COLOMBIA POTENCIA DE LA VIDA

DANE
INFORMACIÓN PARA TODOS
70 AÑOS

Revisión de REFERENTES INTERNACIONALES

DIRECCIÓN DE REGULACIÓN, PLANEACIÓN, ESTANDARIZACIÓN Y NORMALIZACIÓN (DIRPEN)

REVISIÓN DE REFERENTES INTERNACIONALES

(1) Buenas prácticas en la aplicación de operativos que incluyan encuestas telefónicas - CATI

(2) Metodologías de anonimización para datos estructurados y no estructurados

(3) Concepto de datos no estructurados

(4) Segunda Cumbre Anual sobre el Estado de la Política de Datos Abiertos

Julio 2023

Tabla de contenido

Lista de tablas

Lista de figuras

Introducción

Este reporte tiene el propósito de apoyar el conocimiento, la generación de recomendaciones y propiciar acciones acordes a las necesidades de temáticas líderes del Departamento Administrativo Nacional de Estadísticas (DANE) y del Sistema Estadística Nacional (SEN), a partir de una revisión prospectiva que involucra referentes internacionales de diferente naturaleza y el rol en el ecosistema de datos, incluyendo oficinas nacionales de estadística, organizaciones no gubernamentales e institutos de investigación, etc.

Con ello, se busca enriquecer los trabajos que se vienen desarrollando al interior de las áreas técnicas del DANE y las instancias de coordinación del SEN, considerados prioritarios en concordancia con el Plan Estratégico Institucional y las agendas de trabajo e investigación.

Para tal fin, la revisión de referentes constituye una investigación prospectiva de la práctica internacional, en función del tema de análisis, de organizaciones mencionadas anteriormente. Los temas que se abordan en cada reporte se priorizan al considerar la urgencia de la necesidad a partir de una lista de temas construida a partir de la consulta directa realizada a los directivos DANE, los directores técnicos y los coordinadores de las mesas estadísticas del SEN. La profundidad y el detalle de las revisiones está asociada a las preguntas clave, las perspectivas, el alcance y la disponibilidad de información, si bien se pretende dar una adecuada respuesta y generar valor.

En esta versión del reporte se abordan cuatro temas: el primero hace referencia a las buenas prácticas en la aplicación de operativos que incluyan encuestas telefónicas - CATI; el segundo son las metodologías de anonimización para datos estructurados y no estructurados; el tercero abarca los conceptos de datos no estructurados, y el cuarto es una reseña sobre la segunda cumbre anual sobre el estado de la política de datos abiertos. Estos temas buscan apoyar el trabajo que viene realizando el Sistema de Ética Estadística (SETE) y el GIT de Prospectiva y Analítica de Datos.

Revisión de
REFERENTES INTERNACIONALES

1.

Buenas prácticas en la aplicación de operativos que incluyan encuestas telefónicas - CATI

1. Buenas prácticas en la aplicación de operativos que incluyan encuestas telefónicas - CATI

1.1. Resumen

La captura de información por medio telefónico es un método aplicado desde hace mucho tiempo, y ha sido empleado principalmente para la indagación de la opinión pública y estudios de mercado. Ahora bien, es conocido que desde el inicio de la emergencia derivada por la pandemia del COVID-19 del 2020, las oficinas nacionales de estadísticas (ONE) tuvieron que adaptar la recolección de información primaria para no interrumpir su producción estadística, por lo que se implementaron nuevos métodos o métodos mixtos de recolección que incluyeron, por ejemplo, encuestas telefónicas para capturar información en encuestas a hogares.

EL DANE, al igual que otras ONE de la región, vivieron un proceso acelerado de incorporación de encuestas telefónicas en la producción de estadísticas oficiales. Por esta razón y buscando la adaptación de la producción de estadísticas oficiales con nuevos métodos de captura de información con las encuestas telefónicas - CATI[1] y operativos frente a nuevos retos como los vividos con la pandemia del COVID-19, autoridades competentes en la materia como el Banco Mundial[2] o la CEPAL[3] emitieron algunas recomendaciones como:

- Definir la publicación de las estadísticas oficiales provenientes de las encuestas de hogares en la selección probabilística de la muestra y no en modelos predictivos.
- Conformar un panel de seguimiento basado en los meses de recolección presencial más recientes y realizar encuestas de forma telefónica.
- En los casos en los que las tasas de respuesta fueron bajas, se recomendó publicar las estadísticas a nivel nacional, evitando las desagregaciones.
- Utilizar modelos de corrección del sesgo de selección y cobertura que permitieran ajustar los factores de expansión teniendo en cuenta la información auxiliar disponible en el panel de seguimiento.

En general, para poder hacer un buen uso y comprender el potencial de complementariedad, subsidiariedad o sustitución que los métodos de captura de información telefónica pueden brindar a los institutos nacionales de estadística, también se entiende como recomendable considerar periodos de transición donde se sigan realizando operativos telefónicos y evaluar la posibilidad de continuar en

[1] CATI: Computer Assisted Telephone Interview.

[2] Disponible en: https://documents1.worldbank.org/curated/en/189691588696451053/pdf/Guidelines-on-CATI-Implementation.pdf

[3] Disponible en https://rtc-cea.cepal.org/sites/default/files/2020-12/ECLAC_ImpactAssessment_0.pdf

el tiempo levantamientos mixtos (presenciales y telefónicos o asistidos) de información. Por esta razón se considera de vital importancia conocer las buenas prácticas en la implementación de operativos de captura de información usando métodos telefónicos y mixtos para producción de estadísticas oficiales alrededor del mundo.

1.2. Síntesis de hallazgos

A continuación, en la Tabla 1 se presenta una breve descripción de los principales hallazgos de la revisión de referentes internacionales sobre las buenas prácticas en la aplicación de operativos que incluyan encuestas telefónicas - CATI. Para ello, se consultaron seis referentes internacionales, un referente suramericano, dos europeos, dos norteamericanos y uno oceánico.

Tabla 1. Principales hallazgos sobre buenas prácticas en la aplicación de operativos que incluyan encuestas telefónicas - CATI

Referente	¿Cuáles son las buenas prácticas que aplican los referentes en la aplicación de operativos que incluyan encuestas telefónicas - CATI?
Chile	La Subsecretaría de Telecomunicaciones de Chile desarrolló junto a la empresa de investigación CADEM la Encuesta de Satisfacción de Usuarios y Medición del Nivel de Calidad de Servicios de Telecomunicaciones, donde utilizó el Sistema de Entrevistas Telefónicas Asistidas por Computador (CATI). La información fue recolectada por medio de la aplicación de un formulario telefónico diseñado específicamente para identificar la satisfacción de los usuarios frente a la prestación de servicios de telecomunicaciones.
Canadá	Statistics Canada con las evaluaciones de impacto de privacidad, analiza y evalúa los riesgos de privacidad, confidencialidad o seguridad asociados con la recopilación, el uso o la divulgación de información personal. Como parte de estas evaluaciones se utilizó el programa de monitoreo de entrevistas personales asistidas por computadora (CAPI) y la entrevista telefónica asistida por computadora - CATI.
Finlandia	En Finlandia se utilizan las entrevistas telefónicas asistidas por computador - CATI en la recolección de datos de las encuestas del uso del tiempo, condiciones de vida, fuerza laboral, distribución del ingreso y confianza del consumidor[4].
Estados Unidos	La Oficina de Censos y el Departamento de Salud de Alaska han desarrollado encuestas en las que han utilizado la vía telefónica para la captura de los datos. Una de ellas ha sido realizada en primera instancia de manera presencial seguida de una parte telefónica. La otra encuesta se desarrolla únicamente por teléfono y han utilizado estrategias para la elección de las personas que serán entrevistadas.

[4] Disponible en: Search (stat.fi)

Referente	¿Cuáles son las buenas prácticas que aplican los referentes en la aplicación de operativos que incluyan encuestas telefónicas - CATI?
España	El INE de España realiza la Encuesta Continua de Hogares que es una operación estadística con un sistema de recolección multicanal soportado por una aplicación informática, por lo que incorpora controles de rango, flujo, completitud y validez que están en funcionamiento durante toda la recolección de la información.
Australia	La Oficina de Estadísticas de Australia (ABS) emplea CATI en algunas de sus operaciones estadísticas, tanto para encuestas de hogares como de empresas. Este método permite entrevistas telefónicas con respuestas ingresadas directamente en una computadora, lo que conlleva ventajas como costos reducidos, edición inmediata, programación de llamadas y monitoreo del personal entrevistador. La ABS presenta un modelo de idoneidad que considera trece factores al evaluar la elección del modo de recolección de datos.

Fuente: DANE a partir de las revisiones de referentes.

1.3. Revisión de referentes

En esta sección se presentará de forma sintetizada la revisión de referentes internacionales.

1.3.1. Chile

La Subsecretaría de Telecomunicaciones de Chile[5] desarrolló junto a la empresa de investigación CADEM la Encuesta de Satisfacción de Usuarios y Medición del Nivel de Calidad de Servicios de Telecomunicaciones[6]. Este proyecto fue diseñado con el objetivo de medir de forma cuantitativa el grado de satisfacción de los usuarios con los servicios de telefonía móvil, internet móvil, televisión e internet residencial, por cada una de las compañías proveedoras que prestan sus servicios en Chile y cuya muestra es estadísticamente confiable y representativa a nivel nacional y regional y con la utilización de un sistema de entrevistas telefónicas asistidas por computador - CATI.

Para lograr la aplicación de la encuesta por medio del sistema CATI fue necesario el desarrollo de un cuestionario que incluyera:

- Una introducción, un texto estándar donde el encuestador se presenta, pone al usuario en contexto sobre el motivo de la llamada y pide autorización para aplicar el formulario.
- Filtro, donde el encuestador identifica que el usuario es mayor de edad.

[5] Disponible en: https://www.subtel.gob.cl/?s=CATI+

[6] Disponible en: https://www.subtel.gob.cl/wp-content/uploads/2017/03/estudios_satisfaccion_usuarios/2016/Primera_Medicion/Informe_I_2016.pdf

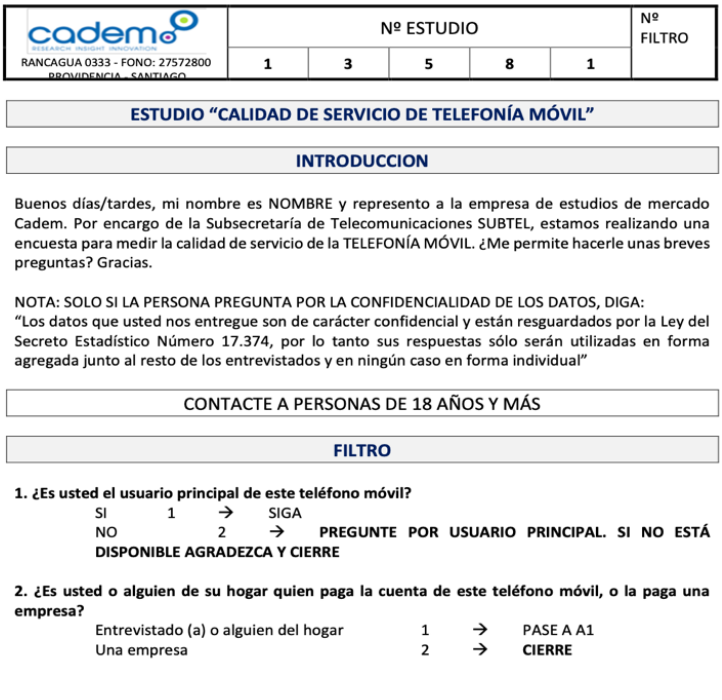
- Datos del proveedor, donde el encuestador realiza preguntas sobre el prestador de servicios de telecomunicaciones, como nombre del prestador de servicios, tiempo de prestación de los servicios de telecomunicaciones, satisfacción con el servicio, etc.
- Problemas recientes con el servicio, el encuestador realizará preguntas acerca de los problemas o fallas que los servicios presentan.
- Derechos del consumidor de telecomunicaciones, el encuestador presenta al usuario consumidor de servicios de telecomunicaciones sus derechos, consulta las empresas con las que se siente más protegido, etc.

A continuación, se presenta un listado con las actividades desarrolladas en el proceso de recolección de información por medio del sistema CATI:

- Diseño de muestra, cuestionarios y establecimiento de plazos de entrega de productos.
- Revisión de cuestionarios y diseño muestral.
- Programación del cuestionario en CATI.
- Reclutamiento de encuestadores.
- Capacitación de manejo del aplicativo.
- Validación del instrumento piloto.
- Preparación de un informe piloto.
- Planificación de campo y plan de análisis de primera encuesta.
- Entrega de resultados.
- Revisión de bases de datos del marco muestral por servicio.
- Programación del cuestionario final en CATI.
- Aplicación de la encuesta.
- Supervisión del proceso de recolección de información.
- Codificación, procesamiento, validación y generación de base de datos.
- Análisis y elaboración de informe.
- Presentación de informe preliminar.
- Revisión de comentarios.
- Presentación de informe y resultados finales.

A continuación, se presentan dos ejemplos del formato de cuestionario implementado para el levantamiento y la captura de la información de satisfacción de los servicios de telecomunicaciones.

Figura 1. Formulario CATI - introducción



cadem
RESEARCH INSIGHT INNOVATION
RANCAGUA 0333 - FONO: 27572800
PROVIDENCIA - SANTIAGO

Nº ESTUDIO					Nº FILTRO
1	3	5	8	1	

ESTUDIO "CALIDAD DE SERVICIO DE TELEFONÍA MÓVIL"

INTRODUCCION

Buenos días/tardes, mi nombre es NOMBRE y represento a la empresa de estudios de mercado Cadem. Por encargo de la Subsecretaría de Telecomunicaciones SUBTEL, estamos realizando una encuesta para medir la calidad de servicio de la TELEFONÍA MÓVIL. ¿Me permite hacerle unas breves preguntas? Gracias.

NOTA: SOLO SI LA PERSONA PREGUNTA POR LA CONFIDENCIALIDAD DE LOS DATOS, DIGA:
"Los datos que usted nos entregue son de carácter confidencial y están resguardados por la Ley del Secreto Estadístico Número 17.374, por lo tanto sus respuestas sólo serán utilizadas en forma agregada junto al resto de los entrevistados y en ningún caso en forma individual"

CONTACTE A PERSONAS DE 18 AÑOS Y MÁS

FILTRO

1. ¿Es usted el usuario principal de este teléfono móvil?

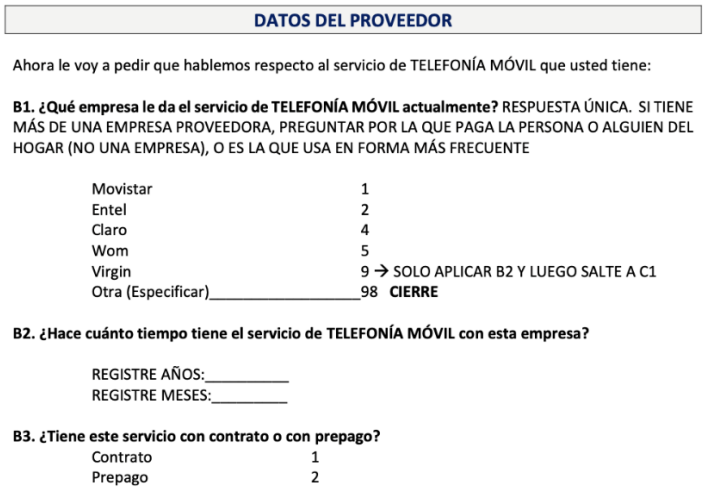
SI 1 → SIGA
NO 2 → **PREGUNTE POR USUARIO PRINCIPAL. SI NO ESTÁ DISPONIBLE AGRADEZCA Y CIERRE**

2. ¿Es usted o alguien de su hogar quien paga la cuenta de este teléfono móvil, o la paga una empresa?

Entrevistado (a) o alguien del hogar 1 → PASE A A1
Una empresa 2 → **CIERRE**

Fuente: DANE a partir de Encuesta de Satisfacción de Usuarios y Medición del Nivel de Calidad de Servicios de Telecomunicaciones.

Figura 2. Formulario CATI – datos del proveedor



DATOS DEL PROVEEDOR

Ahora le voy a pedir que hablemos respecto al servicio de TELEFONÍA MÓVIL que usted tiene:

B1. ¿Qué empresa le da el servicio de TELEFONÍA MÓVIL actualmente? RESPUESTA ÚNICA. SI TIENE MÁS DE UNA EMPRESA PROVEEDORA, PREGUNTAR POR LA QUE PAGA LA PERSONA O ALGUIEN DEL HOGAR (NO UNA EMPRESA), O ES LA QUE USA EN FORMA MÁS FRECUENTE

Movistar	1
Entel	2
Claro	4
Wom	5
Virgin	9 → SOLO APLICAR B2 Y LUEGO SALTE A C1
Otra (Especificar)_____	98 **CIERRE**

B2. ¿Hace cuánto tiempo tiene el servicio de TELEFONÍA MÓVIL con esta empresa?

REGISTRE AÑOS:_____
REGISTRE MESES:_____

B3. ¿Tiene este servicio con contrato o con prepago?

Contrato	1
Prepago	2

Fuente: DANE a partir de Encuesta de Satisfacción de Usuarios y Medición del Nivel de Calidad de Servicios de Telecomunicaciones.

1.3.2. Canadá

Statistics Canada ha establecido evaluaciones de impacto en la privacidad[7] que analizan los riesgos de privacidad, confidencialidad o seguridad asociados con la recopilación, el uso o la divulgación de información personal y desarrollar medidas destinadas a mitigar y, cuando sea posible, eliminar los riesgos identificados.

