etnia*". Al aplicar esta técnica, se enmascara la pertenencia al grupo étnico por parte de las unidades de observación, siendo la única técnica que nos permite este tipo de proceso.

   En la base de datos anonimizada no se podrá diferenciar a qué etnia pertenecen las unidades de observación; sin embargo, los usuarios sí podrán calcular el porcentaje de las unidades de observación que pertenecen o no a un grupo étnico por departamento.

9. Para el Riesgo 3, es decir para *Todas las personas que vivan en una vivienda con un número de habitaciones por encima del promedio departamental, se aplica la Técnica de Recodificación a la variable número de habitaciones de la casa* teniendo en cuenta los valores obtenidos en el paso número 5.

10. Se definen los siguientes rangos a la variable número de habitaciones de la casa de "0 – 3"; "4 – 10" y luego se asignó su correspondiente categoría a cada unidad de observación.

    Con esta técnica, se minimiza la identificación de las unidades de observación que viven en una vivienda con un número de habitaciones por encima del promedio departamental. Con la recodificación realizada en la base anonimizada, se puede tener la frecuencia de las viviendas en cada uno de los diferentes rangos establecidos.

11. Para el riesgo 4, esto es, *Todas las personas que hayan viajado fuera del país más veces que el promedio departamental*, se aplica la Técnica de Recodificación para la variable *número de viajes realizados fuera del país.*

    A partir de esto, se definen las siguientes categorías para la variable: entre 0 y 2 viajes "0 - 2" y para más de 2 viajes "más de 2". Con esta técnica, en la base de datos anonimizada se puede tener la frecuencia del número de viajeros en cada uno de los diferentes rangos establecidos.

Al aplicar estas técnicas de anonimización se obtiene una primera versión de la base de datos anonimizada. En este caso, se recomienda tener copias de las bases de datos, e indicar los criterios aplicados para disminuir los riesgos identificados en la base de datos original.

En resumen, para comprender los pasos y recapitular las acciones desarrolladas en la base COL20, es importante destacar que durante el proceso de anonimización se aplicó la técnica de *Recodificación a tres variables (número de viajes fuera del país, grupo étnico, número de habitaciones de la casa)* y se eliminaron el 28,1% de las variables iniciales, esto es, 9 de las 32 variables iniciales.

Con la primera versión de la base de datos anonimizada se procede a evaluar si las unidades de observación identificadas como riesgosas ya no presentan algún riesgo, en caso de presentarse nuevas observaciones con riesgo se deben buscar técnicas alternativas de anoni-

mización. La primera versión de la base de datos anonimizada será insumo para la etapa de evaluación de los resultados obtenidos.

**Recuadro 9. Ejemplo del Proyecto Censo Nacional de Población y Vivienda 2018**

El DANE en 2018 realizó el Censo Nacional de Población y Vivienda (CNPV2018), una operación estadística a gran escala que permitió recopilar información detallada sobre las características demográficas y socioeconómicas de toda la población colombiana.

Dado que los microdatos censales contienen información sensible sobre los individuos, como edad, sexo, estado civil, nivel educativo, ingresos, lugar de residencia, entre otros, existe el riesgo de que esa información pueda ser utilizada para identificar a las personas encuestadas y afectar de esta manera la confidencialidad de los datos.

Por esta razón, antes de publicar y difundir los microdatos del CNPV2018, el DANE implementó un riguroso proceso de anonimización con el fin de minimizar al máximo la posibilidad de reidentificación de los individuos a partir de la información disponible.

**Fases del proceso de anonimización**

El proceso de anonimización del CNPV2018 contempló las siguientes fases:

1. Revisiones previas: se realizó una exploración inicial de las bases de datos para identificar variables sensibles y posibles riesgos de filtración de información confidencial.

2. Análisis de riesgos: se evaluaron escenarios de ataque mediante cruces con bases de datos externas y análisis de unicidad de registros. Esto permitió cuantificar los riesgos de reidentificación en diferentes niveles de desagregación geográfica.

3. Identificación y selección de técnicas: a partir de los análisis previos, se determinaron las técnicas de anonimización más adecuadas, como la supresión y la recategorización de variables.

4. Aplicación de técnicas: se implementaron programáticamente las transformaciones definidas sobre las bases de datos originales para generar las versiones anonimizadas.

5. Evaluación de resultados: se realizaron nuevos análisis sobre las bases anonimizadas resultantes para validar que los riesgos de reidentificación se hubieran minimizado significativamente.

