DANE
INFORMACIÓN PARA TODOS

REVISIÓN DE REFERENTES INTERNACIONALES

El futuro es de todos
Gobierno de Colombia

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

REVISIÓN DE REFERENTES INTERNACIONALES

(1) Implementación de instrumentos, software, aplicativos o sistemas que permitan la anonimización de bases de datos.
(2)Recomendaciones que ofrecen los referentes internacionales respecto a la gobernanza y ética de los datos e interoperabilidad en torno al sistema estadístico nacional.
(3) Conceptos estandarizados de los métodos de recolección de información

Mayo de 2022

Tabla de contenido

Lista de tablas

Lista de ilustraciones

Introducción

El presente reporte de revisión de referentes internacionales hace parte de una iniciativa de la Dirección de Regulación, Planeación, Estandarización y Normalización (DIRPEN) emprendida desde junio de 2021 con el objetivo de apoyar el conocimiento, la generación de capacidades, brindar recomendaciones y propiciar acciones frente a temáticas estratégicas del Departamento Administrativo Nacional de Estadística (DANE) y del Sistema Estadístico Nacional (SEN). 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 establecidas.

Para tal fin, la revisión de referentes constituye una investigación prospectiva de la práctica internacional, en función del tema de análisis, de organizaciones de diferente naturaleza y rol en un ecosistema de datos estadísticos, incluyendo: institutos u oficinas nacionales de estadística, ministerios, organismos, organizaciones no gubernamentales e instituciones académicas o de investigación. Los temas que se abordan en cada reporte mensual se priorizan, considerando la urgencia de la necesidad, de una lista de temas construida a partir de la consulta directa realizada a los directivos del DANE, directores técnicos y coordinadores de las mesas estadísticas del SEN. La profundidad y detalle de las revisiones está asociada a las preguntas clave, perspectivas, alcance y disponibilidad de información.

En esta edición del reporte se abordan tres temas: (1) Implementación de instrumentos, software, aplicativos o sistemas que permitan la anonimización de bases de datos; (2) Principios que rigen el intercambio de información de Registros Administrativos y Fuentes no Estructuradas, ética de interoperabilidad en torno al sistema estadístico nacional y (3) Marco conceptual de los métodos de recolección de información.

Por cada uno de los temas se incluyen, un resumen con la necesidad y objetivo de la revisión, una tabla de síntesis asociada al hallazgo principal o respuesta a la pregunta clave, la revisión de cada referente y las conclusiones en las que se identifican tendencias o buenas prácticas que pueden ser de utilidad para el tema en el DANE y/o el SEN.

1.

Implementación de instrumentos, software, aplicativos o sistemas que permitan la anonimización de bases de datos

1. Implementación de instrumentos, software, aplicativos o sistemas que permitan la anonimización de bases de datos.

1.1 Resumen

El propósito de los Sistemas Estadísticos Nacionales (SEN) es proveer información estadística oficial, relevante, oportuna, confiable y objetiva a los diferentes usuarios, siendo la información el eje fundamental en la toma decisiones políticas, temáticas, técnicas, académicas, entre otras. Para los productores de información, este propósito implica grandes desafíos debido al aumento de la demanda de información desagregada, al mayor uso de microdatos y la necesidad de buscar nuevas fuentes de información.

En Colombia, el SEN establece dentro de sus objetivos, promover entre sus miembros, el acceso y uso de los microdatos para la producción y difusión de estadísticas oficiales, mediante el Decreto 2404[1] de 2019 y el Código Nacional de Buenas Prácticas Estadísticas, que en sus principios 10 y 11 incentiva a los miembros del SEN a i) implementar prácticas que permitan el acceso a las estadísticas y microdatos asociados a todo tipo de usuario con el mayor detalle posible, en diferentes formatos y medios que faciliten la consulta, visualización y uso y ii) establece la necesidad de uso de técnicas para la anonimización de microdatos, garantizando la confidencialidad de la información, reduciendo el riesgo de identificación o localización geográfica de las fuentes empleadas en los procesos estadísticos.

En línea con lo anterior, en el año 2018, el DANE, como ente rector del SEN, presentó la "Guía para anonimización de bases de datos en el Sistema Estadístico Nacional"[2], cuyo objetivo es orientar a los integrantes del SEN sobre el proceso de anonimización de bases provenientes de registros administrativos y operaciones estadísticas, el software estadístico utilizado por la entidad en los procesos de anonimización son Stata, SAS y R.

Debido a los cambios en las dinámicas sociales derivados de la pandemia del COVID-19 y el constante desarrollo de las fuentes de información, sumado al aumento en la demanda de acceso a la misma, el equipo encargado del desarrollo metodológico del proceso de anonimización del DANE, plantea la necesidad de contar con una revisión internacional sobre software, herramientas o sistemas que permitan la anonimización de datos para revisar si el alcance de dichos elementos pueden adaptarse a las operaciones estadísticas desarrolladas actualmente, con el propósito de mejorar el proceso, conocer y adaptar nuevas técnicas/herramientas en el análisis y difusión de datos.

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 la implementación y desarrollo de instrumentos,

1 Art. 2.2.3.1.2. Disponible en https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=104952

2 Disponible en https://www.dane.gov.co/files/sen/registros-administrativos/guia-metadatos.pdf

software, aplicativos o sistemas que permitan la anonimización de bases de datos, se revisaron dos organismos internacionales, ocho países europeos y dos países de América del Norte.