Como parte de estas evaluaciones se presenta el Programa de monitoreo de entrevistas personales asistidas por computadora - CAPI: resumen de la evaluación del impacto en la privacidad[8]. Este es un nuevo programa de seguimiento de entrevistas de encuestas de hogares realizadas por entrevistadores de campo y con este programa se espera:

- Facilitar la evaluación de la calidad de los datos de las encuestas realizadas en los hogares.
- Proporcionar un medio para identificar mejor las necesidades de capacitación de los entrevistadores de campo.
- Identificar más fácilmente problemas potenciales relacionados con la herramienta de recopilación o el cuestionario.

El Programa de Monitoreo CAPI emplea un programa de computadora que realiza una grabación de audio de las entrevistas a medida que el entrevistador ingresa las respuestas del encuestado en su computadora portátil. La grabación sirve como medio para que la entrevista sea evaluada después de que se esté realizando, con el fin de determinar si existen problemas de privacidad, confidencialidad y seguridad asociados con el programa y, de ser así, hacer recomendaciones para su resolución o mitigación

En el caso de la entrevista telefónica asistida por computadora - CATI[9], se presenta la metodología y los desafíos que se pueden encontrar específicamente en las encuestas de salud en la vigilancia de la salud pública. Por esto, se muestran los méritos y las limitaciones en el uso del CATI en la Tabla 2, ya que permite la entrada directa de datos en un formato electrónico (reduciendo el tiempo y los costos de procesamiento) y, por lo tanto, la entrega rápida de datos.

---

[7] Disponible en: https://www.statcan.gc.ca/en/about/pia/pia

[8] Disponible en: https://www.statcan.gc.ca/en/about/pia/capi

[9] Disponible en: https://www.canada.ca/en/public-health/services/reports-publications/health-promotion-chronic-disease-prevention-canada-research-policy-practice/vol-25-no-2-2004/computer-assisted-telephone-interviewing-cati-health-surveys-public-health-surveillance.html#tab5

Tabla 2. Méritos y limitaciones de la entrevista telefónica asistida por computadora - CATI

Tipo	Méritos	Limitaciones
Entrevista	1. Manera eficiente de obtener información. 2. Única forma de obtener opinión e información sobre actitudes. 3. Puede dirigirse a subgrupos específicos de la población.	1. El autoinforme presenta problemas. 2. La capacidad de las personas para comprender las preguntas. 3. La capacidad de identificar su estado. 4. Voluntad de informar sobre su estado. 5. Necesidad de validar los datos, por ejemplo, mediante la búsqueda de registros.
Teléfono	1. Económico. 2. Proporciona muestras grandes. 3. Proporciona resultados instantáneos. 4. La marcación aleatoria de dígitos puede proporcionar una muestra razonablemente aleatoria. 5. Facilita preguntas delicadas.	1. No aplica si el porcentaje de hogares con teléfono es bajo. 2. Tendencia a entrevistar a personas que se quedan en casa. 3. La marcación aleatoria de dígitos puede llegar a números no válidos. 4. Problema de capacidad de atención. 5. No se pueden mostrar ayudas visuales. 6. La gente está cansada de las encuestas telefónicas de mercado.
Asistido por computadora	1. Se puede vincular con un SMS. 2. Ayuda a los entrevistadores (instrucciones en pantalla). 3. Facilita la recopilación de datos (preguntas aleatorias, ediciones y comprobaciones de coherencia, omisión de preguntas). 4. Facilita la entrada de datos (entrada directa, codificación programada).	1. Gran costo inicial. 2. Requiere software específico. 3. Más tiempo para desarrollar y probar el cuestionario. 4. Más tiempo para capacitar a los entrevistadores. 5. Posibles errores de entrada de datos (papeles de mecanografiar).

Fuente: Statistics Canada.

1.3.3. Estados Unidos

Encuesta de población actual

La Oficina de Censos de los Estados Unidos muestra en la metodología de la Encuesta de Población Actual[10] (CPS, por sus siglas en ingles), su desarrollo, el cual incluye representantes de campo (FR) y

[10] Disponible en: https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/collecting-data.html

entrevistadores telefónicos asistidos por computador - CATI. Lo primero que realizan es el envío de una carta describe la CPS, anuncia la próxima visita y brinda a los encuestados información sobre sus derechos bajo la Ley de Privacidad, la naturaleza voluntaria de la encuesta y las garantías de confidencialidad de la información que brindan. Después de eso se pasa a una fase de clasificación de hogares, para dar paso a las entrevistas iniciales donde se recopila información sobre algunas características adicionales después de completar la parte de la entrevista sobre la fuerza laboral.

Los hogares que siguen siendo parte de la muestra en el segundo, el tercer y el cuarto mes, tienen la opción de realizar la entrevista por teléfono. El uso de este modo de entrevista debe ser aprobado por el encuestado. Dicha aprobación se obtiene al final de la entrevista del primer mes, después de completar la fuerza laboral y cualquier pregunta complementaria. Las entrevistas telefónicas son el método preferido para recopilar los datos; es mucho más eficiente en tiempo y costo. Por lo general obtienen el 85% de las entrevistas en los tres meses de muestra (MIS, por sus siglas en ingles). Los FR deben intentar realizar una entrevista de visita personal para la entrevista del quinto mes. Después de un intento se puede realizar una entrevista telefónica siempre que el hogar original todavía ocupe la unidad de muestra.

El uso de las entrevistas telefónicas asistidas por computador - CATI ha sido un hito en la informatización de la CPS quien la ha usado desde 1983. El primer caso fue la prueba Tri-Cities en 1987, desde allí, han enviado varios casos que actualmente son 8000 casos por mes. Una de las razones principales para usar CATI es facilitar el esfuerzo de reclutamiento y contratación en áreas difíciles de enumerar. Es más fácil contratar a una persona que trabaje en las instalaciones de CATI que contratar a personas para que trabajen como FR. Los FR son más provechosos en las zonas rurales porque deben atender a menos personas y quienes trabajan en los CATI trabajan con las grandes ciudades por el volumen de personas a atender.

Sistema de Vigilancia de Factores de Riesgo del Comportamiento de Alaska

El Departamento de Salud de Alaska desarrolla una encuesta denominada Sistema de Vigilancia de Factores de Riesgo del Comportamiento de Alaska[11] (BRFSS, por sus siglas en ingles), la cual anualmente recolecta información de los adultos sobre la salud y los comportamientos que podrían afectarla. La metodología[12] que emplean para desarrollar la encuesta es recopilar los datos de entrevistas telefónicas mediante marcado aleatorio de dígitos (RDD). Los números de teléfono se muestrean utilizando un diseño de muestreo estratificado (DSS), teniendo en cuenta las siete regiones de salud pública de Alaska.

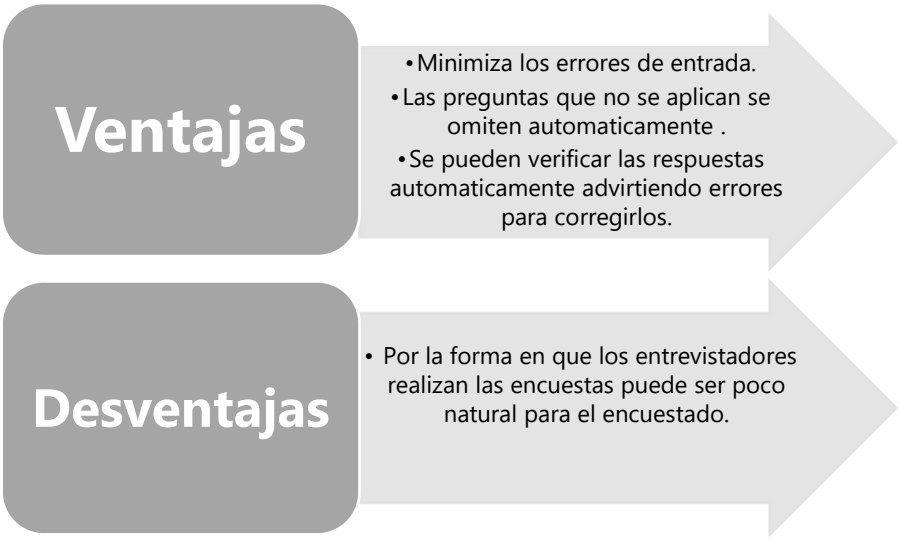
[11] Disponible en: https://health.alaska.gov/dph/Chronic/Pages/brfss/default.aspx

[12] Disponible en: https://health.alaska.gov/dph/chronic/pages/brfss/method.aspx

Los números telefónicos se eligen de bloques de números potenciales de un área, diferenciados en listas de números fijos y de celulares. Estas llamadas las realizan toda la semana, de día y de noche. Cuando marcan a un número fijo y contestan, eligen a uno de los adultos de la casa al azar para que realice la encuesta. Al llamar a los números de celular, se entiende que quien contesta es el usuario principal de la línea por lo que aplican la encuesta directamente. Los participantes se eligen al azar para garantizar que los datos recopilados puedan generalizarse al estado en su conjunto. El número de entrevistas que realizan varía y depende del presupuesto que tengan disponible.

Alaska tiene un contrato con ICF Macro, Inc. para recopilar datos BRFSS procurando capturar la misma cantidad de entrevistas cada mes para reducir el sesgo causado por la variación estacional en los comportamientos y las condiciones de salud. ICF utiliza un software de entrevista telefónica asistida por computador - CATI, el cual muestra el cuestionario en la pantalla durante la entrevista y el entrevistador ingresa las respuestas directamente en el computador.

Figura 3. Ventajas y desventajas del uso de CATI



Fuente: DANE a partir de información del Departamento de Salud de Alaska.

1.3.4. España

El INE de España realiza la Encuesta Continua de Hogares que es una operación estadística que ofrece información anual sobre las características demográficas básicas de la población, los hogares que componen y las viviendas que habitan. La información se ofrece desagregada por comunidades autónomas y provincias. Sobre la población, facilita datos por sexo, edad, estado civil, país de nacimiento, nacionalidad, situación en el hogar. Para los hogares aporta información sobre su tamaño y composición y para las viviendas sobre su régimen de tenencia, superficie útil, habitaciones, año de edificación y tipología del edificio.

La encuesta tiene un único cuestionario estructurado en tres apartados: identificación de los posibles miembros del hogar, cuestionario de vivienda y cuestionarios individuales. Se determina si en la dirección seleccionada hay alguien residiendo de forma habitual, en cuyo caso se trata de una vivienda principal y se incluyen los datos correspondientes a las características de la vivienda y los correspondientes a las características de cada persona residente[13].

Existen tres formatos diferentes del cuestionario ya que la recolección de información de esta encuesta es multicanal. Los formatos de este cuestionario son:

- CAWI: recolección a través de Internet y autodiligenciamiento del cuestionario por el propio informante.
- CATI/CAPI: recolección de la información a través de entrevistador, por teléfono o presencial en el domicilio del informante mediante ordenador portátil tipo tableta.
- Papel: el informante diligencia el cuestionario que se envía previamente por correo certificado y se devuelve por el mismo medio.

Los formatos CAWI, CATI y CAPI son un cuestionario electrónico soportado por una aplicación informática, por lo que incorpora controles de rango, flujo, completitud y validez que están en funcionamiento durante toda la recolección de la información, mientras que los cuestionarios recibidos en papel deben pasarse a la aplicación y es entonces cuando se realizan los controles.

No obstante, una unidad puede iniciar el diligenciamiento de la información por un canal y finalizarla por otro diferente; por ejemplo, un cuestionario recibido en papel que no cumpla los requisitos establecidos en los controles estipulados puede ser derivado a recolección telefónica - CATI, si se dispone de teléfono para esa unidad o recolección presencial a domicilio con tableta - CAPI en caso contrario.

La recolección de la información se realiza trimestralmente de acuerdo con un calendario prefijado que contempla las peculiaridades de un sistema de recolección multicanal secuencial de la siguiente manera:

- Envió de cartas a las unidades seleccionadas para que puedan diligenciar la información por internet - CAWI: la información se recolecta a través de un cuestionario electrónico disponible en www.iria.ine.es, al que tienen acceso todas las unidades seleccionadas en la muestra para poder acceder al mismo. En las cartas que se envían a las unidades seleccionadas se incluyen las claves necesarias para el diligenciamiento electrónico y es el propio informante el que diligencia el cuestionario; este canal se encuentra disponible durante todo el periodo de recolección de información de la encuesta.

[13] Disponible en: https://www.ine.es/inebaseDYN/ech30274/docs/metodologia_ech.pdf

- Una vez diligenciado el cuestionario, la misma aplicación web determina la completitud y la validez de este y pasa la información a la base central.
- También se realiza la apertura del teléfono de consulta gratuita, que incluye la posibilidad de contestar el cuestionario vía telefónica.
- Envío de carta de reclamación a las unidades muestrales que no hayan respondido al cuestionario por internet transcurrido un plazo sin respuesta.
- Envío de nueva reclamación con cuestionario en papel para su devolución por correo certificado.
- Entrevistas telefónicas - CATI para unidades pendientes de respuesta para las que se dispone de teléfono: es un entrevistador-grabador quien realiza la entrevista telefónica diligenciando un cuestionario electrónico disponible en la aplicación ARZ, el procedimiento es controlado en todo momento por la aplicación y es el siguiente: cuando la unidad pasa a la fase CATI se procede a realizar el proceso de contacto con la unidad desde el centro telefónico correspondiente a la zona. El entrevistador llama a la unidad que seleccione la aplicación o a uno de los que tenga asignados; esa llamada puede ser contestada o puede ocurrir que salte un fax o un contestador, etc. Para cada llamada el entrevistador o la aplicación asigna un resultado de llamada.

  La aplicación genera llamadas sucesivas en diferentes momentos hasta conseguir el contacto. Una vez que se ha contactado con alguien, el entrevistador debe comprobar que la unidad a la que se está llamando es la dirección seleccionada, después debe localizar un informante adecuado y por último realizar la entrevista. Según las situaciones que se presenten, el entrevistador o la aplicación asignarán una incidencia la unidad. Además, la aplicación en cada momento, dependiendo de los resultados de las llamadas a la unidad, en su caso, de las incidencias producidas, asignará un resultado temporal o final a cada unidad.
- Envío de una carta anunciándoles la próxima visita a su domicilio de un agente entrevistador: el día que se realiza la visita se recoge la información mediante entrevista personal CAPI con dispositivo portátil para unidades pendientes que no disponen de teléfono o no se contactan. Se realiza entrevista personal en las unidades en las que no se haya conseguido ninguna información en las fases anteriores. La recolección por este método se efectúa en dos momentos: simultáneamente a la recolección CATI, se investigan por CAPI las unidades de los que no se haya obtenido teléfono, y al finalizar el trimestre se visitan todas las unidades pendientes de investigar. Las unidades muestrales que hayan diligenciado el cuestionario por Internet o por papel y no esté completo o presente inconsistencias, son llamadas (si tienen teléfono), o visitadas (si no lo tienen), para completar o depurar el cuestionario mediante CATI o CAPI, respectivamente. Durante toda la fase de recolección está a disposición de los informantes un número de teléfono gratuito.

1.3.5. Australia

La ABS (Australian Bureau of Statistics, Oficina de Estadísticas de Australia) para algunas de sus operaciones estadísticas maneja las entrevistas telefónicas y se utilizan tanto en las encuestas de hogares como de empresas de ABS y pueden utilizarse junto con entrevistas cara a cara. Por ejemplo, en la LFS (Labour Force Survey, Encuesta de Población Activa), la primera entrevista generalmente se realiza cara a cara y las entrevistas restantes se realizan por teléfono[14]. Para las encuestas empresariales, el uso de CATI está centralizado mientras que se encuentra descentralizado para encuestas de hogares.

Las CATI implican que el entrevistador ingresa las respuestas directamente en una computadora a medida que se hacen las preguntas de la encuesta a los proveedores por teléfono. Esta técnica permite:

- Costos reducidos en comparación con las entrevistas cara a cara, ya que se necesitan menos entrevistadores y no hay costos de viaje involucrados.
- Entrevistas telefónicas que potencialmente produzcan resultados más oportunos.
- Parte de la edición debe realizarse de inmediato (lo que mejora la calidad de los datos y reduce el tiempo de procesamiento), las ediciones y las verificaciones automáticas ayudan a reducir los errores de ingreso de datos y alertan al entrevistador sobre respuestas incoherentes o poco probables, lo que mejora la calidad de los datos y reduce el tiempo de procesamiento.
- Programación de llamadas para llevar a cabo. Se puede llamar a los encuestados en momentos convenientes o cuando los datos estén disponibles. Además, si el teléfono está ocupado, el sistema reprogramará la llamada y los seguimientos para obtener información adicional son relativamente rápidos y económicos.
- Las preguntas se secuenciarán de modo que solo las preguntas relevantes sean visibles para el entrevistador. La secuenciación puede integrarse automáticamente en el sistema, lo que significa que los entrevistadores no tienen que seguir manualmente instrucciones de secuenciación complejas, reduciendo así los errores del entrevistador.
- Monitoreo del personal entrevistador para que la consistencia del desempeño sea mayor.
- Los datos se ingresan durante la propia entrevista, por lo que no hay una fase separada de ingreso de datos (ahorro de tiempo y recursos).

Al igual que con otros métodos de recopilación de datos, existen algunos inconvenientes asociados con este enfoque. Hay límites en el número y la complejidad de las preguntas que se pueden hacer y debido a la facilidad con la que el encuestado puede terminar la entrevista, tanto la falta de respuesta

[14] Disponible en: https://www.abs.gov.au/statistics/detailed-methodology-information/concepts-sources-methods/labour-statistics-concepts-sources-and-methods/2021/methods-four-pillars-labour-statistics/household-surveys

como la falta de respuesta parcial puede ser mayor que en las entrevistas cara a cara. Entre las desventajas se encuentra:

- Los costos de instalación de la interfaz pueden ser altos.
- Los entrevistadores deben tener habilidades informáticas y estar capacitados para utilizar correctamente el sistema informático.
- En comparación con otras entrevistas telefónicas, el tiempo de la entrevista aumenta porque la edición se realiza mientras se realiza la entrevista; esto puede aumentar la falta de respuesta.