**Medición de riesgos de reidentificación**

Uno de los pasos más importantes fue la medición rigurosa de los riesgos de reidentificación mediante la simulación de diferentes escenarios de ataque sobre los datos originales. Para ello se utilizaron técnicas como el análisis de unicidad de registros y los cruces con bases de datos externas que contenían información real sobre un subconjunto de la población colombiana.

En concreto, se cruzó la información censal con los registros administrativos del SISBEN, focalizando el análisis en variables clave como edad, sexo, estado civil, nivel educativo y parentesco. Este ejercicio se repitió para diferentes niveles de desagregación geográfica, desde el total nacional hasta el nivel de manzana.

Los resultados de estos análisis permitieron concluir que sin aplicar técnicas de anonimización, una proporción importante de los registros podrían ser vulnerables a ataques de reidentificación si se publicaran datos a nivel de sección o manzana.

**Ejercicios de ataque a las bases de datos**

Además de los cruces con bases externas, se implementaron otros ejercicios de ataque sobre los datos originales para evaluar escenarios de reidentificación potenciales.

Por ejemplo, se demostró que a partir de ciertos dígitos del número de identificación personal era posible estimar con precisión variables como el sexo, la edad o la región de origen de los individuos. Otra posibilidad evaluada fue la geolocalización de direcciones postales informadas durante la encuesta para derivar la ubicación geográfica a nivel de sector, sección o manzana. Aunque en la práctica este escenario tenía limitaciones, sirvió para dimensionar riesgos en caso de mejorar la calidad de esa información auxiliar en el futuro.

En paralelo se revisaron técnicas alternativas que podrían facilitar la reidentificación, como la ingeniería social, el rastreo web, la suplantación de identidad o los ataques informáticos a la base de datos. Si bien estos escenarios eran menos probables, también debían ser contemplados.

En definitiva, la simulación de múltiples estrategias de ataque desde las capacidades de un hipotético atacante permitió obtener una estimación más realista de los verdaderos riesgos de reidentificación inherentes a los microdatos originales del CNPV2018.

## Etapa VI. Evaluación de resultados del proceso de anonimización

En esta etapa el equipo de trabajo evaluará los resultados del proceso de anonimización y verificará que los riesgos de identificación de las unidades de observación se hayan minimizado y que las variables de la base de datos conserven las propiedades estadísticas deseadas.

Esta etapa se divide en tres subprocesos:

- Revisión de propiedades estadísticas de la base de datos original contra la base de datos anonimizada.
- Reevaluación de riesgos de identificación.
- Creación del Informe Final del Proceso de Anonimización (IFPA).

### Revisión de propiedades estadísticas de la base de datos original contra la base de datos anonimizada

Después de la aplicación de técnicas de anonimización se obtiene una primera versión de lo que sería la base de datos anonimizada. En esta etapa se comparan las propiedades estadísticas de la base de datos anonimizada con respecto a la base de datos original y en caso de que no se cumplan las propiedades esperadas se proponen medidas correctivas (verificación del proceso, aplicación de nuevas técnicas de anonimización, entre otros).

El equipo de trabajo calculará las principales medidas estadísticas sobre todas las variables de la base de datos anonimizada. Con estos resultados, se podrá comparar y concluir, si el proceso de anonimización conservó o no las propiedades estadísticas esperadas con respecto a la base original. Estas propiedades pueden variar según el objetivo que el equipo de trabajo se haya trazado; en algunos casos, se requiere que las medidas globales de las variables (media, varianza, coeficiente de variación, entre otros) se conserven en la base de datos anonimizada, por lo que el equipo de trabajo verificará que las diferencias entre estas medidas no sean significativas.

Por otro lado, existen casos en que el objetivo del proceso es que las propiedades estadísticas de la base de datos anonimizada conserven la tendencia de ciertas variables, la relación lineal entre una variable de estudio y algunas variables explicativas (regresión lineal), o medidas descriptivas sobre segmentaciones (o subpoblaciones) de interés en la población analizada. En estos casos el equipo de trabajo definirá y calculará las medidas que le permitan comparar si las propiedades estadísticas se conservan en la base de datos anonimizada.