Tabla 1. Principales hallazgos sobre implementación y desarrollo de instrumentos, software, aplicativos o sistemas que permitan la anonimización de bases de datos

Referente	¿Manejan un software, aplicativo o herramienta que permita realizar anonimización? ¿Cuenta con una metodología?
Organizaciones internacionales	CEPAL: La Comisión Económica para América Latina y el Caribe – CEPAL desarrolló el software REDATAM (Recuperación de Datos para Áreas pequeñas por Microcomputador), es un sistema computacional de carácter social e interactivo que facilita el procesamiento, análisis y diseminación web de la información de censos, encuestas, registros administrativos, indicadores nacionales/regionales y otras fuentes de datos, es una herramienta que permite el procesamiento en línea de la información censal y estadística producida por las instituciones gubernamentales, divulgando dicha información de manera segura y sin ningún costo, el usuario interactúa a través de páginas predefinidas seleccionando tabulados e indicadores para luego enviar una solicitud al servidor, posteriormente este devuelve el resultado en forma de tabulado, gráfico o mapa estático; la solicitud del usuario puede tomar formatos como frecuencias, cruces de variables, promedios, conteos o listas por área geográfica. EUROSTAT: Cuenta con un software que permite realizar anonimización, mediante el manual de estadística empresarial europeo hace énfasis en softwares como T-ARGUS y Sdc-Table que permiten la anonimización de información o control de divulgación estadística. Son herramientas que realizan sus procedimientos sobre dos tipos de salidas estadísticas, como los datos confidenciales presentados en tablas y datos confidenciales en archivos de microdatos.
Austria	En la oficina federal estadística de Austria no se establece públicamente el uso de softwares de anonimización específicamente. Sin embargo, la unidad de métodos de la oficina junto con la Universidad de Ciencias Aplicadas de Zúrich participan en la construcción de un paper donde hacen un análisis de una interfaz gráfica llamada ScdApp que tiene como base una herramienta de anonimización de R llamada scdMicro. Este software no solo aplica técnicas de anonimización, sino que facilita todo el proceso de carga y preparación de microdatos, configurando y definiendo una instancia del problema. Adicionalmente, brinda facilidades para crear informes resumido y permite la reproducibilidad del proceso, ya que el código R que se requiere se puede ver y descargar de la interfaz.
España	El ayuntamiento de Barcelona, anonimizó los datos personales que tienen a su disposición para proteger la información sensible de los ciudadanos con un software llamado Nymiz. Este es un software que detecta los datos personales en archivos no estructurados y también en datos estructurados, y anonimiza o seudonimiza de forma reversible o irreversible los datos de acuerdo con las necesidades del tratamiento de la información. Es un programa de pago y su costo depende del volumen de datos a procesar. En la página oficial del INE de España no se encontró información relacionada con software de anonimización utilizado por el Instituto.

Referente	¿Manejan un software, aplicativo o herramienta que permita realizar anonimización? ¿Cuenta con una metodología?
Reino Unido	La guía de control de divulgación de GSS/GSR para tablas producidas a partir de registros administrativos de la Oficina de Estadísticas del Reino Unido describe el enfoque que deben seguir los proveedores de datos al producir resultados estándar y cualquier solicitud ad-hoc basada en un marco general para abordar la cuestión de la protección de la confidencialidad. Esta guía incluye las reglas y métodos que afectan el diseño de la tabla, aquellas que modifican sus valores y aquellas que ajustan los datos antes que las tablas sean diseñadas. Algunos de los métodos recomendados son el redondeo controlado y la supresión de celdas, para lo cual se implementa el software µ-ARGUS (versión 3.5.0). Se destaca el artículo titulado “Una solución de software escalable para anonimizar datos biomédicos de alta dimensión”, publicado en La revista GigaScience de OXFORD Academic, el cual presenta una estrategia para garantizar la privacidad y fomentar la reutilización de datos para conjuntos de datos complejos que contienen una gran cantidad de atributos. Se expone un caso de uso, para mejorar el soporte de conjuntos de datos biomédicos de alta dimensión, a través de la ampliación del software de código abierto ARX.
Canadá	Statistics Canada desarrolló el software de control de divulgación automatizado G-Confind con el fin de proporcionar el nivel adecuado de protección para celdas confidenciales y minimizar la perdida de información. G-Confind es un sistema generalizado que evita la divulgación de información confidencial en datos tabulares por medio del método de supresión de celdas, en el cual se identifican y suprimen las celdas sensibles y complementarias a proteger, además, puede usarse para auditar patrones de supresión de celdas, encontrar agregados confidenciales y datos tabulares redondos.
Finlandia	Statistics Finland basa su metodología de anonimización y protección de datos en un marco legal sustentado en la ley de estadística de Finlandia, cuyas bases normativas son tomadas de la reglamentación establecida por la Unión Europea. La oficina de estadísticas cuenta con aplicativo denominado FIONA, el cual garantiza la anonimización, el uso seguro y el tratamiento de los datos estadísticos recolectados, producidos y establecido por la ONE y demás organismo del gobierno finlandés.
Países Bajos	Para el control de divulgación estadística, Estadísticas de Países Bajos fue pionero en el desarrollo del software µ-ARGUS, el cual es un programa interactivo flexible que guía a los usuarios en el proceso de protección de los datos, basado en el enfoque de gestión de riesgos. Este software incorpora algunos métodos sofisticados tradicionales para la anonimización de microdatos como p. ej. métodos de enmascaramiento perturbativo y métodos de enmascaramiento no perturbativo y adicionalmente incorpora otros como microagregación, PRAM, redondeo, codificación superior e inferior, intercambio de rango y adición de ruido. Una de las buenas prácticas que pueden ser tenidas a partir de la experiencia de Países Bajos es la definición e implementación de un conjunto de métodos y reglas para lograr que los archivos de microdatos sean usados tanto por investigadores como por el público en general.