La ABS presenta un modelo de idoneidad para determinar el modo principal de recolección. Este modelo analiza algunos factores para comparar y evaluar cada uno de estos modos de recopilación, según sus características específicas y los objetivos de recopilación de datos. Estos factores son:

1. **Frecuencia:** si la encuesta se realiza con frecuencia puede ser más beneficioso optar por modos más eficientes en términos de costos.
2. **Tamaño de la muestra:** si se está considerando utilizar un modo de alto costo de instalación, como IVR o CATI, las encuestas con tamaños de muestra más grandes suelen ser más convenientes desde el punto de vista del costo-beneficio. Sin embargo, para otros modos, como las encuestas por correo, el tamaño de la muestra tiene un impacto menor en la elección del modo.
3. **Capacidad de generalización de la encuesta:** se analiza cuánto contenido de la encuesta coincide con otras encuestas, lo que puede facilitar la adaptación del modo de recopilación y así ahorrar tiempo de desarrollo. Se excluyen las preguntas de cobertura y periodo de referencia que son comunes a casi todas las encuestas.
4. **Tiempo empleado/duración de la entrevista:** se considera el tiempo estimado necesario para completar la encuesta, ya sea para los encuestados o para los entrevistadores. En particular, se debe tener en cuenta la duración de la entrevista en las encuestas telefónicas y por fax, ya que no son adecuadas para encuestas largas debido a problemas de carga del proveedor o del encuestado.
5. **Número total de elementos de datos:** este factor se agrega al tiempo empleado como medida de la carga del proveedor. Los "elementos de datos" se interpretan aproximadamente como una respuesta individual o una decisión que el encuestado debe tomar. Las encuestas telefónicas y por fax son menos adecuadas para formularios con una gran cantidad de elementos de datos debido a la mayor carga que impone a los encuestados.
6. **Cantidad de secuencias:** se evalúa la cantidad de secuencias en el formulario, es decir, la obligación de seguir instrucciones de sucesión. Los formularios con muchas secuencias se manejan mejor electrónicamente dado que la secuencia puede automatizarse.
7. **Proporción de elementos que requieren verificación de registros:** se distingue entre preguntas que pueden responderse de forma improvisada y preguntas que requieren

verificación de registros o aportes de otras personas. Las preguntas que requieren datos numéricos son menos adecuadas para encuestas telefónicas, que deben ser lo más breves posible.

8. **Inquietudes sobre sensibilidad/confidencialidad:** se considera la disposición de los encuestados para proporcionar respuestas honestas a preguntas delicadas o confidenciales. Las encuestas telefónicas son el peor método para recopilar datos confidenciales, mientras que las encuestas por correo autoadministradas son las mejores.
9. **Alfabetización de los encuestados:** se tiene en cuenta la alfabetización de los encuestados en formularios autoadministrados. Los formularios dirigidos a la población en general deben estar en un nivel de lectura de quinto grado.
10. **Ubicación geográfica de los encuestados:** se considera la ubicación de la mayoría de los encuestados y la dificultad para comunicarse con ellos. Las encuestas por correo son más adecuadas para llegar a encuestados en diversas áreas geográficas.
11. **Presupuesto:** las encuestas administradas por el entrevistador son más costosas que las autoadministradas. Las entrevistas cara a cara son las más caras, seguidas de las entrevistas telefónicas. Las encuestas por correo son generalmente las menos costosas.

Tabla 3. Calificaciones de idoneidad CATI

Calificaciones de idoneidad CATI		
Frecuencia de la encuesta	**Mensual**	**Trimestral**
Tamaño de la muestra	10 mil más.	Menos de 10 mil a 5 mil.
Capacidad de generalización de la encuesta	Todas las preguntas comunes a muchas otras encuestas.	La mayoría de las preguntas comunes a muchas otras encuestas.
Tiempo empleado/duración de la entrevista	5 minutos o menos.	6 - 10 minutos.
Complejidad de las preguntas	Menos del 20%.	20% a menos del 40%.
Cantidad de secuencia	Cualquier nivel, bajo, moderado, alto o ninguno.	
Proporción de elementos que requieren verificación de registros	Ninguno.	Bajo.
Inquietudes sobre sensibilidad/confidencialidad	Muy bajo.	Bajo.
Alfabetización de los encuestados	85% - 100%.	75% o menos.
Ubicación geográfica de los encuestados	Metropolitano (interior y exterior), con cobertura de teléfono/internet/fax.	
Presupuesto	Grande.	

Fuente: DANE, basado en ABS[15].

La idoneidad del uso de CATI se basa en varios factores, ya que las encuestas que se realizan con frecuencia y tienen muestras grandes son más adecuadas para la conversión a CATI desde una perspectiva de costo-beneficio. Además, aquellas encuestas que comparten muchas preguntas con otras (por ejemplo, usando redacción estándar) son más apropiadas para CATI ya que el instrumento puede adaptarse fácilmente a futuras encuestas. Por tanto, la elección de CATI se recomienda para maximizar la eficiencia y el control de costos en dichas situaciones.

La ABS recomienda el uso de CATI centralizada debido a sus beneficios en términos de procedimientos de entrevista estandarizados, un mejor control de calidad y facilidades para el reclutamiento, la capacitación, la supervisión y el seguimiento de los entrevistadores. Esto ayuda a evitar variaciones en las tasas de respuesta y las estimaciones causadas por diferentes enfoques, además de permitir un acceso más efectivo al soporte técnico para los sistemas CATI utilizados en las encuestas[16].

1.3.6. Finlandia

Los datos de la Encuesta de Uso del Tiempo realizada de 2020 a 2021 se recopilaron con entrevistas telefónicas asistidas por computadora - CATI y con diarios web y en papel. Los datos sobre el empleo, el estudio, el trabajo voluntario y los pasatiempos se investigaron en la parte de la entrevista de la encuesta. En las encuestas anteriores sobre el empleo del tiempo, en 1979, de 1987 a 1988 y de 1999 a 2000, los datos se recogieron durante entrevistas cara a cara, sin embargo, en la encuesta de 2009 a 2010, algunas de las entrevistas se llevaron a cabo como entrevistas telefónicas, con el interés de reducir costos. En la encuesta de 2020 a 2021 todas las entrevistas se realizaron como entrevistas telefónicas.

El método principal de recolección para los datos recogidos con las entrevistas de las estadísticas sobre condiciones de vida hasta el año de referencia 2021 es la entrevista telefónica asistida por computador - CATI.

Los datos de la Encuesta de Población Activa se recopilan mediante entrevistas telefónicas asistidas por computador, realizadas por los entrevistadores de Statistics Finland y en parte por entrevistas cara a cara y a partir de 2021 con un cuestionario web.

El método principal de recolección de datos para los datos recolectados con las entrevistas de las estadísticas sobre condiciones de vida (distribución del ingreso) es una entrevista telefónica asistida

[15] Disponible en: https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/7D4A19CDFFFC83E4CA2576B300126A8D?opendocument

[16] Disponible en: https://www.abs.gov.au/ausstats/abs@.nsf/Products/1530.0~2010~Chapter~Mode+Suitability+Framework?OpenDocument#CATI

por computadora - CATI administrada por un entrevistador. Solo una pequeña parte de las entrevistas (alrededor del uno al dos por ciento) se recopila con una entrevista personal asistida por computadora - CAPI.

1.4. Conclusiones

A partir de la revisión de referentes internacionales se concluye lo siguiente:

- Es necesario la implementación de tecnología ya sea a través de aplicaciones o software, dentro del proceso de recolección de información por vía telefónica, con la finalidad de realizar el seguimiento oportuno a las fuentes y contar con información oportuna, coherente y de calidad para los procesos de producción de operaciones estadísticas.
- El uso de este tipo de encuestas permite la corrección rápida por digitación o mal entendimiento de las respuestas, lo que aumenta la confiabilidad en los resultados presentados. Asimismo, acorta los desplazamientos para las personas que hacen las encuestas en campo y el tiempo en la captura de datos, especialmente en ciudades muy grandes, donde la muestra es considerablemente diferente a la de un espacio rural.
- A pesar de que la información acerca de la metodología de las CATI en Finlandia y su uso es limitado, este tipo de encuestas telefónicas ha aumentado con el tiempo debido a la motivación de reducción de costos. Aunque aún se realizan entrevistas cara a cara o con la asistencia de formularios web, estas representan un pequeño porcentaje de las entrevistas totales de las encuestas, siendo las CATI el principal método para llevar a cabo las entrevistas.
- En Nueva Zelanda se implementan las CATI en algunas encuestas, sin embargo, la información respecto a la metodología o motivaciones para su uso es escasa.

1.5. Recomendaciones

Se recomienda a nivel general en el diseño y el desarrollo de las operaciones estadísticas:

1. Se recomienda la implementación de software o aplicaciones que puedan medir, seguir, controlar y sistematizar la recolección de la información por vía telefónica. Por ejemplo, una aplicación que mida los tiempos de las llamadas, autogestione las llamadas con las unidades que están por consultar y donde el entrevistador tenga un campo para incluir la información y las observaciones de la conversación o aplicaciones que trascriban lo que se habla en las llamadas.
2. Se recomienda que quienes realicen las encuestas tengan una capacitación adecuada para la lectura de las preguntas, ya que el tono y la forma en la que las apliquen puede afectar la confianza del encuestado.

3. Se recomienda fortalecer los operativos telefónicos adelantados desde la DRA con el objetivo de garantizar la confiabilidad de la información recolectada y reducir los tiempos de recolección establecidos actualmente.

Revisión de
REFERENTES
INTERNACIONALES

2.

Metodologías de anonimización de datos estructurados y no estructurados

2. Metodologías de anonimización de datos estructurados y no estructurados

2.1. Resumen

El SEN busca promover entre sus miembros el acceso y el uso de los microdatos para la producción y la difusión de estadísticas oficiales, como se encuentra consignado en el Decreto 1743 de 2016 en el artículo 2.2.3.1.2. Además, el Código de Buenas Prácticas Estadísticas del SEN en el principio 10 incentiva a los miembros del SEN a implementar prácticas que permitan el acceso de las estadísticas y los microdatos asociados a todo tipo de usuarios y en el principio 11, sobre la confidencialidad, insta a los productores de información estadística a utilizar técnicas para la anonimización de microdatos para así garantizar la confidencialidad de la información de las fuentes empleadas en el proceso estadístico.

Tomando estas premisas, en 2018 el DANE desarrolló la Guía para la Anonimización de Bases de Datos en el SEN, cuyo propósito es orientar a los integrantes del SEN en el proceso de anonimización de bases de datos que provienen de registros administrativos y operaciones estadísticas.

En el marco del proyecto de infraestructura de datos liderado por MINTIC y donde el DANE está involucrado para su desarrollo, se requiere diseñar una solución común sobre el proceso de anonimización de conjuntos de datos que pueda aportar los lineamientos básicos sobre este proceso a las entidades que forman parte del Sistema Estadístico Nacional de una forma eficiente y actualizada. Para ello, es necesario conocer los conceptos relacionados con la temática en mención que permitan al usuario la comprensión de los elementos técnicos y faciliten la implementación de las herramientas que serán dispuestas para la anonimización efectiva y confiable de textos y conjuntos de datos.

Para cumplir este objetivo se tomó como punto de partida la Guía de Anonimización elaborada en 2018 por el DANE y los referentes nacionales e internacionales que aportan información sobre este proceso. La nueva Guía se actualizará con el trabajo articulado que se está adelantando entre el Archivo General de la Nación (AGN) y el DANE.

2.2. Síntesis de hallazgos

A continuación, la Tabla 4 presenta una breve descripción de los principales hallazgos de la revisión de referentes internacionales sobre metodologías de anonimización de datos estructurados y no estructurados. Para esta versión se consultaron ocho referentes, un referente nacional, uno sudamericano, uno oceánico, dos norteamericanos y tres europeos.

Tabla 4. Principales hallazgos sobre metodologías de anonimización para datos estructurados y no estructurados

Referente	¿Cuáles son las metodologías de anonimización que aplican los referentes para datos estructurados y no estructurados?
Colombia	El Centro Nacional de Memoria Histórica desarrolló la Guía para la Anonimización de Datos e Información No Estructurada, como una guía para el desarrollo de procesos de anonimización de información no estructurada como imágenes, videos, audio, información de redes sociales, entre otros. En la guía se presenta de una forma concreta cuáles son las actividades y los procesos que debe adelantar el usuario para conservar la confidencialidad de la información, en términos de almacenamiento, eliminación, modificación o duplicidad de documentación.
Estados Unidos	El Departamento de Salud y Servicios Humanos de Estados Unidos tiene la Ley de responsabilidad y portabilidad de seguros médicos. Allí incluye información de los procesos de desidentificación, los cuales generan información anonimizada. Estos procesos son: por determinación de un experto y puerto seguro.
Canadá	La información encontrada en Canadá se centra en datos estructurados. El Gobierno de Canadá presenta una orientación sobre el uso de la desidentificación, a través del Aviso de implementación de privacidad 2023, una técnica para preservar la privacidad de la información personal bajo el control de instituciones gubernamentales, además de presentar métodos técnicos para la desidentificación. Statistics Canada garantiza una conexión segura entre el servidor y un navegador con el uso de un cifrado de datos llamado Secure Socket Layer - SSL.
España	En España la información sobre las metodologías de anonimización de datos se enfocan básicamente en datos estructurados. Desde la Agencia Española de Protección de datos y el Ministerio de Asuntos Económicos y Transformación Digital se establecen los conceptos y las orientaciones para los procedimientos de anonimización de datos, partiendo de la continua relevancia que ha venido adquiriendo la información en la toma de decisiones y múltiples beneficios que aporta a la sociedad, metodología que se aplica con la finalidad de preservar y mantener la protección y la privacidad de los datos y el respeto por los derechos de las personas.
Reino Unido	La Oficina del Comisionado de Información (Information Commisioner's Office) del Reino Unido, mediante el documento Anonimización: gestionando el riesgo de protección de los datos: código de práctica (Anonymisation: managing data protection risk: code of practice[17]), establece que se necesitan técnicas de anonimización de datos no estructurados o cualitativos diferentes a las utilizadas con los datos estructurados o cuantitativos. Algunas de las técnicas de anonimización de datos no estructurados son: ocultar los nombres de las personas de los documentos, desenfocar las imágenes de video para disfrazar rostros, disfrazar electrónicamente o volver a grabar material de audio y cambiar los detalles en un informe (nombres de lugares precisos, fechas precisas, etc.).

[17] Disponible en: anonymisation-code.pdf

Referente	¿Cuáles son las metodologías de anonimización que aplican los referentes para datos estructurados y no estructurados?
	Por otro lado, el Servicio de Datos del Reino Unido[18] (UK Data Service) establece que al anonimizar datos cualitativos (como entrevistas transcritas), se deben utilizar datos textuales o audiovisuales, seudónimos o descriptores genéricos para editar la información de identificación y no de borrar la información. Esta información incluye las mejores prácticas de anonimización de datos y anonimización de datos audiovisuales.
Nueva Zelanda	De acuerdo con Digital Government NZ, estos son los métodos para desidentificar los datos de los neozelandeses: confidencialidad, agregación, desidentificación y seudo anonimización.
Países Bajos	La Autoridad de Protección de Datos de los Países Bajos (Autoriteit Persoonsgegevens) es responsable de supervisar la protección de datos personales en el país y proporciona información y recursos relevantes sobre privacidad y protección de datos. Se siguen las directrices del Reglamento General de Protección de Datos (RGPD) establecidas por la Unión Europea, las cuales abordan temas clave en el ámbito de internet y tecnología, así como en los medios de comunicación social. Estas directrices son fundamentales para establecer normas claras y salvaguardar la información no estructurada, como audio y video, así como los datos personales asociados. Además, el informe anual de la Autoridad destaca preocupaciones importantes relacionadas con el almacenamiento de datos gubernamentales en la nube y la necesidad de claridad en la responsabilidad de proteger los datos personales. Se enfatiza la importancia de la anonimización de datos como una medida para proteger la privacidad y la seguridad de los datos personales. Estas directrices son esenciales para garantizar la protección de datos en un entorno tecnológico en constante evolución y promover el cumplimiento de las regulaciones de protección de datos en los Países Bajos.

Fuente: DANE a partir de las revisiones de referentes.

2.3. Revisión de referentes

En esta sección se presenta la revisión de referentes internacionales de forma sintetizada.

[18] Disponible en: Anonymising qualitative data — UK Data Service

2.3.1. Colombia

El Centro Nacional de Memoria Histórica de Colombia (CNMH) desarrolló la Guía para la Anonimización de Datos e Información No Estructurada[19], 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, en términos de proceso de digitalización, se deben 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, aunque hayan sido reproducidos o almacenados en cualquier medio. Debe controlarse el acceso a los documentos, 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). En el caso de tratarse 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

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

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 por lo que 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, con el fin de 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 un proceso de comprimir los datos y empaquetarlos en un formato de audio. 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 telemonitoreo, sistema de red móvil, etc.
- Análisis de video: los videos representan el 80% de los datos no estructurados. No obstante, 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, y en las dimensiones del análisis el tamaño del video al ser mayor utiliza mayor capacidad de la red y del servidor en tiempo de procesamiento, 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, en el tráfico, en negocios, en seguridad, en el análisis para inteligencia empresarial, en el 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, análisis de sonido, análisis de texto, etc., han permitido un avance en términos de aplicación metodológica para este proceso.