Cuando el equipo de trabajo revise la base de datos anonimizada y considere que cumple las propiedades estadísticas procederá al subproceso de reevaluación de riesgos de identificación. Sin embargo, cuando considere que la base de datos anonimizada no las cumple, el equipo debería:

1. Revisar detalladamente la aplicación de las técnicas de anonimización propuestas en la aplicación de técnicas de anonimización. Verificar que no haya errores de procesamiento o que no se esté usando inadecuadamente la técnica.

2. Cuando tenga certeza de que las técnicas planteadas han sido aplicadas adecuadamente, volverá a la Etapa III, donde identificará y aplicará técnicas de anonimización alternativas[18] que no perturben demasiado la base de datos original y permitan el objetivo estadístico planteado.

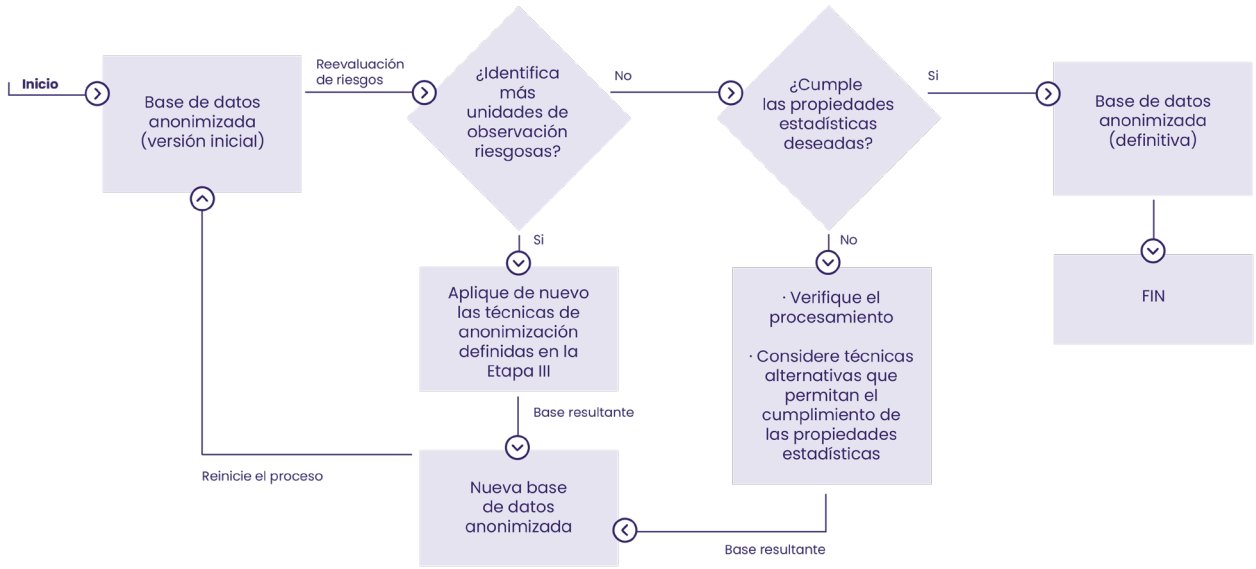
## Reevaluación de los riesgos de identificación

Un criterio que el equipo de trabajo debe tener en cuenta para evaluar el proceso de anonimización es la reevaluación de riesgos de identificación. Para ello, el equipo de trabajo, con base al listado de riesgos priorizados planteado en el Análisis de Riesgos (Etapa II), hará un nuevo análisis de riesgos de identificación, tal y como se hizo con la base original, con el objetivo de identificar nuevas unidades de observación riesgosas.

Es probable que las técnicas de anonimización aplicadas enmascaren la información sensible identificada, pero a su vez conviertan nuevas unidades de observación en unidades riesgosas.

En la Ilustración 2 se presenta un flujograma de cómo el equipo de trabajo puede realizar los dos primeros subprocesos de esta etapa.

**Ilustración 2.** Flujograma reevaluación de riesgos de identificación




Fuente: DANE - DIRPEN.

[18] En la etapa selección de técnicas de anonimización se muestran las técnicas más adecuadas por tipo de variable y nivel de perturbación de la información.

## Creación del Informe Final del Proceso de Anonimización (IFPA)

La etapa de evaluación de resultados finaliza cuando el equipo de trabajo obtiene la base de datos anonimizada en su versión definitiva. Esta base de datos final debe cumplir las propiedades estadísticas esperadas y no permitir la identificación de Información sensible de las unidades de observación que se había previsto en la base de datos original.