Referente	¿Manejan un software, aplicativo o herramienta que permita realizar anonimización? ¿Cuenta con una metodología?
Bélgica	En el Instituto Nacional de Estadística de Bélgica no se evidencia públicamente el uso de softwares de anonimización. Sin embargo, You-Get e IBM ofrecen una solución única para las autoridades públicas que buscan anonimizar documentos de forma segura y escalable. Con el respaldo de IBM Cloud Pak for Automation e IBM Cloud Pak for Data, la solución se integra con la infraestructura y los sistemas de TI existentes para permitir a las autoridades públicas anonimizar sus datos de manera eficiente mientras comparten información con sus ciudadanos de manera transparente, amigable para los ciudadanos y confiable.
Estados Unidos	Desde el Instituto Nacional de Estándares y Tecnología tienen el programa de ingeniería de privacidad, en él describen 15 herramientas de desidentificación y el software ARX, estas técnicas son aplicadas a un conjunto de datos para prevenir o limitar los riesgos de la privacidad de los individuos, grupos protegidos y establecimientos, estas técnicas pueden ser introducción de ruido como la privacidad diferencial, el enmascaramiento de datos y la creación de conjuntos de datos sintéticos que se basan en modelos que preservan la privacidad.
Francia	El Instituto Nacional de Estadística y Estudios Económicos de Francia - INSEE – no tiene público el uso específico de software para la anonimización de datos confidenciales, sin embargo, publicó el documento “Gestión de la privacidad de los datos individuales” el cual se centra en las metodologías implementadas en µ-Argus el software de protección de microdatos de las estadísticas oficiales (más usado en Europa) y el paquete sdcMicro del software R. Asimismo, el “Manual de análisis espacial” el cual aborda la confidencialidad de datos espaciales, usa el software de confidencialidad µ-Argus para gestionar el secreto secundario del problema de diferenciación geográfica.
Noruega	Statistic Noruega empleo una nueva metodología de anonimización de datos en el 2015 a la Encuesta de condiciones de vida de 2015, esta consiste en matchind estadístico de datos sintéticos y los datos recolectados. Así pues, esta metodología consiste en crear conjuntos de datos “satélite” que se puedan cruzar con los datos recolectados y de esta manera se eliminen aquellas variables que impiden que la información esté anonimizada.

Fuente: DANE a partir de las revisiones de referentes.

1.3 Revisión de referentes

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

1.3.1 Comisión Económica para América Latina – CEPAL

La Comisión Económica para América Latina y el Caribe – CEPAL desarrolló el software REDATAM[3] (Recuperación de Datos para Áreas pequeñas por Microcomputador), es un sistema computacional de carácter social e interactivo que facilita el procesamiento, análisis y diseminación web de la información de censos, encuestas, registros administrativos, indicadores nacionales/regionales y otras fuentes de datos, a comienzos de 2015 fue lanzada la última versión, denominada “REDATAM7 Fastandfriendly”.

3 Disponible en https://redatam.org

Los países del caribe y algunos territorios generalmente publican reportes tradicionales de censos y algunos han desarrollado herramientas de tabulación en línea usando dicho software, REDATAM Webserver es una herramienta que permite el procesamiento en línea de la información censal y estadística producida por las instituciones gubernamentales, divulgando dicha información de manera segura y sin ningún costo, el usuario interactúa a través de páginas predefinidas seleccionando tabulados e indicadores para luego enviar una solicitud al servidor, posteriormente este devuelve el resultado en forma de tabulado, gráfico o mapa estático; la solicitud del usuario puede tomar formatos como frecuencias, cruces de variables, promedios, conteos, listas por área geográfica, etc. Por ejemplo, en el año 2003, el grupo de Análisis y Liberación de Resultados de Registros Administrativos, del Instituto Nacional de Estadísticas y Geografía – INEGI, utilizó el software REDATAM, versión REDATAMSP+ Process en el procesamiento de bases de datos para las encuestas de Salud, Cultura, Intentos de suicidio y Suicidio. En el año 2005 en Institutó lo utilizó en el conteo de Población y Vivienda, permitiendo el desarrollo y mejoramiento de las estadísticas vitales. Dentro de las funciones de la aplicación web se encuentra la posibilidad de ver la estructura de la base de datos, sus variables y definiciones, además es posible el procesamiento remoto de la base de datos para programar tabulados had-hoc, utilizando la sintaxis de la versión REDATAM+SP.

Motor Redatam Webserver[4]

Cuenta con 2 niveles de acceso, el primer nivel se utiliza como plataforma de desarrollo local y es útil para entrar a varias aplicaciones que se encuentran en desarrollo a partir de una tabla de contenidos, el segundo nivel permite el acceso directamente a la aplicación, utilizando una serie de indicadores disponibles para procesar en línea de acuerdo a unas bases de datos específicas por medio de un archivo controlador llamado “app_main.lnl”. Para desarrollar una aplicación solo es necesario modificar y crear archivos INL. Los archivos HTML se utilizan solo como plantillas con parámetros estándares, solo en aquellos casos que se requiera personalizar alguna plantilla se deberá modificar el archivo HTML correspondiente.

Estructura de una aplicación

Una vez que se ha instalado el motor web de Redatam descomprimiendo el archivo Red_Webserver_V6XXX.exe en el disco C:\servers y se ha conectado con el servidor local IIS o Apache, ya se puede comenzar a desarrollar una aplicación personalizada de acuerdo a las necesidades de información.