2.3.2. Estados Unidos

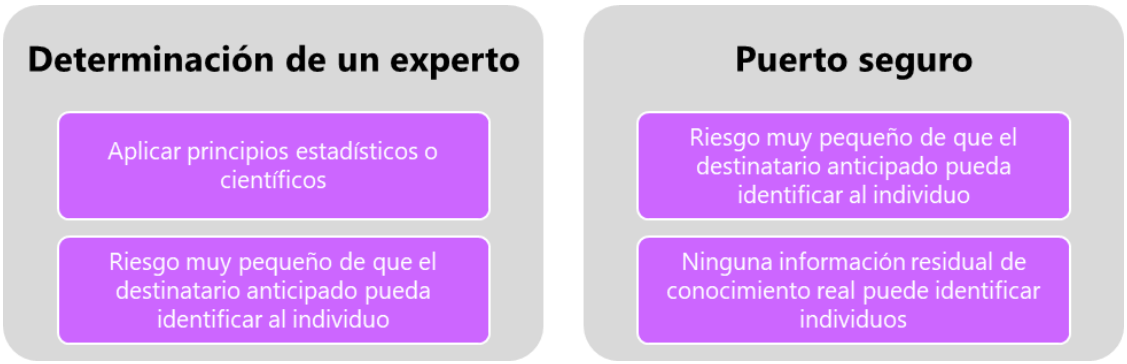
Una de las entidades que realiza la desidentificación de la información es el Departamento de Salud y Servicios Humanos[20] (HHS, por sus siglas en ingles), quienes en la Ley de responsabilidad y portabilidad de seguros médicos (HIPAA, por sus siglas en ingles), exponen dos métodos de desidentificación, que son: determinación de expertos y puerto seguro. La información que manejan son datos en su mayoría desestructurados:

- La salud o la condición física o mental pasada, presente o futura del individuo.
- La prestación de atención médica a la persona.
- El pago pasado, presente o futuro por la prestación de atención médica a la persona, y que identifica a la persona o para el cual existe una base razonable para creer que puede usarse para identificar a la persona. La información de salud protegida incluye muchos identificadores comunes (p. ej., nombre, dirección, fecha de nacimiento, número de Seguro Social) que se pueden asociar con la información de salud mencionada anteriormente.

La importancia de la desidentificación es que mitiga los riesgos de privacidad para las personas y, por lo tanto, respalda el uso secundario de datos para estudios de eficacia comparativa, evaluación de políticas, investigación en ciencias de la vida y otros esfuerzos. Al usar cualquiera de estos métodos se generan datos anonimizados que igualmente tienen cierto riesgo de ser identificados y asociados con un paciente.

[20] Disponible en: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#rationale

Figura 4. Métodos para lograr la desidentificación de acuerdo con la regla de privacidad de HIPAA



Fuente: hhs.gov.

Método de determinación de expertos

La sección 164.154(b) de la Ley HIPPA indica:

> (b) Especificaciones de implementación: requisitos para la desidentificación de la información de salud protegida. Una entidad cubierta puede determinar que la información de salud no es información de salud identificable individualmente solo si:
>
> (1) Una persona con conocimiento y experiencia adecuados de los principios y los métodos científicos y estadísticos generalmente aceptados para generar información no identificable individualmente: (i) Aplicando dichos principios y métodos, determina que el riesgo es muy pequeño de que la información pueda ser utilizada, sola o en combinación con otra información razonablemente disponible, por un receptor anticipado para identificar a una persona que es el sujeto de la información; y
>
> (ii) Documente los métodos y los resultados del análisis que justifiquen tal determinación.

Método de determinación de puerto seguro

La sección 164.154(b)(2) de la Ley HIPPA indica que:

> (2)(i) Se eliminan los siguientes identificadores del individuo o de familiares, empleadores o miembros del hogar del individuo:
>
> (A) Nombres.
>
> (B) Todas las subdivisiones geográficas más pequeñas que un estado, incluida la dirección, la ciudad, el condado, el recinto, el código postal y sus códigos geográficos equivalentes, excepto los tres dígitos iniciales del código postal si, de acuerdo con los datos disponibles al público actuales de la Oficina del Censo: (1) la unidad geográfica formada por la combinación de todos los códigos postales con los mismos tres dígitos iniciales contiene más de 20.000 personas; y

(2) los tres dígitos iniciales de un código postal para todas las unidades geográficas que contengan 20.000 personas o menos se cambien a 000.

(C) Todos los elementos de las fechas (excepto el año) para las fechas que están directamente relacionadas con una persona, incluida la fecha de nacimiento, la fecha de admisión, la fecha del alta, la fecha de la muerte y todas las edades mayores de 89 años y todos los elementos de las fechas (incluido el año) que indiquen dicha edad, excepto que dichas edades y elementos pueden agregarse en una sola categoría de 90 años o más.

(D) Números de teléfono.

(L) Identificadores de vehículos y números de serie, incluidos los números de matrícula.

(E) Números de fax.

(M) Identificadores de dispositivos y números de serie.

(F) Direcciones de correo electrónico.

(N) Localizadores de recursos universales web (URL).

(G) Números de seguro social.

(O) Direcciones de Protocolo de Internet (IP).

(H) Números de registros médicos.

(P) Identificadores biométricos, incluidas huellas dactilares y de voz.

(I) Números de beneficiarios del plan de salud.

(Q) Fotografías de rostro completo y cualquier imagen comparable.

(J) Números de cuenta.

(R) Cualquier otro número, característica o código de identificación único, excepto lo permitido por el párrafo (c) de esta sección [El párrafo (c) se presenta a continuación en la sección "Reidentificación"]; y

(K) Números de certificado/licencia

(ii) La entidad cubierta no tiene conocimiento real de que la información podría usarse sola o en combinación con otra información para identificar a un individuo que es un sujeto de la información.

Al realizar cualquiera de los métodos y generar información anonimizada no está protegida por la Ley HIPPA. Asimismo, lleva a la pérdida de datos que limita la utilidad de la información de salud resultante en ciertas circunstancias.

2.3.3. Canadá

El Gobierno de Canadá publicó el Aviso de implementación de privacidad 2023-01: Desidentificación[21], el cual proporciona información y orientación a las instituciones gubernamentales sobre el uso de la desidentificación como técnica de preservación de la privacidad. El objetivo es reforzar las protecciones de privacidad en relación con la información personal bajo el control de estas instituciones, en cumplimiento de sus obligaciones sobre protección de la privacidad.

Este aviso presenta consideraciones sobre el contexto y el riesgo de reidentificación, es decir, cómo la información desidentificada puede ser potencialmente vinculada nuevamente a individuos. Además, ofrece una introducción a los métodos para la desidentificación y proporciona definiciones de trabajo de términos clave relacionados con este proceso.

Se puede utilizar más de un método de desidentificación para reducir el riesgo de reidentificación. Esto, a su vez, reduce la probabilidad de uso indebido y minimiza el impacto de divulgaciones inapropiadas u otras violaciones de la privacidad. Según el grado de desidentificación aplicado, pueden ser necesarios otros controles técnicos y administrativos.

Las instituciones pueden utilizar y divulgar información anonimizada para diversas actividades, como investigación, análisis estadístico, revisión por pares, análisis de tendencias, toma de decisiones basadas en evidencia y evaluación de sesgos y daños. Estas actividades pueden involucrar la participación de otras instituciones gubernamentales o jurisdicciones y están destinadas a respaldar políticas y programas, así como a mitigar sesgos sistémicos en la información institucional.

Existe el riesgo de reidentificación, por ello se requiere determinar caso por caso y deben aplicarse las salvaguardas de privacidad adecuadas. El riesgo de reidentificación se puede considerar en términos de la información misma, la probabilidad de reidentificación y el daño potencial si se hace un mal uso de la información.

La desidentificación de la información puede realizarse manualmente o mediante el uso de herramientas como códigos o algoritmos. Los métodos manuales pueden implicar suprimir columnas de datos que son identificadores directos o altamente confidenciales. Otra opción es reducir la granularidad de los conjuntos de datos para hacer que las unidades sean menos únicas, como utilizar rangos de edad en lugar de edades exactas.

[21] Disponible en: https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/access-information-privacy-notices/2023-01-de-identification.html

Los métodos técnicos más sofisticados pueden emplear algoritmos para crear o enmascarar datos artificialmente. La elección de los métodos depende de varios factores, como el contexto, los datos y el nivel de riesgo residual aceptable de reidentificación. También es posible utilizar más de un método en un conjunto de datos.

Es importante destacar que los métodos técnicos son solo una parte de la preservación de la privacidad. Además de la desidentificación, se requieren otros controles administrativos, como acuerdos y arreglos, controles de acceso y auditorías, para reducir el riesgo de divulgación inadvertida, acceso no autorizado, reidentificación o inferencia. Estos controles contribuyen a preservar y promover la privacidad de las personas en general.

Los métodos utilizados por el Gobierno de Canadá en la desindentificación de datos no estructurados son:

- Pequeño número de individuos: en el PIN 2020-03[22] se aborda medidas de protección contra la reidentificación, como la supresión, el enmascaramiento y la redacción de identificadores directos, así como la evaluación del riesgo de reidentificación a través de identificadores indirectos. También ofrece orientación sobre la publicación de tablas agregadas, incluyendo la determinación del tamaño mínimo de las celdas y la modificación de los datos para mitigar el riesgo de reidentificación.
- Enmascaramiento y ofuscación: son métodos para borrar, eliminar, reemplazar, alterar u ocultar identificadores en un conjunto de datos. Algunos ejemplos son la anulación (suspensión de campo), el redondeo o la generalización, el regex, la perturbación, el uso de funciones criptográficas y el agregar valores aleatorios.
- Seudonimización: es un proceso de enmascaramiento de identificadores directos mediante el uso de alias consistentes en varios conjuntos de datos. Aunque los identificadores directos se reemplazan, se debe anonimizar también aquellos con una "clave" pueden volver a identificar a las personas. Es útil para usos internos donde se requiere un identificador único, como archivos jerárquicos o vinculación de fuentes de información.
- Generación de datos sintéticos: este proceso implica que un algoritmo cree un nuevo conjunto de datos con valores completamente falsos, manteniendo las mismas propiedades estadísticas del conjunto de datos original. Esto puede ser útil en el entrenamiento de herramientas de inteligencia artificial y aprendizaje automático, así como en estrategias de datos abiertos para capacitación, divulgación y familiarización con datos.

[22] Disponible en: https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/access-information-privacy-notices/2020-03-protecting-privacy-releasing-information-about-small-number-individuals.html

Para el caso de Statistics Canada, utilizan un cifrado de datos en el sitio web equipado con Secure Socket Layer (SSL)[23] para garantizar una conexión segura entre el servidor y un navegador compatible con SSL. El protocolo SSL proporciona un paso seguro para transmitir y autenticar datos mediante el cifrado de la información. Los datos no pueden verse comprometidos cuando SSL está en uso. Esta es la forma de encriptación más segura comúnmente disponible en América del Norte.

Además, de un monitoreo de tráfico de red que emplea programas de software para monitorear el tráfico de la red y para identificar intentos no autorizados de cargar o cambiar información o causar daños de otro modo. Estos programas también se utilizan para recopilar información anónima, como estadísticas para mejorar la funcionalidad del sitio web.

Desde el Comisionado de Información y Privacidad de Ontario se publicaron las Directrices para la desidentificación de datos estructurados[24], en el que se presenta que la desidentificación es el proceso de eliminación de cualquier información que identifique a un individuo o para la que exista una expectativa razonable de que la información podría utilizarse, ya sea sola o con otra información, para identificar a un individuo.

Proceso de desidentificación de datos estructurados

Para proteger la privacidad de las personas preservando al mismo tiempo la mayor utilidad posible de la información, la cantidad y los tipos de desidentificación deben determinarse mediante un análisis sistemático del nivel y los tipos de riesgo de reidentificación que implica la divulgación de un conjunto de datos. A continuación, se presentan los pasos a tener en cuenta:

1. Determinar el modelo de divulgación y la forma en que se divulga un conjunto de datos desidentificados puede variar entre público, semi-público o no público: cada modelo de divulgación tiene diferentes niveles de disponibilidad y protección de la información. La elección del modelo depende de los propósitos y los requisitos legislativos de la divulgación. El modelo seleccionado determinará la cantidad de desidentificación necesaria.
2. Clasificar las variables dependiendo del tipo de información, dado que algunas variables pueden usarse para identificar a los individuos, ya sea directa o indirectamente, mientras que otras no. La desidentificación se ocupa únicamente de las variables que pueden usarse para identificar a los individuos.
3. Determinar un umbral de riesgo de reidentificación aceptable, puesto que la cantidad de desidentificación requerida está en proporción al riesgo de reidentificación asociado con la divulgación del conjunto de datos. Cuanto mayor sea el riesgo de reidentificación, mayor será

[23] Disponible en: https://www.statcan.gc.ca/en/reference/privacy

[24] Disponible en: https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

la cantidad de desidentificación necesaria. Para determinar un nivel aceptable de riesgo de reidentificación para un conjunto de datos, es necesario evaluar en qué medida la divulgación del conjunto de datos invadiría la privacidad de una persona. El resultado de esta evaluación debe ser un valor cualitativo típicamente clasificado como "bajo", "medio" o "alto".

4. Medir la cantidad de riesgo de reidentificación en un conjunto de datos es un proceso de dos pasos: se debe calcular la probabilidad de reidentificación de cada fila y aplicar el método de medición de riesgo adecuado según el modelo de divulgación utilizado.
5. Medir el riesgo de contexto: analizar el riesgo de reidentificación en términos de posibles ataques produce el riesgo contextual. Junto con el riesgo de los datos, este valor se utiliza para calcular el riesgo general de reidentificación en la divulgación de un conjunto de datos. Los ataques de reidentificación y los adversarios pueden variar según el modelo de divulgación utilizado.
6. Calcular el riesgo global: una vez medidos el riesgo de los datos y el riesgo del contexto, puede calcularse el riesgo global de reidentificación. El riesgo global es igual al riesgo de datos multiplicado por el riesgo de contexto.
7. Desidentificar los datos: es necesario eliminar cualquier información que pueda identificar a una persona o que pueda utilizarse en combinación con otra información para identificarla. Esto se puede lograr mediante diversas técnicas, dependiendo del tipo y la naturaleza de los identificadores. Para eliminar información identificable incluyen el enmascaramiento de identificadores directos, la modificación del tamaño de las clases de equivalencia y garantizar que el riesgo global de reidentificación sea inferior o igual al umbral establecido.
8. Evaluar la utilidad de los datos: cuanto más se desidentifiquen las variables, mayor es el riesgo de pérdida de utilidad del conjunto de datos. La generalización y la supresión se pueden utilizar de diferentes formas y combinaciones para reducir el riesgo de reidentificación.
9. Documentar el proceso: elaborar un informe que documente el proceso de desidentificación y sus resultados ofrece beneficios como la demostración de cumplimiento, generación de confianza y mayor transparencia, concienciación y comprensión en las prácticas de gestión de la información de la organización.

2.3.4. España

La información es un activo invaluable para la sociedad actual, ya que nos permite tomar decisiones en todos los aspectos y los sectores; por lo tanto, el acceso, el tratamiento y el análisis de información brindan múltiples beneficios. Sin embargo, se debe preservar y mantener la protección y la privacidad de los datos y el respeto por los derechos de las personas.

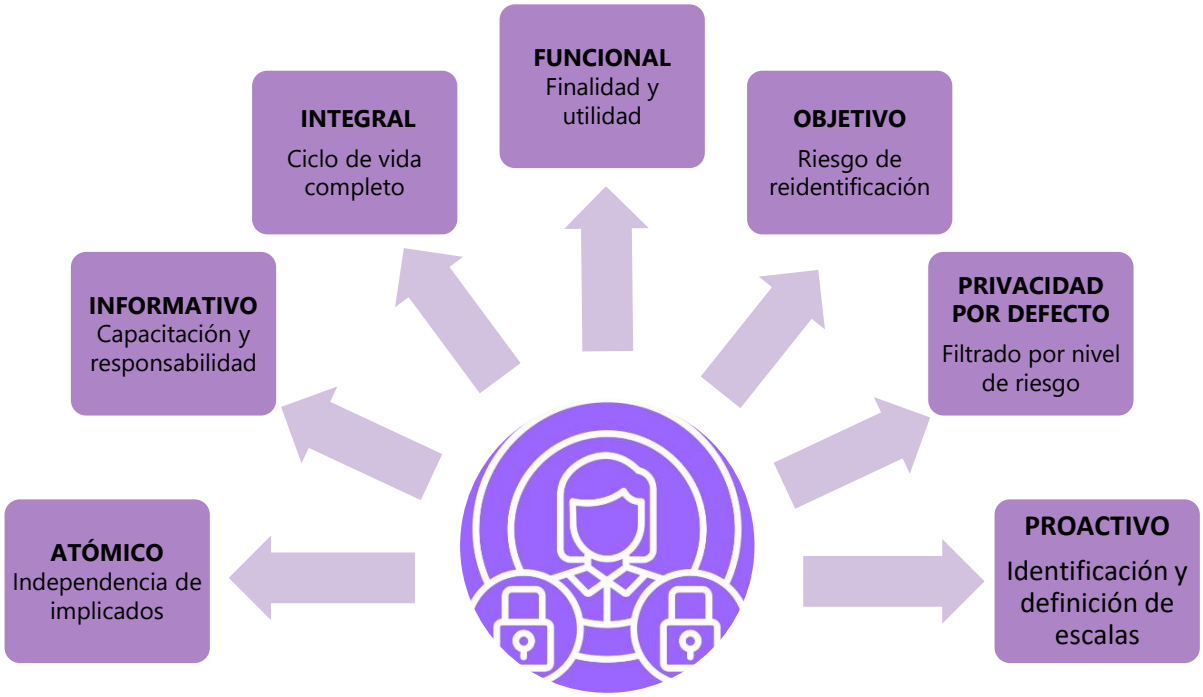
Con el fin de garantizar que la sociedad cuente con acceso de la información y esta se maneje correctamente sin menoscabar el respeto a la protección de datos se recurre a la anonimización,

metodología que adquiere un papel fundamental como una forma de eliminar las posibilidades de identificación de la información por medio de ocultación, enmascaramiento o disociación de los datos.