El equipo de trabajo debe construir el IFPA y que debe seguir la siguiente estructura:

1. **Características del equipo de trabajo e insumos**

   En esta sección es importante que el equipo describa con qué insumos inicia el proceso de anonimización. Debe tener en cuenta:

- El equipo de trabajo encargado del proceso de anonimización: incluir el rol que tiene cada persona en el equipo, esto es, persona con conocimiento temático de la base de datos o la persona encargada del procesamiento de la base de datos.
- Paquete de software utilizado.
- Descripción de la base de datos original: esta sección puede incluir las dimensiones de la base de datos, su formato, periodo, la operación estadística o el registro administrativo a que corresponde la información.
- Diccionario de datos (en caso de que cuente con uno).

2. **Revisiones previas al proceso de anonimización**

   En esta sección el equipo documentará los hallazgos encontrados en la Etapa I del proceso. Debe tener en cuenta:

- Análisis exploratorio de la base de datos.
- Fundamentos legales que respalden o impidan la publicación de la información.
- Historial de solicitudes de Información por parte de los usuarios.
- Propiedades estadísticas que se espera que la base de datos anonimizada conserve.

3. **Análisis de riesgos de identificación de las unidades de observación**

   En esta sección el equipo documentará todos los riesgos planteados en la Etapa II. Debe incluir el Informe de Riesgos, como se indica en la sección 5.2.4.

4. **Selección de técnicas a implementar**

   En esta sección el equipo documentará las técnicas que escogió para cada uno de los riesgos planteados en la Etapa II. Es suficiente con que el equipo incluya la Tabla 18. Planteamiento de técnicas de anonimización para cada uno de los riesgos identificados.

**5. Análisis de viabilidad**

En esta sección se documentará el concepto de viabilidad sobre el proceso de anonimización al que el equipo llegó. Debe justificar claramente las razones para considerar viable o no viable el proceso.

**6. Aplicación de las técnicas de anonimización**

En esta sección el equipo documentará las rutinas utilizadas en la programación del proceso de anonimización.

**7. Evaluación de resultados**

En esta sección el equipo documentará los hallazgos encontrados en la evaluación de resultados. Puede tener en cuenta:

- ¿Las propiedades estadísticas esperadas se cumplen en la base de datos anonimizada con respecto a la base de datos original?
- ¿Debido al incumplimiento de las propiedades estadísticas esperadas se verificó el procesamiento de las técnicas de anonimización? ¿se encontró algún error?
- ¿Se replantearon las técnicas de anonimización propuestas debido al incumplimiento de las propiedades estadísticas esperadas?
- En la reevaluación de los riesgos de identificación, ¿se encontraron nuevas unidades de observación riesgosas?, ¿cuántas?

Finalmente, con la creación del IFPA el proceso de anonimización quedará debidamente documentado. En caso de que el proceso no sea viable el informe incluye las razones normativas, temáticas y procedimentales. Por otro lado, si el proceso es viable, el informe permite la continuación del proceso de anonimización en próximas bases de datos de la misma operación estadística o del mismo registro administrativo.

### Ejemplo de Etapa VI: Evaluación de resultados

Continuando con el ejemplo de **COL20**, después de aplicar las técnicas de anonimización definidas en la sección anterior, en esta etapa se verifica el cumplimiento de las propiedades estadísticas esperadas y se reevalúan los riesgos de identificación.

Como se definió en la Etapa I, la propiedad estadística que se desea mantener en la base de datos anonimizada, es que los promedios de las variables numéricas por departamento no presenten una variación superior al 5%. Para esto, se calculan las variaciones entre los promedios departamentales de las variables que en la base de datos anonimizada continúan siendo cuantitativas. Algunas variables, como número de habitaciones en la vivienda y número de viajes fuera del país se convirtieron en variables categóricas para disminuir el riesgo de identificación. Las variaciones en las variables se presentan a continuación:

**Tabla 27.** Variaciones de los promedios de las variables numéricas a nivel departamental

| Departamento | Ingresos anuales | Ingresos mensuales | Cuántas personas componen el hogar | Número de bienes raíces |
|---|---|---|---|---|
| **AMAZONAS** | 0,00% | 0,00% | -2,44% | -3,57% |
| **ANTIOQUIA** | 0,00% | 0,00% | 1,18% | 0,00% |
| **ARAUCA** | 0,00% | 0,00% | -2,22% | 4,00% |
| **SAN ANDRÉS** | 0,00% | 0,00% | -3,45% | 0,00% |
| **ATLÁNTICO** | 0,00% | 0,00% | -1,23% | 0,00% |
| **BOGOTÁ, D.C.** | 0,00% | 0,00% | 1,30% | -2,13% |
| **BOLÍVAR** | 0,00% | 0,00% | 0,00% | 3,85% |
| **BOYACÁ** | 0,00% | 0,00% | 0,00% | -2,70% |
| **CALDAS** | 0,00% | 0,00% | 0,00% | 0,00% |
| **CAQUETÁ** | 0,00% | 0,00% | 2,63% | 6,67% |
| **CASANARE** | 0,00% | 0,00% | 0,00% | -3,57% |
| **CAUCA** | 0,00% | 0,00% | -3,12% | 0,00% |
| **CESAR** | 0,00% | 0,00% | 0,00% | 0,00% |
| **CHOCÓ** | 0,00% | 0,00% | 0,00% | 0,00% |
| **CÓRDOBA** | 0,00% | 0,00% | -1,79% | -2,94% |
| **CUNDINAMARCA** | 0,00% | 0,00% | 1,08% | 1,82% |
| **GUAINÍA** | 0,00% | 0,00% | -4,17% | -3,70% |

| Departamento | Ingresos anuales | Ingresos mensuales | Cuántas personas componen el hogar | Número de bienes raíces |
|---|---|---|---|---|
| **GUAVIARE** | 0,00% | 0,00% | -3,23% | 0,00% |
| **HUILA** | 0,00% | 0,00% | 1,45% | 2,38% |
| **LA GUAJIRA** | 0,00% | 0,00% | 3,23% | -3,13% |
| **MAGDALENA** | 0,00% | 0,00% | 0,00% | 4,55% |
| **META** | 0,00% | 0,00% | 0,00% | 0,00% |
| **NARIÑO** | 0,00% | 0,00% | -2,94% | 0,00% |
| **NORTE DE SANTANDER** | 0,00% | 0,00% | 0,00% | 0,00% |
| **PUTUMAYO** | 0,00% | 0,00% | 0,00% | 5,88% |
| **QUINDIO** | 0,00% | 0,00% | 0,00% | -2,17% |
| **RISARALDA** | 0,00% | 0,00% | 1,85% | 3,13% |
| **SANTANDER** | 0,00% | 0,00% | 0,00% | 0,00% |
| **SUCRE** | 0,00% | 0,00% | 0,00% | 0,00% |
| **TOLIMA** | 0,00% | 0,00% | -1,41% | 1,47% |
| **VALLE DEL CAUCA** | 0,00% | 0,00% | -1,05% | -1,85% |
| **VAUPÉS** | 0,00% | 0,00% | -4,17% | 5,26% |
| **VICHADA** | 0,00% | 0,00% | 0,00% | -2,94% |

Fuente: DANE - DIRPEN.

Se observa que las variaciones en las variables ingresos anuales e ingresos mensuales es de 0% para todos los departamentos, esto se debe a que las variables son continuas. En cambio, las variables número de bienes raíces y número de personas que componen el hogar, que fueron micro agregadas y además redondeadas (pues deben ser un número entero) presentan variaciones mayores al 0%.

Por otro lado, es evidente que la variable "número de bienes raíces" es la única que no conserva en todos los departamentos una variación inferior al 5%. En el ejercicio se verificó que la variación no corresponde a un error de procesamiento.

El equipo de trabajo considera que las variaciones superiores al 5% para los departamentos de Caquetá y Putumayo son permitidas desde el punto de vista temático, por lo tanto, se considera que con las técnicas utilizadas se mantienen las propiedades estadísticas establecidas al inicio del proceso de anonimización

Un paso que debe tenerse en cuenta para la evaluación de resultados es la reidentificación de unidades de observación riesgosas. En este caso, sobre la base de datos anonimizada se buscan unidades de observación que cumplan con los riesgos planteados en la Etapa II. Se obtiene la información presentada en la siguiente tabla:

**Tabla 28.** Reidentificación de unidades de observación riesgosas

| Riesgo | Unidades de observación riesgosas | Porcentaje |
|---|---|---|
| **1** | 0 | 0% |
| **2** | 0 | 0% |
| **3** | 0 | 0% |
| **4** | 0 | 0% |
| **5** | 8 | 1,61% |

Fuente: DANE - DIRPEN.