En las aplicaciones web de Redatam, los parámetros que rigen la base de datos a leer, la apariencia, el contenido, la estructura y los tabulados a ejecutar se definen a través de un lenguaje propio de Redatam Webserver de tipo paramétrico, el cual consta de un número de parámetros definidos en bloques en un archivo de tipo ASCII con la extensión INL. Toda aplicación tiene un archivo maestro del cual se desprenden otros archivos INL complementarios. Se propone agrupar en los archivos complementarios parámetros de un mismo tipo para simplificar su definición y localización; a los bloques de parámetros se les denomina nodos, cada nodo tiene un nombre propio irrepetible, el cual se define bajo corchetes; para activar los parámetros, es necesario hacer la llamada al bloque

4 Disponible en https://redatam.org/cdr/manuales/webserver/esp/RedWeb-05-LenguajeINL-Esp.pdf

específico haciendo referencia al nombre que se le dio al bloque. Bajo cada sección o nodo se definen controles o parámetros los cuales representan las propiedades que componen el grupo, y que deben escribirse en letras mayúsculas y cada uno de estos en líneas separadas.

Los parámetros de un proceso son llamados controles. Hay controles específicos de cada proceso, llamados controles propios, y controles comunes, que pueden ser usados en más de un proceso. Por ejemplo, los procesos pueden usar un filtro para seleccionar los casos, una selección geográfica o la definición del factor de ponderación a ser usado.

Procesos básicos y de indicadores requieren una base de datos con los datos originales de la fuente, a partir de los cuales se realiza el proceso y cálculo según el requerimiento. Se utilizan las mismas tablas que existen en la versión PC de Redatam: AREALIST (Cuadro que entrega la distribución de las categorías de una variable seleccionada dado un nivel geográfico de salida), AVERAGE (Entrega el promedio de la variable seleccionada, puede combinarse con otras variables en las filas y columna), FREQUENCY (Distribución de frecuencias de una o más variables.

Es similar al proceso Cruz, pero de una sola dimensión), CRUZ (Cruce de variables hasta 5 dimensiones), COUNT (Cuadro que entrega un conteo de elementos dado un nivel geográfico de salida), MEDIAN (Cuadro con la mediana de la variable seleccionada, puede combinarse con otras variables en las filas y columna), MAXIMUM (Cuadro con el valor máximo de la variable seleccionada, puede combinarse con otras variables en las filas y columna), MINIMUM (Cuadro con el valor mínimo de la variable seleccionada, puede combinarse con otras variables en las filas y columnas) y STATS (Cuadro que entrega varios descriptores estadísticos de una variable NO categórica, como, por ejemplo, el valor máximo, mínimo, la sumatoria, la desviación estándar, la amplitud, la moda). Suelen ser procesos muy utilizados para la publicación y procesamiento de microdatos de censos de población.

Procesos de indicadores FRACTION (Cuadro que entrega una razón resultado de la selección del numerador y denominador según un nivel geográfico seleccionado) y QTS (Cuadro que entrega un porcentaje para las categorías seleccionadas de una variable según un nivel geográfico seleccionado) son procesos de razón, que generan una lista de áreas geográficas con resultados porcentuales a partir de la selección de un numerador y denominador. Toma una variable cualquiera y usa como numerador una combinación de categorías de esa variable, y como denominador el número total de casos que respondieron a esa variable. Con el QTS se calcula, por ejemplo, el porcentaje de viviendas sin luz eléctrica.

Procesos de indicadores agregados CNTP (nodos contenedores), es decir, para los procesos de datos agregados, los datos a utilizar ya deben estar calculados, y vienen estructurados en una base de datos Redatam con todas sus desagregaciones: áreas, períodos, indicadores, sexo, edad, zona, etc. A partir de esto, en la aplicación se realiza una búsqueda de datos, según las solicitudes del usuario, extrayendo la información requerida. Generalmente, este tipo de base de datos está relacionada con un sistema de indicadores.

1.3.2 EUROSTAT

La oficina europea de estadística, destaca, mediante el manual de estadística empresarial europeo[5], algunos instrumentos y softwares que permiten la anonimización de información. La oficina lo denomina control de divulgación estadística (SDC, por sus siglas en inglés), y lo declara como un proceso complejo, ya que tiene que garantizar el anonimato de las unidades estadísticas y al mismo tiempo, limitar la pérdida innecesaria de información a través de supresiones/modificaciones excesivas. Se han desarrollado herramientas de control de divulgación estadística para hacer frente a estos problemas. Hay dos familias de herramientas divididas por el tipo de salida estadística:

1. Herramientas que protegen los datos confidenciales presentados en tablas.
2. Herramientas de protección de datos confidenciales en archivos de microdatos.

Herramientas para proteger datos en tablas
Las herramientas estándar para los datos presentados en tablas incluyen tau Argus y sdcTable basado en R. Se precisa que la mayoría de las autoridades estadísticas nacionales estaban familiarizadas con tau Argos[6]. Estas herramientas estándar a menudo se complementan con otras herramientas (SAS, STATA o Excel) y algunos procedimientos manuales.

Metodología
Las herramientas especializadas de control de divulgación estadística identifican las celdas confidenciales primarias de acuerdo con las reglas definidas por el usuario. Cuanta más información se proporciona a la herramienta, mejor se protegen los datos. Idealmente, los datos de entrada son microdatos. Las herramientas identifican celdas confidenciales secundarias según el método elegido; tanto las celdas confidenciales primarias como las secundarias están protegidas con el método seleccionado. Los datos pueden ser protegidos por supresión, redondeo u otros métodos.

Herramientas para proteger archivos de microdatos
Estos incluyen tau Argus y sdcMicro basado en R, estas herramientas se complementan con herramientas estadísticas estándar (como SAS, STATA o SPSS). Las herramientas especializadas de protección de microdatos aplican métodos de control de divulgación estadística sobre los microdatos, en línea con la metodología aplicada a las tablas.

Soluciones de código abierto
Eurostat apoya la migración de herramientas de control de divulgación estadística hacia soluciones de código abierto. Las herramientas de Argus tienen código abierto desde 2015 y códigos Argus desde 2016.