La finalidad del proceso de anonimización es eliminar o reducir al mínimo los riesgos de reidentificación de los datos anonimizados manteniendo la veracidad de los resultados del tratamiento de estos, es decir, además de evitar la identificación de las personas, los datos anonimizados deben garantizar que cualquier operación o tratamiento que pueda ser realizado con posterioridad a la anonimización no conlleva una distorsión de los datos reales. Un análisis masivo de los datos o los macrodatos que puedan derivar de los datos anonimizados no debería diferir del análisis que pudiera obtenerse si hubiera sido realizado con datos no anonimizados[25].

Los procesos de anonimización se deben regir por el concepto de protección de datos desde el diseño según el Reglamento General de Protección de Datos (RGPD), es decir que los requisitos de privacidad se tienen en cuenta desde las etapas iniciales y durante todo el ciclo de vida del producto o el sistema de información, teniendo en cuenta siete principios:

Figura 5. Principios proceso de anonimización de datos



Fuente: DANE, a partir de información de: https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

[25] Disponible en: https://datos.gob.es/sites/default/files/doc/file/orientaciones_y_garantias_anonimizacion_0.pdf

Descripción de los principios de anonimización:

- **Proactivo:** el diseño debe plantearse desde las etapas iniciales de conceptualización, identificando microdatos, datos de identificación indirecta y datos sensibles, estableciendo escalas de sensibilidad las cuales pueden ser cualitativas o cuantitativas que servirán de referencia dentro de una organización. Estas escalas deben divulgarse a todos los actores que participan en el proceso de anonimización y será fundamental en el análisis de riesgos o la Evaluación de Impacto en la Protección de los Datos Personales (EIPD).
- **Privacidad por defecto:** el objetivo del diseño de un sistema de información anonimizada es garantizar la confidencialidad de los interesados, por lo cual es necesario establecer el grado de detalle o granularidad de los datos anonimizados con el objetivo de preservar la confidencialidad al eliminar variables no esenciales para el estudio a realizar y teniendo en cuenta factores de riesgo y beneficio.
- **Objetivo:** en el proceso de anonimización es muy difícil lograr una anonimización absoluta, por lo tanto, es necesario evaluar el nivel de riesgo de reidentificación asumido y establecer las políticas adecuadas de contingencia.
- **Funcional:** para garantizar la utilidad de los datos anonimizados, es necesario tener clara la necesidad de información de los usuarios de la información anonimizada y definir claramente la finalidad del análisis que se va a realizar sobre los datos una vez anonimizados.
- **Integral:** el proceso de anonimización va más allá de la generación del conjunto de datos, siendo aplicable también durante el estudio de estos, a través de contratos de confidencialidad y uso limitado, validados mediante las auditorías pertinentes durante todo el ciclo de vida del proceso de anonimización[26].
- **Informativo:** todo el personal con acceso a los datos anonimizados o no anonimizados deben estar debidamente formados, capacitados e informados respecto a sus responsabilidades, sus obligaciones y los riegos asociados.
- **Atómico:** el equipo de trabajo se debe definir con personas independientes para cada una de las funciones dentro del proceso.

Fases de la anonimización

En un proceso de anonimización es favorable definir un protocolo de actuación, es de vital importancia destacar que el protocolo es dinámico y se debe ajustar a las particularidades del proceso de anonimización que se trabaje. Algunos de los elementos que se deben tener en cuenta son:

1. Definición del equipo de trabajo.

[26] Disponible en: https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

2. Evaluación de riesgos de reidentificación.
3. Definición de objetivos y finalidad de la información anonimizada.
4. Viabilidad del proceso.
5. Preanonimización: definición de variables.
6. Eliminación/reducción de variables.
7. Selección de técnicas de anonimización.
8. Segregación de la información.
9. Proyecto piloto.
10. Anonimización.
11. Formación e información al personal implicado.
12. Garantías jurídicas.
13. Auditoría del proceso de anonimización.

Tipos de riesgos

Es necesario realizar un análisis de riesgos del proceso de anonimización teniendo en cuenta que ninguna técnica de anonimización garantiza en totalmente la imposibilidad de la reidentificación, ya que existirá siempre una probabilidad de reidentificación.

Es importante señalar que el riesgo de reidentificación aumenta con el paso del tiempo, debido a la posible aparición de nuevos datos o el desarrollo de nuevas técnicas, avances en computación cuántica, que podrían conllevar la ruptura de claves de cifrado[27].

Existen los siguientes tres vectores de riesgo asociados a la reidentificación:

[27] Disponible en: https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

Figura 6. Vectores de riesgo anonimización de datos

Fuente: DANE, basado en información de: https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

Evaluación de riesgos de reidentificación

Algunas de las fases necesarias para el análisis y la evaluación de riesgos de reidentificación son:

- Identificación y categorización de activos implicados en el proceso de anonimización.
- Constitución del equipo de trabajo.
- Identificación de riesgos.
- Valoración de los riesgos existentes.
- Salvaguardas.
- Cuantificar el impacto.
- Informe de riesgos.
- Determinación del umbral de riesgos aceptable.
- Gestión de los riesgos asumibles.
- Informe final.
- Revisión de riesgos.

Definición de objetivos y finalidad de la información anonimizada

El proceso de anonimización debe de garantizar la privacidad de los interesados y a su vez determinar los objetivos que deberá cumplir la información anonimizada en función de las necesidades de su

destinatario. El diseño del proceso de anonimización estará condicionado por el objetivo final de la información anonimizada dando lugar a información de uso restringido o a datos abiertos.

Preanonimización: definición de variables de identificación

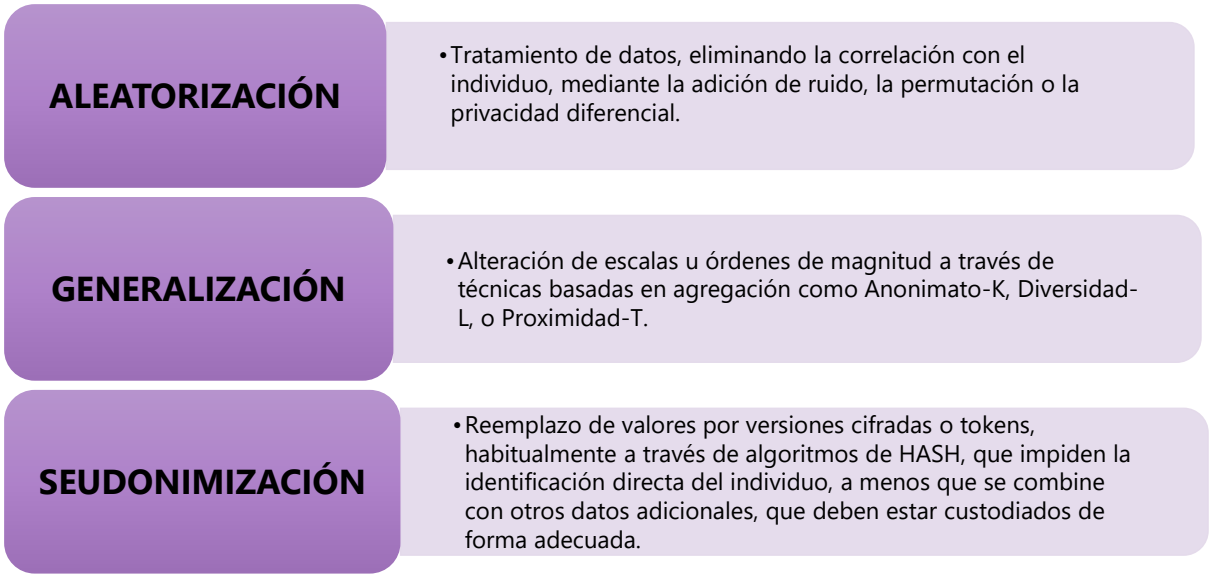
En esta fase se definen las posibles variables de identificación (directas e indirectas) que se utilizaran durante el proceso de anonimización y se debe tener en cuenta:

- La determinación de variables: datos personales, identificadores directos e indirectos, datos especialmente protegidos y otros datos con carácter confidencial.
- La clasificación y la sensibilidad de las variables por categorías: de identificación directa, de identificación geográfica, de carácter especialmente protegido, numéricas, temporales, metadatos, etc.
- Variables de identificación que no puedan ser anonimizadas y que sea preciso eliminar del proceso de anonimización.
- Variables anonimizadas que sean imprescindibles para la finalidad a la que se van a destinar los datos anonimizados.

Es imprescindible que las variables de identificación que se determinen sean las realmente necesarias para la finalidad a la que será destinada la información anonimizada.

Técnicas de anonimización

Figura 7. Técnicas de anonimización de datos

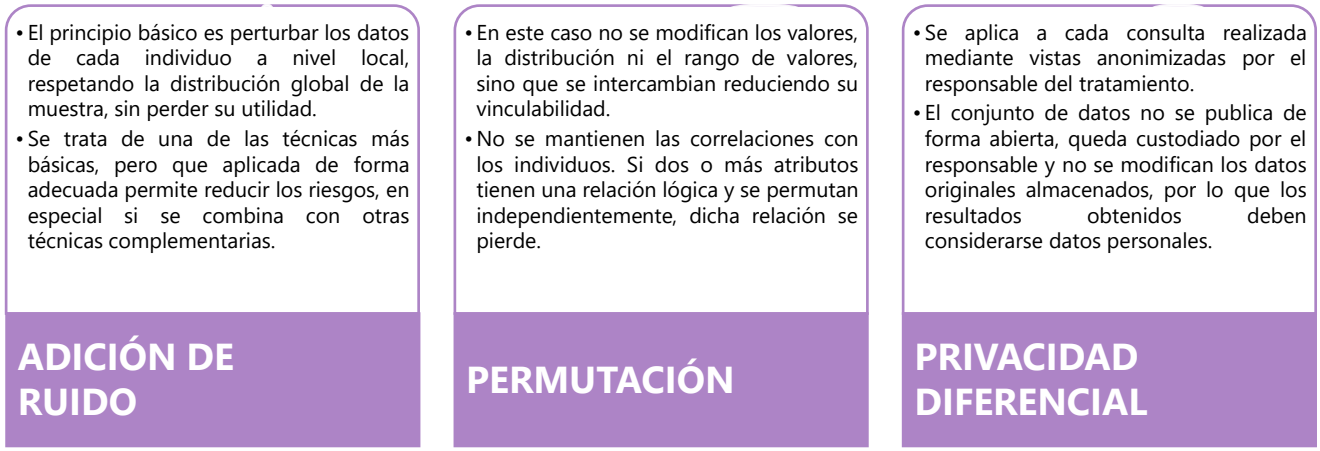


Fuente: DANE, basado en información de https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

Técnicas de aleatorización

Con esta técnica lo que se hace es modificar o alterar la veracidad de los datos a nivel individual, respetando la distribución global de estos, con la finalidad de reducir la vinculabilidad y la inferencia. La aleatorización empleada de forma aislada no es efectiva frente a la singularización. Siempre deben combinarse, al menos, con un proceso de filtrado explícito de atributos obvios o identificadores indirectos o indirectamente mediante técnicas de generalización.

Figura 8. Técnicas de aleatorización de anonimización de datos



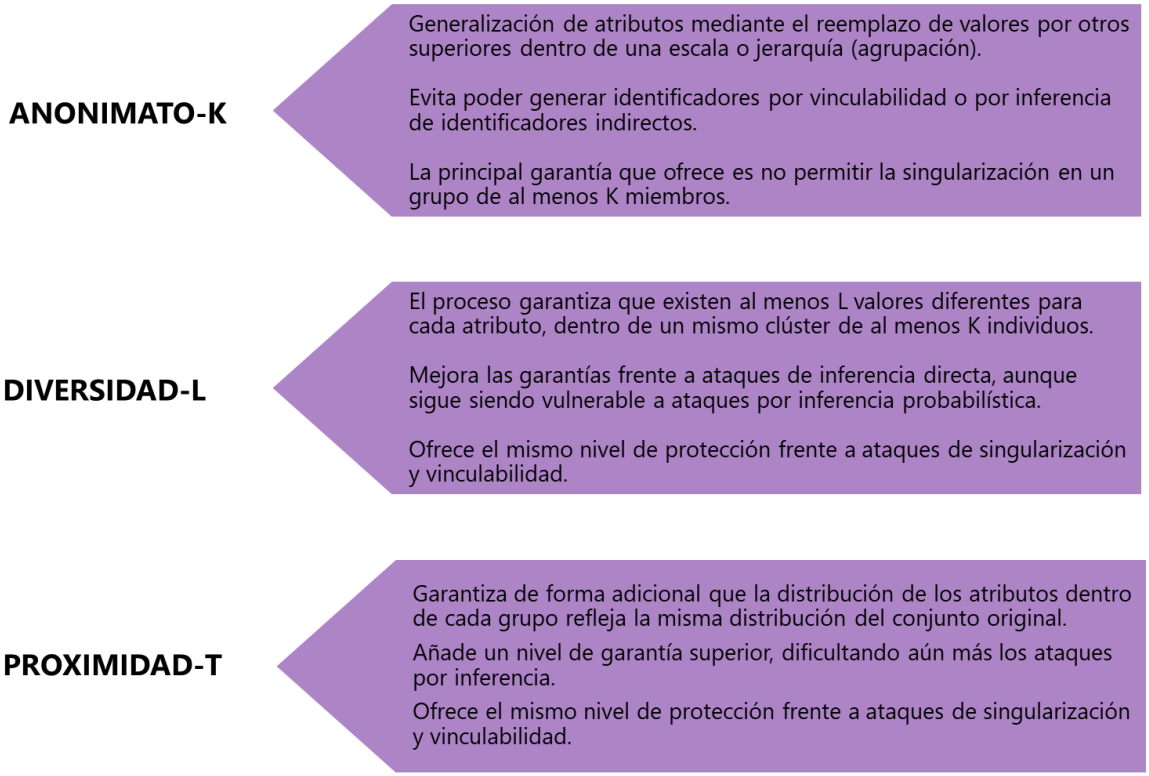
Fuente: DANE, basado en información de https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

Técnicas de generalización

Tiene por objetivo generalizar algunos atributos críticos de forma que se evite la singularización, por ejemplo, modificando las escalas o los órdenes de magnitud, reemplazando valores por categorías superiores en una jerarquía.

Es necesario aplicar el proceso de forma adecuada y aplicar otras técnicas de forma conjunta para garantizar la protección ante ataques de inferencia o vinculabilidad.

Figura 9. Técnicas de generalización de anonimización de datos



Fuente: DANE, basado en información de https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

Técnicas de seudonimización

La seudonimización no se considera un método de anonimización (Dictamen 05/2014), ni tampoco los conjuntos de datos resultantes pueden considerar conjuntos de datos anónimos. Si bien, resultan medidas útiles para mejorar la seguridad al reducir la vinculabilidad del conjunto de datos obtenido. El problema principal de estas técnicas es que los individuos del conjunto siguen siendo vulnerables a ataques por singularización, dado que son identificables por los pseudónimos y/o tokens. Además, tampoco protegen frente a vinculabilidad o inferencia, especialmente en aquellos casos donde se reutilicen atributos seudonimizados en diferentes conjuntos de datos, por ejemplo, por reutilización de claves. También podría realizarse vinculación a partir de otros atributos del conjunto[28].

[28] Disponible en: https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

Figura 10. Técnicas de seudonimización de anonimización de datos

Fuente: DANE, basado en información de https://datos.gob.es/sites/default/files/doc/file/informe-anonimizacion-es.pdf

Segregación de la información

Con el fin de garantizar la confidencialidad de la información anonimizada, es vital elaborar un mapa de sistemas de información que garantice entornos separados para cada tratamiento de datos. La segregación de entornos para los tratamientos de la información implicará también la segregación del personal que accede a la información y a los datos personales.

Proyecto piloto[29]

Es recomendable la realización de un proyecto piloto con una pequeña muestra de datos de prueba (no reales) en el que puedan obtenerse de forma objetiva conclusiones acerca de la viabilidad de todas las propuestas de los miembros del equipo de anonimización. Algunos de los objetivos que pueden ser tenidos en cuenta para la realización del proyecto piloto son los siguientes:

- Arrojar luz sobre los resultados de las técnicas de anonimización propuestas.
- Comprobar la fortaleza de los resultados frente a posibles intentos de reidentificación.
- Cuantificar los costes del proceso de anonimización.
- Asegurar que los objetivos que pretenden los procedimientos de anonimización son viables.

[29] Disponible en: https://datos.gob.es/sites/default/files/doc/file/orientaciones_y_garantias_anonimizacion_0.pdf

- Garantizar el diferencial de privacidad mediante pruebas encaminadas a evaluar y cuantificar las desviaciones o las distorsiones resultantes del proceso de anonimización.
- Evaluar los resultados del proyecto frente a los resultados del análisis de riesgos.

Los resultados del proyecto piloto se deben compartir con el destinatario de la información anonimizada junto con los datos no reales que hayan sido utilizados para la realización del proyecto piloto, de tal forma que sea posible valorar si el resultado del proceso de anonimización se ajusta a los objetivos a los que se va a destinar la información. Igualmente, permitir a su legítimo destinatario valorar si el diferencial de privacidad es aceptable o si, por el contrario, la distorsión generada por el proceso de anonimización genera un diferencial de privacidad que hace que la información resultante no pueda ser utilizada con la finalidad que se persigue.