La reidentificación de nuevas unidades de observación riesgosas muestra que la ocurrencia de los riesgos 1,2, 3 y 4 fue minimizada correctamente. Sin embargo, dado que aún aparecen ocho unidades de observación riesgosas por el riesgo 5, el equipo propone nuevas categorías para la variable nivel de escolaridad, unificando las categorías técnicos y profesionales en “técnico o profesional” y dejar la categoría Bachiller y así minimiza el riesgo de identificación.

Finalmente, después de que se tuvieron en cuenta los cambios que sugiere la etapa de evaluación de resultados se obtiene la base de datos anonimizada en su versión final. Esta base cumple con las propiedades estadísticas esperadas y no tiene riesgo de que las unidades de observación sean identificadas.

# 6. Recomendaciones

A partir de los puntos revisados en la guía, como un paso final para un proceso de anonimización efectivo, el DANE y el AGN presentan algunas recomendaciones que se pueden tener en cuenta:

1. Si bien las entidades deben garantizar la transparencia y el acceso a la información pública como cumplimiento de derechos de los ciudadanos, antes de iniciar el proceso de anonimización se debe realizar un análisis del contexto de producción de información atendiendo a los parámetros dados en la Ley 1712 de 2014 y tomar en cuenta sus excepciones e instrumentos de gestión de información pública que garanticen su implementación.

2. Es necesario identificar el tipo de información a la que se refieren los datos personales, ya que estos pueden ser sensibles, semiprivados, privados o públicos, acorde con la Ley 1581 de 2012, y visibilizar las acciones previas y pertinentes en la realización del proceso de anonimización.

3. La base de datos que será anonimizada debe cumplir con los requerimientos iniciales previstos en la primera etapa de la guía. Esto permitirá que el proceso de anonimización no se vea afectado por errores de captura de la información o la falta de reglas de validación.

4. Las rutinas del software estadístico utilizadas en el proceso de anonimización deben estar debidamente explicadas para permitir que el proceso pueda repetirse cuando se cuenten con actualizaciones de las bases de datos.

5. La documentación correspondiente a la aplicación de las técnicas de anonimización debe ser información restringida para el equipo de trabajo, debido a que esta información permitiría a terceros revertir el proceso de anonimización y exponer la Información de las unidades de observación que son anonimizadas.

6. Si la base de datos que se dispone para uso de los diferentes usuarios relaciona cuadros de salida publicados por la entidad del SEN a través de sus diferentes medios, es importante que genere una documentación que permita entender la información contenida en la base de datos anonimizada, para que los usuarios puedan replicar las cifras que son publicadas.

7. Utilizar diferentes medios de difusión y estrategias de publicidad para visualizar la información anonimizada que la entidad posee ante las demás entidades pertenecientes al SEN, permite el acceso y el uso de microdatos para la producción y la difusión de estadísticas oficiales.

8. Se recomienda fomentar en las entidades la elaboración de datos abiertos, en el que se asegure su calidad

y armonización de la información, toda vez que estos son fundamentales en los ejercicios ciudadanos de control social que contemplen contempla las características de la información como su oportunidad, gratuidad y facilidad de acceso, entre otras (CONPES 4070 de 2021).

Pautas a tener en cuenta para mantener el proceso de información de bases de datos:

- Actualizar la revisión de normatividad asociada a la privacidad de la información.
- Conocer las regulaciones y familiarizarse con las leyes de privacidad aplicables.
- Evaluar los riesgos y estimar el riesgo de reidentificación de los datos anonimizados.
- Seleccionar métodos robustos y utilizar métodos sólidos de anonimización para el proceso.
- Validar el proceso e implementar técnicas de validación y monitorearlas continuamente.
- Realizar un registro detallado y mantener un registro preciso de actividades de anonimización realizadas.
- Capacitar al personal y asegurarse de que el personal esté bien capacitado en términos de prácticas de privacidad.
- Privacidad por diseño e integrar principios de privacidad desde el diseño.
- Adaptabilidad y ajustar el proceso ante cambios en el contexto o las amenazas de reidentificación.
- Colaborar con expertos y buscar asesoramiento de expertos en privacidad y seguridad.