1.3.3 Austria

La unidad de metodologías de la oficina federal estadística de Austria, junto con el instituto de Análisis de Datos y Diseño de Procesos de la Universidad de Ciencias Aplicadas de Zúrich,

5 Disponible en https://ec.europa.eu/eurostat/documents/3859598/12453409/KS-GQ-21-001-EN-N.pdf/f67631e8-c728-e650-d777-de0d9079bf18

6 Según un cuestionario realizado en todo el ESS (en 2016).

realizaron un estudio y análisis sobre el proceso de anonimización de datos en una interfaz gráfica *sdc app* basada en el paquete de programación de R, conocido como sdcMicro, el cual brinda herramientas de software para realizar anonimización de microdatos[7].

En esta publicación que realizan las dos oficinas, liderada por los investigadores Meinfl y Templ realizan una descripción de las facilidades de Sdcapp, presentando la metodología de uso de la interfaz, donde muestra opciones como modificación, análisis de la data, anonimización, definición del problema, anonimización de datos categóricos y de datos continuos. Adicionalmente, los autores hablan sobre el contraste Riesgo/utilidad, asi como de la opción de exportar los datos y la posibilidad de reproducir la información. Paralelamente, realizan un marco comparativo con los softwares más conocidos en materia de anonimización, mostrando ventajas y desventajas de cada uno de ellos.

SdcMicro

Existen varias herramientas de software de anonimización como sdcMicro, el cual es un paquete de R para la anonimización de datos optimizado para grandes conjuntos de datos. Existen dos panoramas 1. Los usuarios que se sienten cómodos con el uso de R y su lenguaje de programación, este paquete proporciona una herramienta para la aplicación de un conjunto integral de métodos comúnmente utilizados y descritos en la literatura sobre control de divulgación y 2. Los usuarios que no se sienten cómodos con R, donde la interfaz es una aplicación de estos métodos que demuestra ser difícil para los no expertos en R creando conjuntos de datos seguros y anónimos.

La GUI[8] interactiva basada en la web de sdcMicro

Los autores declaran que sdcApp o GUI es un software especializado que permite a los no expertos en un software en particular y sin habilidades de programación anonimizar conjuntos de datos. Así mismo, la GUI no se limita a aplicar técnicas de limitación de divulgación, sino que facilita todo el proceso de anonimización. Esta interfaz permite subir datos al sistema, modificarlos y crear un objeto que defina el escenario de divulgación. Una vez que se ha definido un problema de control de divulgación estadística (SDC), los usuarios pueden aplicar técnicas de anonimización a este objeto y obtener comentarios instantáneos sobre el impacto en el riesgo y la utilidad de los datos después de que se hayan aplicado los métodos SDC. Se destaca que sdcApp proporciona la lista más completa de métodos populares.

Metodología de sdcApp

Todo el proceso de anonimización se integra de forma natural en sdcApp basado en la herramienta de anonimización de sdcMicro de R, teniendo como objetivo general mapear e integrar todo el proceso de anonimización de un conjunto de datos.

Tabla 2. Metodología de componentes sdcApp

Incorporación de los datos.
• La incorporación al sistema de los microdatos originales que se catalogan como “posiblemente inseguros”. Donde sdcApp admite la importación y exportación de datos en

7 Meindl, B., & Templ, M. (2019). Feedback-Based Integration of the Whole Process of Data Anonymization in a Graphical Interface. Algorithms, 12(9), 191. Disponible en https://doi.org/10.3390/a12090191

8 Interfaz gráfica del usuario.

varios formatos (STATA, SAS, SPSS y R), adecuados para el uso de usuarios de otro software. • Cualquier conjunto de datos cargado en el espacio de trabajo de R se puede acceder e importar fácilmente en la aplicación.
Fase de inspección y modificación. • En este campo se habilita la posibilidad de mostrar los microdatos, explorar y restablecer las variables, usar un subconjunto de los microdatos, conversión numérica a factor, conversión de variables a términos numéricos, establecer datos jerárquicos, entre otros.
Proceso de anonimización interactivo.
Se selecciona un escenario de riesgo. • Según el escenario, la utilidad de los datos y las medidas de riesgo se actualizan sobre la marcha tan pronto como se aplica una técnica de control de divulgación y/o se modifican los datos. • El usuario puede revertir fácilmente el último paso si el análisis de riesgo o los indicadores de utilidad de datos muestran que la elección de parámetros para el método seleccionado no es óptima. Esto abre la posibilidad de poder probar diferentes valores de parámetros y/o métodos. En cualquier momento, las comparaciones de variables entre el conjunto de datos sin procesar (no seguro) y el estado actual del conjunto de datos se pueden analizar gráficamente utilizando un conjunto de estadísticas adecuadas. • La propia sdcApp tiene acceso a una gran colección de algoritmos discutidos y desarrollados en la literatura académica (incluidos métodos para la recodificación global, la supresión local, la aleatorización posterior, la adición de ruido, la microagregación y muchos más); y todos están optimizados para grandes conjuntos de datos.
Reproducibilidad
Para obtener este paso, se destaca que todos los pasos de anonimización aplicados dentro de la GUI se almacenan internamente. El código R que se necesita para recrear el estado actual, se rastrea internamente y se puede ver en la aplicación y de igual manera se puede descargar a un archivo.
Preparación de informes
La GUI igualmente ayuda a los usuarios y agencias a preparar informes sobre el proceso adecuado para audiencias internas y externas. Dichos informes se pueden crear y descargar desde la interfaz.