2.3.5. Reino Unido

La Directiva de Protección de Datos de 1995 establece que los principios de protección de datos se aplicarán a los datos anonimizados, de tal manera que el sujeto no sea identificable. También menciona que un código de práctica puede proporcionar orientación sobre las formas en que los datos pueden convertirse en anónimos y conservados de tal forma en que la identificación del sujeto no sea posible. Hasta donde se tiene conocimiento, el código de práctica ('Anonimización: gestionando el riesgo de protección de los datos) de la Oficina del Comisionado de Información (ICO por sus siglas en inglés) es el primer código de la práctica en materia de anonimización en ser publicado por cualquier autoridad de protección de datos europea. Los problemas relacionados con las técnicas de anonimización y el estado de los datos anónimos se está convirtiendo en una cuestión clave, ya que el debate sobre la propuesta de la Comisión Europea del marco de protección de nuevos datos continúa.

El código de práctica publicado por la ICO menciona algunas técnicas clave de anonimización. Entre estas se encuentran:

Enmascarar datos: esto implica eliminar identificadores personales obvios como nombres de una pieza de información para crear un conjunto de datos en los que no hay identificadores de persona.

Eliminación parcial de datos: resultados en datos donde algunos identificadores personales, por ejemplo, el nombre y la dirección se eliminan, pero otros, como las fechas de nacimiento, permanecen.

Cuarentena de datos: es la técnica de solo suministrar datos a un destinatario que es poco probable que tenga acceso o no puede tenerlo a los demás datos necesarios para facilitar la reidentificación. Puede implicar la divulgación de identificadores personales únicos, por ejemplo, la referencia a números, pero no la "clave" necesaria para vincularlos a individuos particulares.

Estas son técnicas de riesgo relativamente alto porque los datos anónimos todavía existen en forma individual. Los datos del censo electoral, por ejemplo, podrían usarse para reintroducir nombres que

se han eliminado en el conjunto de datos con bastante facilidad. Sin embargo, este tipo de datos también es relativamente "rico" en términos de permitir que un individuo sea rastreado como parte de un estudio longitudinal, por ejemplo.

- **Seudoanonimización**

Desidentificación de datos para que una referencia codificada o seudónimo sea adjunto a un registro para permitir que los datos se asocien con un individuo particular sin que el individuo sea identificado.

La modificación determinista es una técnica similar. En este caso, "determinista" significa que siempre se reemplaza el mismo valor original por el mismo valor modificado. Esto significa que, si múltiples registros de datos están vinculados, en el sentido en que el mismo nombre (o dirección, o número de teléfono, por ejemplo) está presente en todos esos registros, los registros correspondientes en el conjunto de datos modificado también estarán vinculado de la misma manera. Esto facilita ciertos tipos de análisis de datos.

Igualmente, esta es una técnica de riesgo relativamente alto, con fortalezas y debilidades similares a las del enmascaramiento de datos.

- **Agregación**

Los datos se muestran como totales, por lo que no se muestran los datos relacionados con ningún individuo. Los números pequeños en los totales son a menudo suprimidos por "borrosidad" o se omiten por completo.

- **Supresión de celdas:** si los datos provienen de una encuesta por muestreo, entonces puede ser inapropiado liberar salidas tabulares con celdas que contienen un pequeño número de individuos, por ejemplo, por debajo de 30. Esto se debe a que el error de muestreo en tales estimaciones de celdas normalmente sería demasiado grande para que las estimaciones sean útiles para fines estadísticos. En este caso, la supresión de celdas con números pequeños para fines de calidad actúa en conjunto con la supresión con fines de divulgación.
- **Control de inferencia:** algunos valores de celda (por ejemplo, pequeños como 1-5) en datos estadísticos puede presentar un mayor riesgo de reidentificación. Dependiendo de las circunstancias, se pueden suprimir los números pequeños o los valores pueden ser manipulados (como en el método Barnardisation). Si un gran número de celdas se ven afectadas, el nivel de agregación podría cambiar. Por ejemplo, los datos podrían vincularse a una geografía más amplia. Podrían ampliarse las zonas o las franjas de edad.
- **Perturbación:** como en Barnardisation es un método de control de divulgación para tablas o recuentos. Implica sumar o restar 1 aleatoriamente de ciertas celdas de la tabla. Esta es una forma de perturbación.

- **Redondeo:** redondear una figura hacia arriba o hacia abajo para disimular la precisión estadística. Por ejemplo, si una tabla tiene una celda con valor de 10.000 para todas las personas que realizan alguna actividad hasta la fecha actual. Sin embargo, al mes siguiente, la cifra en esa celda se eleva a 10.001. Si un intruso compara las tablas sería fácil deducir una celda de 1. El redondeo evitaría esto.
- **Muestreo**: en algunos casos, cuando un gran número de registros están disponibles puede ser adecuado, para propósitos estadísticos, publicar una muestra de registros seleccionados a través de un procedimiento aleatorio. Al no liberar detalles de la muestra, los propietarios de los datos pueden minimizar el riesgo de reidentificación.
- **Datos sintéticos:** combinar los elementos de un conjunto de datos, o crear nuevos valores basados en los datos originales, de modo que todos los totales generales y valores del conjunto se conserven, pero no se relacionen con ningún individuo en particular.
- **Informes tabulares:** es un medio para producir datos tabulares (agregados), que protegen contra la reidentificación.

Estas son técnicas de riesgo relativamente bajo porque generalmente es difícil averiguar algo sobre un individuo particular, mediante el uso de datos agregados. Estos datos no pueden apoyar la investigación a nivel individual, pero, por ejemplo, pueden ser suficientes para analizar las tendencias sociales a nivel regional.

Elementos de datos derivados y bandas

Los datos derivados son un conjunto de valores que reflejan el carácter de los datos de origen, pero que ocultan los valores originales exactos. Esto generalmente se realiza mediante el uso de técnicas de bandas para producir descripciones de valores más gruesas que en la fuente conjunto de datos, por ejemplo, reemplazar las fechas de nacimiento por edades o años, direcciones por zonas de residencia, utilizando códigos postales parciales o redondear las cifras exactas para que aparezcan en forma normalizada.

Esta es una técnica de riesgo relativamente bajo porque las técnicas de bandas dificultan la coincidencia de datos o la hacen imposible. Los datos resultantes pueden ser relativamente sustanciales porque pueden facilitar la investigación a nivel individual, pero presenta un riesgo de reidentificación relativamente bajo.

El Servicio de Datos del Reino Unido sugiere que, al anonimizar datos cualitativos, se deberían utilizar seudónimos o descriptores para editar información de identificación vez de borrarla. Se debe tener en cuenta el nivel de anonimato requerido para satisfacer las necesidades acordadas durante el proceso de consentimiento informado. La planificación previa y el acuerdo con los participantes durante el proceso de consentimiento, sobre lo que puede y no puede ser registrado o transcrito, puede ser una forma mucho más efectiva de crear datos que representen con precisión el proceso de investigación y

la contribución de los participantes. A modo de ejemplo, si el nombre de un empleador no puede ser revelado, se debe acordar de antemano que no se mencionará durante una entrevista. Esto es más fácil que pasar tiempo más tarde eliminándolo de una grabación o transcripción.

Las mejores prácticas para anonimizar datos:

- No recopilar datos reveladores a menos que sea necesario. Por ejemplo, no solicitar nombres completos si no se pueden utilizar en los datos.
- Planificar la anonimización en el momento de la transcripción o la redacción inicial (los estudios longitudinales pueden ser una excepción si las relaciones entre oleadas de entrevistas requieren atención especial para la edición armonizada).
- Utilizar seudónimos o reemplazos que sean consistentes dentro del equipo de investigación y durante todo el proyecto. Por ejemplo, utilizar los mismos seudónimos en publicaciones e investigaciones de seguimiento.
- Utilizar técnicas de "búsqueda y reemplazo" con cuidado, para que no se realicen cambios no deseados y no se pierdan palabras mal escritas.
- Identificar claramente los reemplazos en el texto, por ejemplo, con [corchetes] o utilizando etiquetas XML, como <seg>palabra para anonimizar </seg>.
- Mantener versiones sin editar de los datos para su uso dentro del equipo de investigación y para su preservación.
- Crear un registro de anonimización de todos los reemplazos, las agregaciones o las eliminaciones realizadas y almacene dicho registro por separado de los archivos de datos anónimos.
- Considerar redactar declaraciones donde exista un mayor riesgo de daño o divulgación.

La herramienta ayudante de anonimización de datos[30] del Servicio de Datos del Reino Unido puede ser útil para encontrar información reveladora para eliminar o seudonimizar en archivos de datos cualitativos. La herramienta no anonimiza ni realiza cambios en los datos, pero utiliza macros de MS Word para encontrar y resaltar números y palabras que comienzan con letras mayúsculas en el texto. Los números y las palabras en mayúscula son a menudo reveladores, por ejemplo, nombres, empresas, fechas de nacimiento, direcciones, instituciones educativas y países.

Anonimización de datos audiovisuales

Las grabaciones de voz y audio casi siempre se refieren a una persona identificable y constituyen datos personales según la legislación de protección de datos. La anonimización de los datos audiovisuales

[30] Disponible en Anonymising qualitative data — UK Data Service

debe hacerse con sensibilidad, teniendo en cuenta la legislación aplicable. Para los datos de audio, las mejores prácticas incluyen pitar nombres reales o nombres de lugares y la eliminación o la alteración de cualquier discurso o vocalización identificable que pueda conducir a la reidentificación de individuos. Para los datos de video la información identificable, como rostros, nombres y marcas de identificación, debe eliminarse u ocultarse.

Los investigadores siempre deben considerar para qué se utilizarán los datos y si las técnicas de anonimización reducirían significativamente la usabilidad de los datos. Si la anonimización resultara en demasiada pérdida de datos, obtener el consentimiento del participante para usar y compartir los datos sin alteraciones puede considerarse como una mejor estrategia.

2.3.6. Nueva Zelanda

La Infraestructura Integrada de Datos[31] (IDI) de Stats NZ es una base de datos de investigación grande de datos des identificados de los neozelandeses. Contiene microdatos anónimos sobre personas y hogares.

Los datos son sobre eventos de la vida, como educación, ingresos, beneficios, migración, justicia y salud. Provienen de agencias gubernamentales, encuestas de Stats NZ y organizaciones no gubernamentales (ONG). Los datos están vinculados entre sí, o integrados, para formar el IDI.

El IDI complementa la Base de Datos Longitudinal de Empresas (LBD), que contiene microdatos vinculados sobre las empresas. Las dos bases de datos están vinculadas a través de datos fiscales.

De acuerdo con Digital Government NZ, estos son los métodos[32] que pueden utilizarse para reducir la cantidad de información personal identificable:

- **Confidencialidad:** este método aplica técnicas estadísticas a los datos para evitar la identificación individual. Por ejemplo, puede incluir el rango de edad y dar la altura promedio del grupo, pero luego eliminar datos sobre el color del cabello porque un individuo podría ser identificado por él (según lo determinado por el análisis estadístico). Se tienen en cuenta los métodos estadísticos utilizados para proteger contra la divulgación de información confidencial a personas que no están autorizadas a tener acceso a ella, de una manera que pueda identificar a una persona, hogar u organización. Los métodos estadísticos utilizados proporcionan un nivel de protección contra la identificación que no puede obtenerse de la desidentificación.
- **Agregación:** este método combina datos individuales de un grupo de personas en rangos de edad, cabello y altura. Este método tiene en cuenta datos combinados a partir de varias

[31] Disponible en: Integrated Data Infrastructure | Stats NZ

[32] Disponible en: Making personal information safe for reuse | NZ Digital government

mediciones, pero sin el uso adicional de métodos estadísticos para proteger contra la reidentificación.

- **Desidentificación:** este método elimina el nombre y reemplaza la fecha de nacimiento con su edad. Es el proceso de eliminación de información de microdatos para reducir el riesgo de reconocimiento espontáneo. Por lo general, incluye la eliminación de nombres, fechas exactas de nacimiento o muerte y direcciones exactas.
- **Seudo anomización:** este método asigna a la persona un nombre diferente u otra forma de identificarla, como asignarle un grupo de letras, como XYZ, junto con su fecha de nacimiento y color de cabello. El proceso de sustitución de identificadores directos por otros diferentes en microdatos para reducir el riesgo de reconocimiento espontáneo.

2.3.7. Países Bajos

La Autoridad de Protección de Datos de los Países Bajos, conocida como Autoriteit Persoonsgegevens[33], es la agencia gubernamental encargada de supervisar la protección de datos personales en el país. Esta autoridad proporciona información y recursos relevantes relacionados con la privacidad y la protección de datos.

En los Países Bajos, se siguen las directrices del Reglamento General de Protección de Datos (RGPD) establecidas por la Unión Europea. Estas directrices, publicadas por el Consejo Europeo de Protección de Datos (EDPB por sus siglas en inglés)[34] aclaran varios aspectos del RGPD y ayudan a las organizaciones a aplicar las leyes de privacidad en su trabajo. Dentro de las directrices del RGPD[35] se destacan aquellas relacionadas con el ámbito de "Internet y tecnología"[36]. Estas directrices abordan temas como el circuito cerrado de televisión, los coches conectados, el reconocimiento facial y el perfilado. Además, en el ámbito de los "Medios de comunicación social"[37], se tratan temas como el diseño engañoso, la segmentación de usuarios y los asistentes de voz.

---

[33] Disponible en: https://www.autoriteitpersoonsgegevens.nl/

[34] Disponible en: https://www.autoriteitpersoonsgegevens.nl/themas/internationaal/internationale-samenwerking/european-data-protection-board-edpb

[35] Disponible en https://www.autoriteitpersoonsgegevens.nl/themas/internationaal/internationale-samenwerking/overzicht-van-avg-guidelines

[36] Disponible en: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_201903_video_devices.pdf

[37] Disponible en: https://edpb.europa.eu/system/files/2023-02/edpb_03-2022_guidelines_on_deceptive_design_patterns_in_social_media_platform_interfaces_v2_en_0.pdf

Estas directrices son fundamentales para garantizar la protección de datos y la privacidad en el contexto del avance tecnológico y la proliferación de las redes sociales. Al proporcionar orientación clara sobre aspectos específicos, como la seguridad de los sistemas de vigilancia, el uso ético de la tecnología de reconocimiento facial y la transparencia en el diseño de interfaces de redes sociales, estas directrices ayudan a las organizaciones y a los usuarios a comprender y cumplir con las regulaciones de protección de datos en los Países Bajos.

Las directrices revisadas abordan varias preocupaciones relacionadas con la protección de datos y la privacidad en el ámbito de internet, la tecnología y las redes sociales. Estas directrices son fundamentales para establecer normas claras y salvaguardar la información no estructurada, como audio y video, así como los datos personales asociados. Algunos aspectos relevantes que se regulan incluyen:

- Toma de decisiones automatizadas y elaboración de perfiles: las directrices proporcionan orientación sobre cómo cumplir con las obligaciones de protección de datos en relación con estas prácticas. Esto es especialmente importante debido a los posibles riesgos asociados con la automatización de decisiones y la elaboración de perfiles sin intervención humana adecuada.
- Tecnología de reconocimiento facial en la aplicación de la ley: el documento aborda las aplicaciones de esta tecnología, su marco legal y los riesgos asociados. Además, se brindan recomendaciones para garantizar un uso ético y legal, reconociendo la importancia de equilibrar la seguridad con la privacidad y los derechos individuales.
- Vehículos conectados y aplicaciones de movilidad: las directrices se centran en los riesgos de privacidad y protección de datos en este contexto y proporcionan recomendaciones generales para el procesamiento de datos personales. Dado el crecimiento de los vehículos conectados, estas pautas buscan garantizar la seguridad y la privacidad de los usuarios.
- Uso de cámaras de video y procesamiento de datos personales: las directrices establecen requisitos legales y éticos para la recopilación y el procesamiento de datos personales mediante cámaras de video. Se destacan la importancia del consentimiento del sujeto, la protección de los datos y la necesidad de soluciones de respaldo ante posibles fallos técnicos.
- Asistentes de voz virtuales y redes sociales: las directrices abordan la protección de datos personales y la privacidad en el contexto de los asistentes de voz virtuales y las redes sociales. Se proporciona información sobre las obligaciones legales de los proveedores de servicios y los usuarios, así como recomendaciones para garantizar la transparencia, el consentimiento informado y la seguridad de los datos.

Estas directrices son de suma importancia, ya que buscan proteger los derechos de privacidad de los individuos y establecer estándares claros para el manejo de información no estructurada en el ámbito digital. Al abordar preocupaciones como la toma de decisiones automatizadas, la tecnología de reconocimiento facial, los vehículos conectados y las redes sociales. Estas directrices ayudan a

garantizar que las organizaciones cumplan con las regulaciones de protección de datos y protejan la privacidad de los individuos en un entorno tecnológico en constante evolución.