Si se produce la reidentificación de una base de datos, es esencial tomar medidas inmediatas para mitigar el riesgo y cumplir con las obligaciones legales. A continuación, se mencionan algunos pasos a seguir:

- Detener la divulgación: tan pronto como se detecte la reidentificación se debe detener cualquier forma de divulgación de la información afectada.
- Evaluar el alcance: determinar la extensión del incidente e identificar los datos específicos que han sido comprometidos y cuánto se ha revelado.
- Notificar a las autoridades: si la entidad que custodia la información está sujeta a regulaciones de privacidad, debe notificar a la autoridad de protección de datos correspondiente en el plazo establecido por la ley. Cumplir con los requisitos de notificación es fundamental para evitar sanciones.
- Notificar a los afectados: debe informar a las personas cuyos datos han sido reidentificados. Proporcionarles información clara y concisa sobre la naturaleza del incidente, los datos afectados y las medidas que está tomando para mitigar el riesgo.
- Investigación interna: llevar a cabo una investigación interna para comprender cómo ocurrió la reidentificación y cómo se puede prevenir en el futuro. Evaluar y actualizar tus prácticas de anonimización.
- Mejoras en la seguridad: reforzar las medidas de seguridad, tanto técnicas

como organizativas, para evitar futuros incidentes de reidentificación. Esto podría incluir la revisión de protocolos de anonimización, controles de acceso y auditorías.

- Comunicación transparente: comunicar de manera transparente con las partes interesadas, incluyendo a clientes, empleados y cualquier entidad regulatoria. La transparencia contribuye a la confianza y muestra un compromiso serio con la gestión del incidente.

- Apoyo legal y consultoría: buscar asesoramiento legal para comprender y cumplir con cualquier obligación legal resultante del incidente. Un profesional legal especializado en privacidad puede ayudar a tomar las decisiones correctas.

- Aprendizaje continuo: considerar el incidente como una oportunidad para aprender y mejorar. Implementar lecciones aprendidas en sus prácticas de manejo de datos para fortalecer la seguridad y la privacidad en el futuro.

- Auditorías regulares: establecer procesos de auditoría regulares para evaluar la eficacia de sus medidas de privacidad y anonimización en el tiempo.

# 7. Bibliografía

CBS. (2008). Manual del usuario de mu-Argus. Disponible en CBS: http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf. Recuperado el 20 de Mayo de 2018

Departamento Administrativo Nacional de Estadística (DANE) (2017). Plan Estadístico Nacional del Sistema Estadístico Nacional (SEN) 2017-022. Disponible en : https://www.dane.gov.co/files/sen/PEN-2017-2022.pdf Consultado el 28 de junio de 2018. (2017). Estrategias del Plan Estadístico Nacional.

Departamento Administrativo Nacional de Estadística (DANE) (2017). Código Nacional de Buenas Prácticas del Sistema Estadístico Nacional. Disponible en: https://www.dane.gov.co/files/sen/bp/Codigo_nal_buenas_practicas.pdf

Departamento Administrativo Nacional de Estadística (DANE) (2017). Norma Técnica de la Calidad del Proceso Estadístico. NTC PE 1000. Disponible en: http://www.dane.gov.co/files/sen/normatividad/NTC_Proceso_Estadistico.pdf. Consultado el 28 de junio de 2018.

Departamento Administrativo Nacional de Estadística (DANE). (2016). Metodología Encuesta Anual de Comercio – EAC. Disponible en: http://www.dane.gov.co/files/sen/normatividad/NTC_Proceso_Estadistico.pdf. Consultado el 28 de junio de 2018.

Departamento Nacional de Planeación (DNP) (2018). Documento Conpes 3918: Estrategia para la implementación de los objetivos de desarrollo sostenible ODS en Colombia. Disponible en: https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3918.pdf Consultado el 20 de junio de 2018. (marzo 2018).

Domingo-Ferrer J., Drechsler J. and Polettini S. (2009) Report on synthetic data files. Technical report, Deliverable of Project ESSNET-SDC

Drechsler J., Bender S. and R ässler S. (2008a) “Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel”. Transactions on Data Privacy 1(3), ): 105–130.

Departamento Administrativo Nacional de Estadística (DANE) (2018) “Guía para la anonimización de bases de datos en el Sistema Estadístico Nacional”. Disponible en: https://www.dane.gov.co/files/sen/registros-administrativos/guia-metadatos.pdf

Congreso de Colombia (2023). Ley Estadística 2335 de 2023 “Por la cual se expiden disposiciones sobre las estadísticas oficiales en el país”. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=221910

www.dane.gov.co

@DANE_Colombia

/DANEColombia

/DANEColombia

@DANEColombia