Fuente: DANE a partir de las revisiones de referentes

Factores a resaltar de la herramienta

Los autores hacen énfasis en que la GUI en todo el proceso toma nota de todos los datos, incluidos los datos sin procesar originales y posiblemente inseguros, se guardan en el computador local y nunca se cargan en ningún lugar. Por lo tanto, no es un problema de seguridad en absoluto cuando el proceso de anonimización se realiza en una aplicación basada en la web. Sin embargo, este enfoque permite brindar a los usuarios información adicional de fácil acceso. Por ejemplo, en toda la GUI se muestran iconos de signos de interrogación. Al pasar el cursor sobre estos pequeños íconos, se abre una ventana emergente que proporciona textos de ayuda e información útil. Las entradas intuitivas, como los menús desplegables, los controles deslizantes, las casillas de

verificación o los botones, se utilizan dentro de la GUI que, por lo general, no necesitan explicaciones específicas para los usuarios.

La disponibilidad de una GUI para aplicar métodos comunes de anonimización tiene el potencial de reducir las barreras para una mayor cantidad de usuarios, pero también los usuarios experimentados de sdcMicro pueden beneficiarse del uso de la GUI debido a su compatibilidad con la recodificación y la generación de informes de utilidad y riesgo después de cada intervención.

Comentarios sobre otros softwares

- Sobre ***µ-Argus*** destacan que es un software que en el año 2019 seguía en desarrollo y estaba siendo soportado por Statistics Netherlands y otros, donde algunas de las extensiones estaban siendo subsidiadas por Eurostat. El software cuenta con una interfaz gráfica de usuario de apuntar y hacer clic, que se basaba en "Visual Basic" hasta la versión 4.2, y ahora (Versión 5.1 y posteriores) está escrita usando "Java". Actualmente, solo se han creado versiones de 32 bits y no hay una interfaz de línea de comandos disponible. La herramienta utiliza diferentes métodos estadísticos de anonimización, como la recodificación global (agrupación de categorías), la supresión local, la aleatorización, la adición de ruido, la microagregación y la codificación superior e inferior. Cuenta con varias herramientas con una interfaz gráfica de usuario para el cálculo básico de la frecuencia y para garantizar el anonimato k, la diversidad l y cálculos de frecuencia similares.
- ***TIAMAT*** es una herramienta visual que permite a los editores de datos seleccionar una transformación de anonimización k adecuada y sus parámetros correspondientes para proteger sus datos. Esta herramienta no admite métodos adicionales. Respecto a ***PARAT*** (versión 6) se basa únicamente en cálculos de frecuencia. OpenAnonymizer se basa únicamente en los conceptos de k-anonimato y l-diversidad.
- ***SECRETA*** posee características limitadas. Se destaca que no hay una estimación de riesgo interna disponible en esta herramienta.
- ***AMNESIA*** (versión 1.0.6) es una herramienta de anonimización de datos desarrollada en el Centro de Investigación Athena. Tiene un creador y editor de jerarquías para la anonimización. Sin embargo, solo admite k-anonimato y km-anonimato.
- ***Arx*** está implementado en Java (versión actual 3.7.1) y se usa principalmente para datos biomédicos sin estructuras de datos especiales. La anonimización en Arx consta de tres pasos básicos: primero, configurar el proceso de anonimización; segundo, explorar el llamado "espacio de soluciones"; y, por último, analizar los datos perturbados. La herramienta es útil para k-anonimato, l-diversidad y enfoques similares como t-cercanía o δ-presencia. Proporciona funciones interactivas para investigar la pérdida de información basándose en resúmenes univariados de los datos originales y perturbados. Arx tiene más métodos integrados que otras herramientas en el área biomédica. Sin embargo, no puede tratar con datos de diseños complejos y tiene características limitadas aparte de los procedimientos puramente basados en frecuencia.

En este punto se destaca que sdcApp posee muchas más posibilidades de gestionar problemas complejos referentes a anonimización de lo que terminan ofreciendo otras herramientas, como Arx, PARAT, OpenAnonymizer, SECRETA, Amnesia, Cornell Anonymization Toolkit y TIAMAT. A partir

de esta publicación[9] y el análisis exhaustivo que realizan los autores, se resalta que la riqueza de métodos, la reproducibilidad y las características como el botón “deshacer”, junto a las amplias funciones de creación de informes, así como los cambios instantáneos y la visualización de riesgos, utilidades y datos, solo se implementan en sdcApp.

1.3.4 España

El ayuntamiento de Barcelona, en cumplimiento del Reglamento General de Protección de Datos (RGPD), anonimizó los datos personales que tienen a su disposición para proteger la información sensible de los ciudadanos con un software llamado Nymiz[10]. Este es un programa que permite: (i) Proteger los datos de la compañía, (ii) Minimizar los riesgos de robo de información, (iii) Cumplir con la normativa RGPD vigente, y (iv) Aprovechar el volumen de los datos para su análisis. Nymiz es un software que detecta los datos personales en archivos no estructurados y también en datos estructurados, y anonimiza o seudonimiza de forma reversible o irreversible los datos de acuerdo con las necesidades del tratamiento de la información.

Tabla 3. Características de Nymiz

Característica	Descripción
Machine Learning inside (PLN)	Procesamiento de lenguaje natural para mejorar la fiabilidad y el alcance del servicio
Servicio Cloud	Soluciones adaptadas en el entorno local como en la nube
Contenido con información estructurada y no estructurada	Disponible para bases de datos y documentos no estructurados
Reconocimiento multilenguaje	Disponible en inglés y español

Fuente: nymiz.com/es/

Namiz opera con la tecnología PLN (Procesamiento del Lenguaje Natural) y, por medio de algoritmos de inteligencia artificial, amplía la fiabilidad de los resultados analizando el contexto y afinando la detección de cualquier término que pueda considerarse un dato personal. Es un software de pago que ofrece diferentes soluciones dependiendo el tamaño de la empresa. Para las entidades de la administración pública que requieren de un despliegue dentro de su infraestructura, Namiz ofrece una versión especial en donde el cliente obtendrá todas las funcionalidades y ventajas del software SaaS pero en local. Tiene una facturación de licencia anual.