Además, las directrices reconocen la importancia de la anonimización de datos como una medida para proteger la privacidad y la seguridad de los datos personales. La anonimización se utiliza para eliminar información identificable y garantizar que los datos ya no estén sujetos a las regulaciones de protección de datos. Aunque no se proporcionan detalles exhaustivos sobre las técnicas de anonimización, se mencionan algunas posibles técnicas en relación con diferentes contextos. Algunos aspectos destacados en las directrices con relación a la anonimización de datos son:

- Importancia de la anonimización: la anonimización de datos se reconoce como una medida clave para proteger la privacidad y la seguridad de los datos personales. Se menciona en varias directrices como una forma de eliminar permanentemente la relación con una persona y garantizar que el sujeto de datos ya no pueda ser identificado.
- Eliminación de información identificable: la anonimización se refiere a la eliminación de información que pueda identificar a una persona, lo que implica que los datos ya no se consideren datos personales y, por lo tanto, no estén sujetos a las regulaciones de protección de datos.
- Uso de la anonimización: la anonimización de datos se menciona en relación con varios contextos, como la transmisión de datos de vehículos conectados, la eliminación de datos personales en las redes sociales y el procesamiento de datos de asistentes de voz virtuales.
- Técnicas de anonimización: aunque no se proporcionan detalles específicos en todas las directrices se mencionan algunas técnicas de anonimización que podrían utilizarse, como la eliminación de información situacional, el filtrado de características de voz, la eliminación de información identificable.
- Tipos de información no estructurada que requiere tratamiento: de las directrices se identifican los siguientes escenarios de anonimización de datos no estructurados que requieren la aplicación de estrategias de anonimización:
    - ♦ En los vehículos conectados y aplicaciones de movilidad, los datos deben ser anonimizados antes de transmitirlos fuera del vehículo.
    - ♦ El uso de cámaras de video plantea la necesidad de anonimizar datos personales, como imágenes faciales, señales de voz y patrones de movimiento.
    - ♦ Sobre asistentes de voz virtuales, se discute la dificultad de anonimizar grabaciones de voz, pero se menciona que se están investigando técnicas para eliminar información situacional y anonimizar la voz.
    - ♦ Con relación a la toma de decisiones automatizadas y elaboración de perfiles, se menciona que, en algunos casos, el derecho a la eliminación de datos puede ser efectivo mediante la anonimización.

- En la protección de datos en las redes sociales, se menciona la anonimización de datos como una forma de eliminar permanentemente la relación con una persona y asegurar que el sujeto de datos ya no pueda ser identificado.

El informe anual de la Autoridad de Protección de Datos de los Países Bajos para 2022[38] destaca varias preocupaciones y recomendaciones importantes en el ámbito de la protección de datos personales. El documento resalta la advertencia de la Autoridad sobre los riesgos de privacidad relacionados con el almacenamiento de datos gubernamentales en servicios en la nube, especialmente en servidores de empresas estadounidenses. Se menciona que la Autoridad ha presentado recomendaciones para mejorar las políticas de almacenamiento de datos gubernamentales en la nube y espera que el gobierno las tome en cuenta.

Además, el informe enfatiza la importancia de la claridad en la responsabilidad de la protección de datos personales, señalando que la ley no permite la división de responsabilidad entre múltiples partes. Es crucial definir claramente quién es responsable de la protección de los datos personales y establecer las tareas y las responsabilidades de cada parte involucrada en el proceso. También se destaca la necesidad de que la toelichting (explicación) en los documentos legales sea concreta y clara para evitar confusiones y malentendidos.

Aunque el informe no especifica técnicas específicas para preservar la confidencialidad de las personas, pone un fuerte énfasis en la protección de datos personales y la importancia de una asignación clara de responsabilidades. Estas recomendaciones resaltan la necesidad de garantizar la seguridad y la privacidad de los datos personales en un entorno en constante evolución y subrayan la importancia de tomar medidas concretas para salvaguardar la información personal de los ciudadanos.

2.4. Conclusiones

A partir de la revisión de referentes internacionales se concluye lo siguiente:

- El proceso de anonimización es una herramienta para garantizar la privacidad de los datos y hay que tener presente los riesgos en los que se incurre al aplicar esta metodología ya que no se garantiza totalmente la imposibibilidad de reidentificación de la información. Esto sujeto a factores como el avance tecnológico, entre otros, en este contexto hace necesario fortalecer los procesos de anonimización y realizar el proceso cuantas veces sea necesario para garantizar la protección de datos personales.
- La anonimización puede generar riesgos en la seguridad de la información ya que por su carácter no está definido en algunas leyes. Asimismo, en los procesos de desidentificación se

[38] Disponible en: https://www.autoriteitpersoonsgegevens.nl/documenten/ap-jaarverslag-2022

pierde información importante y aunque sean datos anonimizados existe el riesgo de que se pueda identificar a la persona que los proporciona.

- El Reino Unido pareciera ser el país o uno de los países con mayor desarrollo de información de la anonimización de datos. El Código de práctica gestionando el riesgo de protección de datos es, según el conocimiento de la Oficina del Comisionado de Información, el primer código de práctica en materia de anonimización publicado por alguna autoridad de protección de datos europea. Mediante este documento se pretende dar manejo al riesgo que implica la posibilidad de revelar información privada de las personas.
- El Reino Unido está poniendo cada vez más datos en el dominio público. Sin embargo, también existe el riesgo de que se pueda reconstruir una imagen de los individuos de sus vidas privadas también. Con cantidades cada vez mayores de personal información en el dominio público, es importante que las organizaciones tengan un enfoque estructurado y metódico para evaluar los riesgos. Este código de prácticas trata sobre la gestión de ese riesgo y no es un manual de ingeniería de seguridad, ni tampoco cubre todas las técnicas de anonimización, pero contiene consejos claros y prácticos y la explicación de algunos conceptos legales complicados y será de utilidad para la libertad de información, para los profesionales que se dedican a la protección de datos y para todos aquellos que están contribuyendo a la creación de una de las economías más transparentes y responsables del mundo.

2.5. Recomendaciones

A partir de la revisión de referentes internacionales se concluye lo siguiente:

- Teniendo en cuenta que los ataques pueden cambiar con el paso del tiempo es necesario revisar las soluciones pensando en estos riesgos. Esto involucra considerar nuevos datos publicados que constituyan un nuevo riesgo de reidentificación. En este contexto, se sugiere revisar los lineamientos de desidentificación de manera constante para reducir dichos riesgos.
- Siempre es importante tener en cuenta el escenario en que van a ser utilizados los datos, para poder evaluar los riesgos de que sean compartidos y con ello la rigurosidad del proceso. Pues, aun cuando la base no sea compartida públicamente, también se debe llevar a cabo este ejercicio de preservación de datos sensibles: es una práctica deseable sin importar el nivel de confidencialidad del ámbito.
- Se recomienda incluir conceptos asociados con la medición de calidad de los datos en la actualización de la Guía para la anonimización, con el propósito de establecer mejores lineamientos de caracterización y utilidad de la información.
- Se recomienda focalizarse en tener la información que sea necesaria y suficiente para el análisis y no acumular información de más, para que los procesos sean más confiables y organizados.

Revisión de
REFERENTES
INTERNACIONALES

3.

Definición de datos no estructurados

3. Definición de datos no estructurados

3.1. Resumen

El SEN tiene como propósito presentar información oficial, comprensiva y confiable que pueda servir como lineamiento en la toma de decisiones por los productores de información estadística, es así como el DANE en su rol de coordinador del SEN desarrolla metodologías y guías que permiten a los usuarios realizar procesos de producción estadística, analítica de datos, desidentificación de fuentes de información, etc.

Actualmente, el DANE está realizando la actualización de la Guía de anonimización de datos que fue publicada en el 2018, en esta ocasión la guía contendrá no solo aspectos relacionados con datos "estructurados", sino que también incluirá la temática de datos "no estructurados". Por lo cual, es necesario establecer los conceptos que se deben manejar desde el punto de vista técnico entre las entidades del SEN que permitan un lenguaje común. Para ello, es indispensable contar con un concepto estandarizado de "datos no estructurados" que se incluya en la actualización de la guía y que permita a las entidades del SEN identificar, caracterizar y aplicar las técnicas adecuadas para los diferentes tipos de información en sus procesos de anonimización.

3.2. Síntesis de hallazgos

A continuación, la Tabla 5 se presenta una breve descripción de los principales hallazgos de la revisión de referentes internacionales sobre la definición de datos no estructurados. Se revisaron ocho referentes internacionales, uno oceánico, un organismo internacional, dos norteamericanos y cuatro europeos.

Tabla 5. Principales hallazgos sobre concepto de datos no estructurados

REFERENTE	ENTIDAD CONSULTADA	DESCRIPCIÓN
Estados Unidos	Resources.data.gov[39]	Datos que tienen una forma más libre, como archivos multimedia, imágenes, archivos de sonido o texto no estructurado. Datos no estructurados no sigue necesariamente ningún formato o secuencia jerárquica, ni sigue ninguna regla relacional. Los datos no estructurados se refieren a masas de (generalmente) información computarizada que no tienen una estructura de datos que una máquina pueda leer fácilmente. Los ejemplos de datos no estructurados pueden incluir audio, video y texto no estructurado, como el cuerpo de un correo electrónico o un documento de

[39] Disponible en https://resources.data.gov/glossary/unstructured-data/

REFERENTE	ENTIDAD CONSULTADA	DESCRIPCIÓN
		procesador de texto. Las técnicas de minería de datos se utilizan para encontrar patrones o interpretar de otro modo esta información.
Estados Unidos	Government Operations Agency[40]	Los datos no estructurados (o información no estructurada) son información que no tiene un modelo de datos predefinido o no está organizada de una manera predefinida, como un archivo sin formato. La información no estructurada suele tener mucho texto, pero también puede contener datos como fechas, números y hechos.
Estados Unidos	National Center for Biotechnology Information[41]	Los datos no estructurados se definen como grandes volúmenes de información, de poblaciones mucho más grandes. Dichos datos se componen de toda la información en poder de un servicio específico o base de datos de las interacciones de salud de un individuo durante un período de tiempo y pueden incluir el número de visitas, las prescripciones de medicamentos, las notas clínicas no estructuradas, la información demográfica, los datos de salud física y los registros hospitalarios. Aunque los datos no estructurados también pueden contener algunos datos estructurados, como información psicométrica, son solo una pequeña parte de los registros clínicos dentro de estos estudios y los datos no estructurados están en gran parte desorganizados. Los datos no estructurados se incluyen comúnmente en registros de salud electrónicos (EHR) y encuestas de población grandes. Por lo tanto, los datos estructurados están dirigidos y son específicos para el riesgo de suicidio mediante el uso de cuestionarios psicométricos estandarizados más fáciles de interpretar, mientras que los datos no estructurados contienen información potencialmente menos dirigida, lo que sugiere un punto de comparación entre estos dos grupos. • Los datos son una mezcla de estructurados, semiestructurados o no estructurados y no tienen que estar organizados. • Los datos a menudo se encuentran en grandes sistemas de registro que se han recopilado a lo largo del tiempo. • Interacción de persona a computadora, el análisis se puede realizar sin la participación del cliente. • La recopilación de datos es rápida y puede extraerse de grandes encuestas y datos de registro.

[40] Disponible en https://data.ca.gov/pages/open-data-glossary

[41] Disponible en https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9378826/

REFERENTE	ENTIDAD CONSULTADA	DESCRIPCIÓN
España	Kyocera Cloud Information Manager (KCIM)[42]	Los datos no estructurados, generalmente son datos binarios que no tienen estructura interna identificable. Es un conglomerado masivo y desorganizado de varios objetos que no tienen valor hasta que se identifican y almacenan de manera organizada. Una vez que se organizan, los elementos que conforman su contenido pueden ser buscados y categorizados (al menos hasta cierto punto) para obtener información. Esta es una lista limitada de tipos de datos no estructurados: •Correos electrónicos. •Archivos de procesador de texto. •Archivos PDF. •Hojas de cálculo. •Imágenes digitales. •Vídeo. •Audio. •Publicaciones en medios sociales.
España	Dell[43]	Los datos no estructurados son, esencialmente, cualquier cosa que no esté en una base de datos estructurada: todo, desde el correo electrónico, las imágenes y los documentos, hasta los videos, el contenido de redes sociales y los datos relacionados con las aplicaciones, como los registros.
Francia	Periódico Dunet[44]	Los datos no estructurados son datos representados sin un formato predefinido que facilitaría su acceso y procesamiento. En la era de los grandes datos, estos son, con mucho, los datos más abundantes. Los datos no estructurados, que se presentan en forma de archivos de texto, fotos, audio y video, pueden ser generados por humanos (publicaciones en redes sociales, por ejemplo) o por máquinas con el auge de la Internet de las cosas Objetos (IoT).
Francia	Lemagit[45]	Los datos no estructurados son una designación genérica que describe cualquier dato fuera de un tipo de estructura. Los datos textuales no estructurados son generados por correos electrónicos, presentaciones de PowerPoint, documentos de Word o incluso software de colaboración o mensajería instantánea.

[42] Disponible en https://www.kyoceradocumentsolutions.es/es/smarter-workspaces/insights-hub/articles/diferencia-entre-datos-estructurados-y-no-estructurados.html#:~:text=Los%20datos%20no%20estructurados%2C%20generalmente,y%20almacenan%20de%20manera%20organizada.

[43] Disponible en https://www.dell.com/es-es/dt/learn/data-storage/unstructured-data.htm

[44] Disponible en https://www.journaldunet.fr/web-tech/guide-du-big-data/1516849-donnees-non-structurees-comment-ca-marche/

[45] Disponible en https://www.lemagit.fr/definition/Donnees-non-structurees

REFERENTE	ENTIDAD CONSULTADA	DESCRIPCIÓN
		Los datos no estructurados no textuales, por otro lado, se generan a través de medios como imágenes JPEG, archivos de audio MP3 o archivos de video Flash. Sin administración, el gran volumen de datos no estructurados generados anualmente dentro de una organización puede resultar costoso en términos de almacenamiento. Y los datos no administrados también pueden plantear un problema de responsabilidad, por ejemplo, si la información no se puede ubicar en el contexto de una auditoría de cumplimiento o una acción legal. La información contenida en los datos no estructurados no siempre es fácil de localizar. Esta localización implica que los datos presentes en los documentos, tanto electrónicos como físicos, sean digitalizados para permitir que una aplicación de búsqueda extraiga conceptos de ellos mediante el análisis según los términos utilizados en contextos específicos. Este proceso se llama búsqueda semántica.
Organismo internacional	Eurostat[46]	Datos que no están ordenados en columnas y filas, no están debidamente titulados o identificados; datos para los que es necesario aplicar herramientas para categorizar, etiquetar, agrupar, identificar elementos clave y hacer correlaciones para obtener información de ellos; el contenido de los datos no estructurados se refiere a correos electrónicos, documentos y otros objetos que se componen de texto de seguimiento libre.
Canadá	Statistics Canada[47]	Datos no estructurados o semiestructurados: datos que se almacenan sin un modelo de datos que permita entender cómo están organizados o su contenido, como correos electrónicos, páginas web, redes sociales, informes, imágenes y audio. Ejemplos de datos no estructurados o semiestructurados en DND/CAF son las imágenes de las "cámaras de combate" o los vídeos de noticias del Equipo de Defensa. Los datos no estructurados son todos aquellos que no están ordenados según un modelo predefinido. Para producir información estadística basada en datos no estructurados, es

[46] Disponible en https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiC8qOBs4SAAxWXlWoFHetdCKAQFnoECA8QAQ&url=https%3A%2F%2Funstats.un.org%2Funsd%2FChina_UNSD_Project%2Fseminars%2F2013_TianjinChina%2FSession%25207%2FS7-3%2520EuroStat-meeting%2520challenges%2520of%2520unstrctured%2520data.ppt&usg=AOvVaw2nGdzu3iNyyAXe6niTF6Nv&opi=89978449

[47] Disponible en https://www.canada.ca/en/department-national-defence/corporate/reports-publications/data-strategy/annex-a-definitions.html

REFERENTE	ENTIDAD CONSULTADA	DESCRIPCIÓN
		necesario un procesamiento adicional para organizar la información contenida en los datos. Se presentan ejemplos de transformación de texto, imágenes y sonidos en datos estructurados que pueden utilizarse para el análisis de texto y el reconocimiento de patrones y del habla. **Ejemplo 1)** **Datos no estructurados:** un texto. **Procesamiento:** análisis sintáctico para dividir el texto en una lista de palabras, agregación para contar cuántas veces aparece la misma palabra y el uso de diccionarios y reglas para clasificar las palabras. **Datos estructurados:** una hoja de cálculo: en cada fila hay una palabra distinta, las tres columnas presentan la palabra, el número de apariciones y la categoría de la palabra. **Ejemplo 2)** **Datos no estructurados:** una imagen. **Procesamiento:** asignación de valores RGB a los píxeles; segmentación de la imagen en bloques de píxeles en función de los componentes rojo (R), verde (G) y azul (B). **Datos estructurados:** una base de datos: cada registro es un grupo de píxeles y las variables resumen los componentes de color de cada grupo. **Ejemplo 3)** **Datos no estructurados:** un registro de la voz de alguien. **Procesamiento:** segmentación del registro en sonidos distintos; medida de la duración y las frecuencias. **Datos no estructurados:** lista de segmentos con duración y frecuencias.
Canada	Bank of Canada[48]	Los datos no estructurados, que constituyen el 90% restante de los macrodatos, incluyen correos electrónicos, tweets, mensajes de Facebook, información sobre el tráfico rodado y datos audiovisuales. Los almacenes de datos tradicionales se tensan bajo la carga de datos no estructurados y normalmente no pueden procesarlos.
Alemania	Gaia-X[49]	La definicion de datos, tanto estructurados como no estructurados en Alemania, están definidos por los conceptos de la Union Europea. Datos no estructurados: hacen referencia a volumen de información que no se puede tratar mediante modelo relacional

[48] Disponible en https://www.bankofcanada.ca/wp-content/uploads/2013/08/boc-review-summer13-armah.pdf

[49] Disponible en https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

REFERENTE	ENTIDAD CONSULTADA	DESCRIPCIÓN
		debido a su volumen, velocidad de transferencia y variabilidad sino mediante datos de clave/valor y modelos paralelos o distribuidos de tipo MapReduce. Proyecto Gaia-X: el objetivo es una infraestructura de datos segura y federada que represente los valores europeos, la soberanía digital de los propietarios de los datos, la interoperabilidad de las diferentes plataformas y el código abierto. Dentro de este ecosistema será posible proporcionar, compartir y usar datos dentro de un entorno confiable. Por lo tanto, estimular la innovación y crear valor agregado para la economía de datos para todos los que comparten datos.
Reino Unido	Indeed[50]	Los datos estructurados están menos definidos que los datos estructurados. Se pueden almacenar estos datos en su formato nativo y procesarlos solamente cuando estén listos para su uso. Carecen de una estructura consistente, lo cual lo representa un reto para que los programas los lean, entiendan y analicen. Usualmente, la experiencia en la ciencia de los datos es necesaria para leer y entender correctamente los datos no estructurados. Se puede almacenar este tipo de datos en bases de datos, aunque es su falta de consistencia lo que hace difícil atraer las percepciones en la forma en la que se puede realizar con los datos estructurados. Los datos no estructurados provienen de diferentes fuentes, incluyendo: • Páginas web. • Videos. • Comentarios en publicaciones y blogs de diferentes fuentes. • Respuestas de encuestas. • Documentos. • Imágenes. • Registros de chats. • Textos no estructurados.
Nueva Zelanda	Nec[51]	Los datos no estructurados se clasifican con mayor frecuencia como datos cualitativos, y no se pueden procesar y analizar utilizando herramientas y métodos de datos convencionales. • Ejemplos de datos no estructurados incluyen texto, archivos de video, archivos de audio, actividad móvil, publicaciones en redes sociales, imágenes satelitales, imágenes de vigilancia. • Los datos no estructurados son difíciles de deconstruir porque

[50] Disponible en: https://uk.indeed.com/career-advice/career-development/structured-vs-unstructured-data

[51] Disponible en: https://www.nec.co.nz/market-leadership/publications-media/what-is-structured-data-vs-unstructured-data/

REFERENTE	ENTIDAD CONSULTADA	DESCRIPCIÓN
		no tienen un modelo de datos predefinido, lo que significa que no se pueden organizar en bases de datos relacionales. En cambio, las bases de datos no relacionales o NoSQL son la mejor opción para administrar datos no estructurados. • Extraer información enterrada dentro de datos no estructurados no es una tarea fácil. Se requiere de análisis avanzados y un alto nivel de experiencia técnica para penetrar realmente en los datos y extraer información valiosa. Tal análisis de datos puede ser costoso para muchas empresas. • Ventajas: aquellos capaces de aprovechar datos no estructurados tienen una ventaja competitiva proporcionando una comprensión mucho más profunda del comportamiento y la intención del cliente. Los datos no estructurados también se acumulan a velocidades mucho más rápidas y se pueden almacenar en lagos de datos en la nube que permiten cantidades masivas de almacenamiento de datos. • Desventajas: los costos asociados al análisis de datos no estructurados es uno de los mayores inconvenientes. Idealmente, se necesitaría un científico de datos especializado para maximizar las oportunidades que presentan los datos no estructurados y de alguien que interprete los dato. Además, es probable que también se necesite invertir en herramientas especializadas para el análisis de datos que son costosas. Si bien los costos están bajando, aún pueden ser prohibitivos para muchas empresas.
Nueva Zelanda	Informática[52]	Los datos no estructurados son datos empresariales no transaccionales, cuyo formato no pueden ajustarse fácilmente a un esquema de base de datos relacional. Los datos no estructurados incluyen muchas fuentes de información empresarial que, hasta hace poco, no se extraían para la inteligencia empresarial. Estos incluyen archivos de audio, archivos de video, correos electrónicos y documentos de Word, entre otros. En el mundo de Big Data, las organizaciones están prestando más atención a la información oculta en sus datos no estructurados y tomando medidas para comprender y utilizar el contenido de esos datos. Hoy en día, muchas fuentes de información están basadas en texto, con un contexto semántico que no es fácilmente procesado por los sistemas de gestión de datos. El contenido de los correos electrónicos no está estructurado, al igual que los datos de las redes sociales, los podcasts, los videos de seguridad, los archivos PDF, los mensajes de texto y las presentaciones de ventas. Según

[52] Disponible en: https://www.informatica.com/nz/services-and-training/glossary-of-terms/unstructured-data-definition.html

REFERENTE	ENTIDAD CONSULTADA	DESCRIPCIÓN
		algunas estimaciones, del 70 al 80 por ciento de todos los datos empresariales actuales no están estructurados. Mientras que el volumen de todos los datos está aumentando rápidamente, los datos no estructurados son los que más aumentan. Por ejemplo, los datos de las redes sociales pueden ser una gran fuente de información sobre las tendencias y la satisfacción del cliente. Las organizaciones que no cultivan una competencia en la comprensión de sus datos no estructurados se encontrarán rápidamente en una desventaja competitiva.

Fuente: DANE a partir de las revisiones de referentes.

Revisión de
REFERENTES INTERNACIONALES

4.

Segunda cumbre anual sobre el estado de la política de datos abiertos

4. Segunda cumbre anual sobre el Estado de la Política de Datos Abiertos

- **Introducción**

Stefan Raholst, co-fundador de Golfland, presentó el segundo Summit Anual del Estado de la Política de Datos Abiertos, destacando la misión de Golfland de transformar la gobernanza y la toma de decisiones a través de la ciencia y la tecnología. Raholst resaltó la importancia de los datos abiertos para hacer que la toma de decisiones sea más legítima y efectiva e introdujo el concepto de la "tercera ola" de los datos abiertos, enfocándose en la liberación intencional de datos, colaboraciones de datos, responsabilidad de datos y la superación de barreras geográficas. El objetivo de la cumbre fue discutir el estado actual de la formulación de políticas de datos abiertos y la implementación de políticas de datos abiertos existentes, con un programa que incluye discursos principales, paneles con representantes del sector público y discusiones sobre el papel de los datos desde la perspectiva del sector privado.

- **Impulsando el uso de Datos Abiertos en Irlanda**

Barry Lowry habló sobre la respuesta estratégica de Irlanda a la política de datos abiertos, en el cual destaca el enfoque basado en datos de Irlanda para proporcionar mejores servicios a sus ciudadanos y su éxito en la iniciativa de datos abiertos, ubicándose consistentemente entre los tres o los cuatro primeros lugares en Europa. Lowry mostró el portal de datos abiertos de Irlanda, que contiene una amplia gama de conjuntos de datos y publicadores. Él enfatiza la asociación entre el gobierno y la sociedad en la utilización de datos abiertos para crear nuevas áreas de negocio y la importancia de poner a disposición datos de alta calidad para fines de investigación e innovación. Lowry también abordó los planes de Irlanda para clústeres de supercomputación y cómo manejan el acceso a datos que no pueden ser completamente abiertos.

Además, proporcionó información sobre como Irlanda cuenta con una respuesta estratégica a la política de datos abiertos. Destacó los sólidos vínculos de Irlanda con Nueva York y su orientación de alta tecnología, con muchas empresas importantes que tienen sus sedes europeas en el país. Lowry también habló sobre el enfoque basado en datos de Irlanda para proporcionar mejores servicios a sus ciudadanos, con un enfoque en el principio de "una vez por todas" y maximizar el valor de los datos. Mencionó la Ley de Compartición y Gobernanza de Datos, así como el éxito de Irlanda en la iniciativa de datos abiertos, ocupando consistentemente los primeros lugares en Europa.

Por otra parte, habló sobre la asociación entre el gobierno, que proporciona los datos, y la sociedad, que utiliza los datos para crear nuevas áreas de negocio. Enfatizó en la importancia de los datos abiertos para comprender nuestro entorno y sistemas, y mencionó varios conjuntos de datos utilizados, como los datos del censo y los datos a nivel de edificios. Animó a las personas a explorar el portal

nacional de datos abiertos de Irlanda y mencionó la abundancia de recursos disponibles. También mencionó los planes de la Unión Europea (UE) para poner los datos a disposición de otras partes a través de legislación y enfatizó la importancia de poner a disposición datos de alta calidad para fines de investigación y negocios.

Asimismo, presentó sobre la importancia de la formulación de políticas basadas en evidencia y cómo los datos abiertos pueden cambiar las opiniones de la sociedad sobre temas como el cambio climático. Destacó el éxito de Irlanda al utilizar su portal de datos como una forma de interactuar con el público y explicar nuevas intervenciones, como la vacuna contra el COVID-19. Mencionó los esfuerzos de Irlanda para establecer una red de oficiales de datos, proporcionar recursos de aprendizaje electrónico y crear una sólida infraestructura de gobernanza y técnica para respaldar los datos abiertos.

- **Políticas de Datos Abiertos en el Sector Público: Modelos Emergentes de Apertura**

Hilda Hardeman, directora general de la Oficina de Publicaciones de la Comisión Europea, analizó el enfoque de la Comisión Europea hacia la política de datos abiertos. La Comisión opera varios portales de datos, como el Portal europeo de datos abiertos, que ofrece 1.6 millones de conjuntos de datos de forma gratuita.

Su objetivo es aprovechar el potencial de los datos abiertos, crear oportunidades de reutilización y construir capacidades y habilidades para comprender los datos. El enfoque de la Comisión se alinea con la "tercera ola" de los datos abiertos, centrándose en el uso práctico y el valor de los datos.

Hilda presentó el enfoque de la Unión Europea hacia los datos abiertos, que se basa en tres aspectos clave: proporcionar un marco regulador y de gobernanza; recopilar y publicar datos para su reutilización, y desarrollar capacidades a través de intermediarios de datos. Además, enfatizó que no se trata solo de poner los datos a disposición, sino también de desarrollar la alfabetización de datos y mostrar casos de uso. Destacó la importancia de respetar los valores europeos, como el estado de derecho, los derechos humanos, la equidad y la transparencia.

Por otra parte, Otavio Moreno, director de Gobierno abierto y transparencia de Brasil, discutió el aspecto vinculante de poner los datos a disposición de forma gratuita. Los Estados miembros y otros actores están obligados a proporcionar los datos en un formato legible por máquina y para su descarga masiva. La Comisión Europea, apoyada por un equipo de expertos, supervisará la implementación de estos requisitos. La Oficina de Publicaciones de la UE también desempeña un papel en la divulgación de los conjuntos de datos y la compartición de casos de uso para fomentar su aplicación práctica.

Asimismo, se destacó el cambio de perspectiva hacia los datos abiertos, centrándose no solo en su publicación, sino también en fomentar su uso y crear impacto. Esto requiere la construcción de ecosistemas sólidos que involucren a agencias gubernamentales, sociedad civil, academia y sector privado, así como la colaboración entre diferentes niveles de gobierno en una federación como Brasil.

Se mencionó la necesidad de comprender qué tipo de información se necesita y está disponible en la comunidad, la academia y el sector privado. También destacó el cambio de enfoque de centrarse únicamente en los datos estructurados a incluir datos no estructurados, como imágenes satelitales y documentos escaneados. Otavio hizo énfasis en el desafío de la transparencia por diseño y la necesidad de pensar en cómo diferentes tipos de datos pueden usarse para el desarrollo de otros procesos.

Adicionalmente, la demanda de datos abiertos en diferentes sectores es alta, lo que también ha generado mayor cultura de compartir información, mientras que los sectores con menos disponibilidad de datos tienen una demanda menor. También destacó la creciente demanda de datos por parte del sector privado, especialmente para comprender perfiles de empresas y planificar estrategias comerciales. Sin embargo, señalaron que es un desafío hacer un seguimiento de quién está utilizando los datos y con qué propósito. Sugieren tener interactividad con los datos con el fin de proporcionar mejores conocimientos sobre las preguntas que las personas intentan responder.

Tanto Hilda como Otavio consideran que los diversos tipos de datos que la UE publica y pone a disposición del público, esto incluye resultados de investigación financiados por la UE, datos de observación de la Tierra y datos relacionados con políticas de la UE, como el clima, la energía y el empleo. Se resaltó la importancia de hacer que estos datos sean accesibles para todos y mencionó que ciertos conjuntos de datos, como los datos de residuos, son particularmente populares entre los usuarios. Además, se destacó que la cantidad de usuarios no es la única medida de éxito, ya que incluso si un conjunto de datos es utilizado por un número reducido de personas, aún puede tener un alto impacto si la persona o el proyecto adecuado lo utiliza.

Otavio también discutió sobre la necesidad de crear una cooperación global en el establecimiento de estándares para la publicación de diferentes tipos de conjuntos de datos. Destacó la importancia de comprender las personas que acceden a los datos abiertos, ya que diferentes temas atraen a diferentes grupos de usuarios, como la academia para los datos de salud y las organizaciones de la sociedad civil para los datos de compras gubernamentales.

- **Datos abiertos y colaboración de datos: datos privados para el bien público**

Se realizó un análisis del estado actual del acceso a los datos del sector privado y su relevancia en diversos sectores. Se hizo hincapié en la importancia de hacer accesibles y disponibles tanto los datos públicos como los privados con el fin de fomentar el bienestar de la sociedad en general. El enfoque se centra en los modelos emergentes de colaboración de datos en el sector público y en las estructuras de gobernanza necesarias para garantizar un intercambio de datos sostenible y responsable. La discusión comienza con una actualización sobre el progreso del acceso a los datos de salud con fines de salud poblacional, reconociendo tanto los avances realizados como los desafíos que aún existen en este campo.

Uno de los temas discutidos es la utilización de datos abiertos en el sector de la salud. Se destacó que una gran cantidad de datos permanece sin utilizarse debido a la existencia de regulaciones y la falta de estándares comunes. Sin embargo, cada vez existe un mayor consenso en que la recopilación, la integración y la compartición de datos personales de salud puede conducir a una atención médica personalizada.

Se resaltó la necesidad de políticas, marcos e infraestructuras que garanticen la calidad y la consistencia en el intercambio de datos entre organizaciones y geografías. También se mencionó la importancia del uso primario de los datos, como los diagnósticos de investigación, y la disposición de los pacientes a compartir sus datos para el progreso médico. Sin embargo, se hizo hincapié en la transparencia, la privacidad y las preocupaciones de seguridad de los datos entre los ciudadanos. Se mencionó como ejemplo la Iniciativa de Medicamentos Innovadores de la UE, que es un gran proyecto de colaboración público-privada en la utilización de grandes datos en el ámbito de la salud.

Por otra parte, se abordó el tema de los intermediarios en el mercado de datos abiertos, dada la importancia de la confianza por parte de las partes interesadas, incluyendo al gobierno y a los posibles usuarios de los datos. Se explicó que ser una organización sin ánimo de lucro ayuda a generar confianza hasta cierto punto, aunque también puede dificultar la colaboración con entidades comerciales que pueden dudar de la capacidad de producir un producto o servicio oportuno y de alta calidad.

Otro punto que presentaron es la importancia de los productos de datos, porque simplemente poner los datos a disposición sin saber quién los está utilizando puede dificultar la medición de su impacto. Se utiliza la metáfora de las redes de carreteras para ilustrar que, si bien puede ser difícil medir el impacto de las carreteras en funcionamiento, se reconoce ampliamente su utilidad y que valen la pena.

Sugiere el uso de almacenamiento de objetos en los principales proveedores de servicios en la nube como solución para hacer que los datos sean accesibles. Asimismo, la importancia de una buena gobernanza y el papel de los administradores de datos para garantizar la precisión, la confiabilidad y la procedencia de los datos.

La discusión también abordó la importancia de la gestión de datos y los administradores de datos en la garantía de la calidad de los datos abiertos. Se destaca la necesidad de considerar las implicaciones de privacidad y conectar a los equipos de datos con otros equipos relevantes dentro de la organización. La pregunta más común que reciben es cómo proteger la privacidad y garantizar el uso responsable de los datos. Se enfatiza la importancia de definir los objetivos y los problemas a resolver con los datos para generar confianza y garantizar una colaboración efectiva.

- **Reflexiones finales**

Como cierre del evento, Stefan Raholst presentó un breve resumen de los cuatro elementos importantes discutidos durante la Cumbre Anual del Estado de la Política de Datos Abiertos:

- El primer elemento es la importancia de la cultura, tanto en términos de culturas en competencia (pública y privada) como de la necesidad de tener una visión.
- El segundo elemento se centra en los procesos, los roles y la necesidad de nuevos profesionales para escalar las iniciativas de datos abiertos.
- El tercer elemento es optimizar el acceso a los datos para áreas prioritarias que beneficiarían a la sociedad.
- El cuarto elemento es la necesidad de establecer redes y mantener una conversación global para romper los silos y aprender de diferentes sectores. La idea principal es la pasión y el compromiso de todos los involucrados en el avance de los datos abiertos.

En la preparación del Reporte de esta edición participamos los siguientes funcionarios:

Alexander González Coca – agonzalezc@dane.gov.co

Alexandra Jane Simpson – ajsimpsons@dane.gov.co

Catherine Avila Alvarado – jcavilaa@dane.gov.co

Gildardo Andrés Vargas – gavargasa@dane.gov.co

Julián David García – jdgarciag@dane.gov.co

Laura Esperanza Beltrán Cardozo – lebeltranc@dane.gov.co

Mónica Andrea Quiroga Rivera – maquirogar@dane.gov.co

Yinneth Mahecha Monsalve - ymahecham@dane.gov.co

Diana Marcela Pinzon Topia - dmpinzont@dane.gov.co

Revisión de estilo por: Sonia Naranjo - smnaranjom@dane.gov.co

Revisión de contenido por: Julieth Alejandra Solano Villa - jasolanov@dane.gov.co

Si tiene dudas comentarios o aportes sobre esta edición por favor no dude en comunicarse a los correos: maquirogar@dane.gov.co y jcavilaa@dane.gov.co

@DANE_Colombia

/DANEColombia

/DANEColombia

@DANEColombia

www.dane.gov.co