9 Disponible en: https://www.mdpi.com/1999-4893/12/9/191/htm

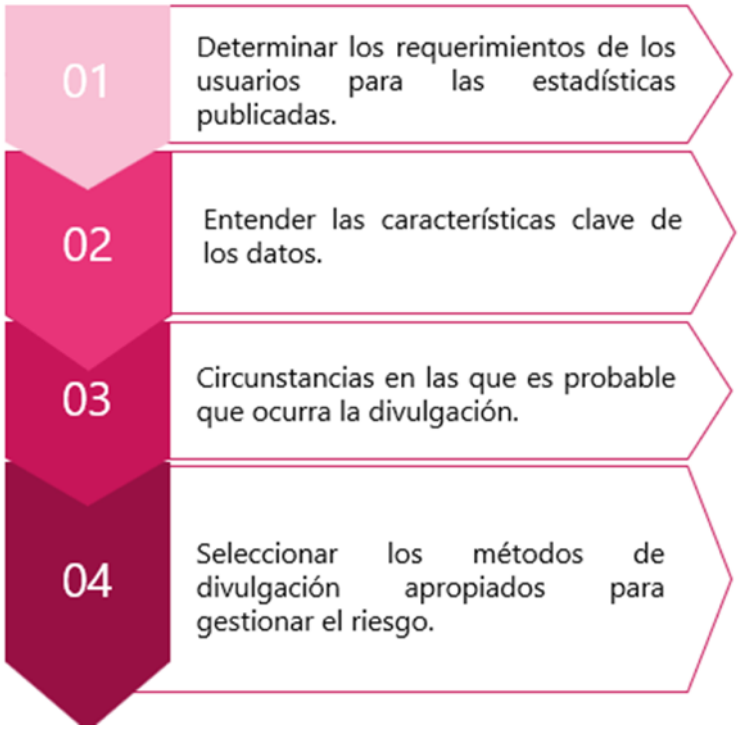
10 Información disponible en https://nymiz.com/es/

1.3.5 Reino Unido

La Oficina de Estadísticas del Reino Unido cuenta con la guía de control de divulgación de GSS/GSR para tablas producidas a partir de registros administrativos[11], la cual describe el enfoque que deben seguir los proveedores de datos al producir resultados estándar y cualquier solicitud ad-hoc basada en un marco general para abordar la cuestión de la protección de la confidencialidad.

La Ilustración 1 presenta los pasos clave que deben ser tenidos en cuenta para implementar procesos de anonimización de datos, garantizando la protección de la confidencialidad.

Ilustración 1. Pasos clave para implementar procesos de anonimización en el Reino Unido



Fuente: DANE basado en la Guía de control de divulgación de GSS/GSR para tablas producidas a partir de registros administrativos.

Adicionalmente, la Tabla 4 sintetiza las reglas y métodos definidos en el Reino Unido para desarrollar procesos de anonimización.

Tabla 4. Selección de reglas y métodos para la anonimización

Categorías	Método	Descripción
Afectan el diseño de la tabla	Rediseño de la tabla	Disfrazar las celdas inseguras por:
		- agrupar categorías dentro de una tabla
		- agregando a una geografía de nivel superior o para un subgrupo de población más grande

11 Disponible en https://gss.civilservice.gov.uk/wp-content/uploads/2018/03/Guidance-for-tables-produced-from-administrative-sources-4.pdf.

		- agregando tablas a través de un
		número de años/meses/trimestres
Modifican los valores en la tabla	Supresión de celdas	Las celdas inseguras no se publican. Se suprimen y se reemplazan por un carácter especial, como '..' o 'X', para indicar un valor suprimido. Tales supresiones se llaman primarias.
		Supresiones. Para asegurarse de que las supresiones primarias no se puedan derivar restando de los totales, puede ser necesario seleccionar celdas "seguras" adicionales para la supresión secundaria.
	Redondeo	El redondeo implica ajustar los valores de todas las celdas de una tabla a una base específica. Esto crea incertidumbre sobre el valor real de cualquier celda y agrega una cantidad pequeña pero aceptable de distorsión a los datos.
Ajustan los datos antes que las tablas sean diseñadas	Intercambio de registros	Intercambie pares de registros dentro de un micro conjunto de datos que coincidan parcialmente para alterar las ubicaciones geográficas adjuntas a los registros, pero deje todos los demás aspectos sin cambios.
	Eliminación de registros arriesgados	Un pequeño número de registros puede ser único en los datos para varias variables. En lugar de proteger las tablas con estas variables, sería más sencillo eliminar el registro

Fuente: DANE basado en la Guía de control de divulgación de GSS/GSR para tablas producidas a partir de registros administrativos.

Para los métodos dentro de la categoría "modificación de los valores de la tabla", si quedan celdas inseguras en la tabulación de salida, se deben considerar métodos de protección adicionales para ocultarlas. Si el rediseño de la tabla no es una solución factible, el método recomendado para la protección posterior para la mayoría de las tablas de frecuencia es el redondeo controlado. Sin embargo, este método requiere un software especializado y, por lo tanto, no siempre será práctico. En algunos casos, si el número de celdas no seguras es bajo, la supresión de celdas puede ser un método alternativo. El redondeo controlado y la supresión de celdas se pueden implementar en el software µ-ARGUS[12].

Adicionalmente, se destaca el trabajo de UK Data Archive, con sede en la Universidad de Essex, es el socio principal Servicio de datos del Reino Unido, proporciona un Repositorio Digital Confiable para recursos de datos nacionales que incluyen muchas encuestas realizadas por la Oficina de Estadísticas Nacionales, centros nacionales de investigación social y datos del Censo.

En la sección de herramientas en uso, se pueden destacar los siguientes servicios:

12 Versión 3.5.0 Disponible en http://neon.vb.cbs.nl/casc

Ilustración 2. Servicios destacados de las herramientas en uso

Fuente: DANE basado en información de herramientas de anonimización de la Universidad de Essex[13]

GigaScience de OXFORD Academic

La revista GigaScience de OXFORD Academic, es una revista científica revisada por pares que se fundó en 2012 en el Reino Unido, tiene como objetivo revolucionar la publicación al promover la reproducibilidad de los análisis y la difusión, organización, comprensión y uso de datos.

La revista publicó un artículo titulado "Una solución de software escalable para anonimizar datos biomédicos de alta dimensión"[14]. El artículo habla sobre la anonimización de datos que es un componente importante para garantizar la privacidad y fomentar la reutilización de datos. Sin embargo, transformar los datos de una manera que preserve la privacidad de los sujetos mientras se mantiene un alto grado de calidad de los datos es desafiante y particularmente difícil cuando se procesan conjuntos de datos complejos que contienen una gran cantidad de atributos. En este artículo, se presenta cómo ampliaron el software de código abierto ARX para mejorar su soporte para conjuntos de datos biomédicos de alta dimensión.

Las versiones de ARX hasta la 3.8.0 solo podían procesar conjuntos de datos con un número limitado de atributos que podían considerarse durante la anonimización (hasta ~15). Hubo 2 razones para esto: (i) el software solo tenía soporte limitado para algoritmos de anonimización capaces de procesar datos de alta dimensión y (ii) la GUI no fue diseñada para trabajar con conjuntos de datos que contienen una gran cantidad de atributos.

Los esfuerzos para superar estas limitaciones se dieron (i) ampliando la interfaz de usuario de ARX con vistas adicionales que simplifican la gestión de datos de alta dimensión, (ii) implementando 2 algoritmos de anonimización heurísticos novedosos y (iii) evaluando los algoritmos novedosos con respecto a su rendimiento para anonimizar conjuntos de datos de baja y alta dimensión. El artículo

13 Disponible en https://www.data-archive.ac.uk/managing-data/digital-curation-and-data-publishing/tools-we-use/

14 Disponible en https://academic.oup.com/gigascience/article/10/10/giab068/6380885

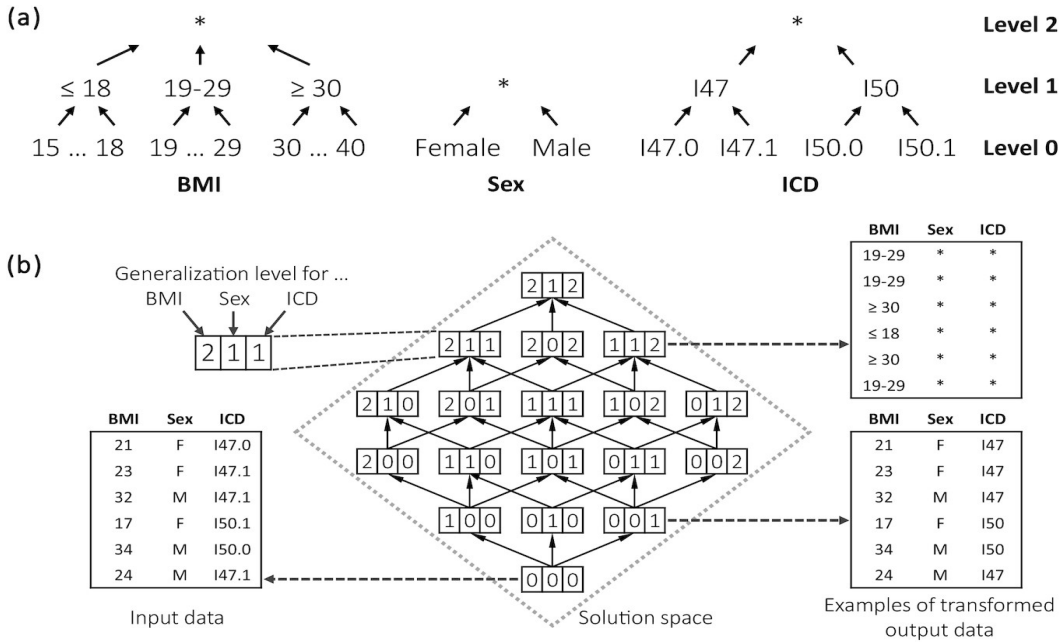
presenta algunas propiedades importantes de la herramienta de anonimización ARX que influyeron en el estudio.

La herramienta de anonimización ARX

ARX admite una variedad de modelos de privacidad, modelos de calidad y esquemas de transformación de datos y permite su combinación arbitraria. Para transformar los datos, se basa en jerarquías de generalización de dominios que describen cómo se pueden transformar los valores para hacerlos menos únicos. Para cada jerarquía es posible definir múltiples niveles de generalización que cubren un rango creciente del dominio del atributo. El espacio de solución básico que utiliza ARX está dado por todas las combinaciones posibles de niveles de generalización definidos por las jerarquías, estas combinaciones se denominan esquemas de generalización.

La Ilustración 3 en la parte “a”, muestra jerarquías de generalización ejemplares para los atributos, índice de masa corporal, sexo y código ICD. En la parte “b” ilustra cómo se estructura el espacio de solución resultante de estas jerarquías y cómo la aplicación de diferentes esquemas de generalización alteraría un conjunto de datos ejemplar.

Ilustración 3. Jerarquías de generalización y la estructura del espacio de solución



Fuente: Gigascience, volumen 10, número 10, octubre de 2021[15]

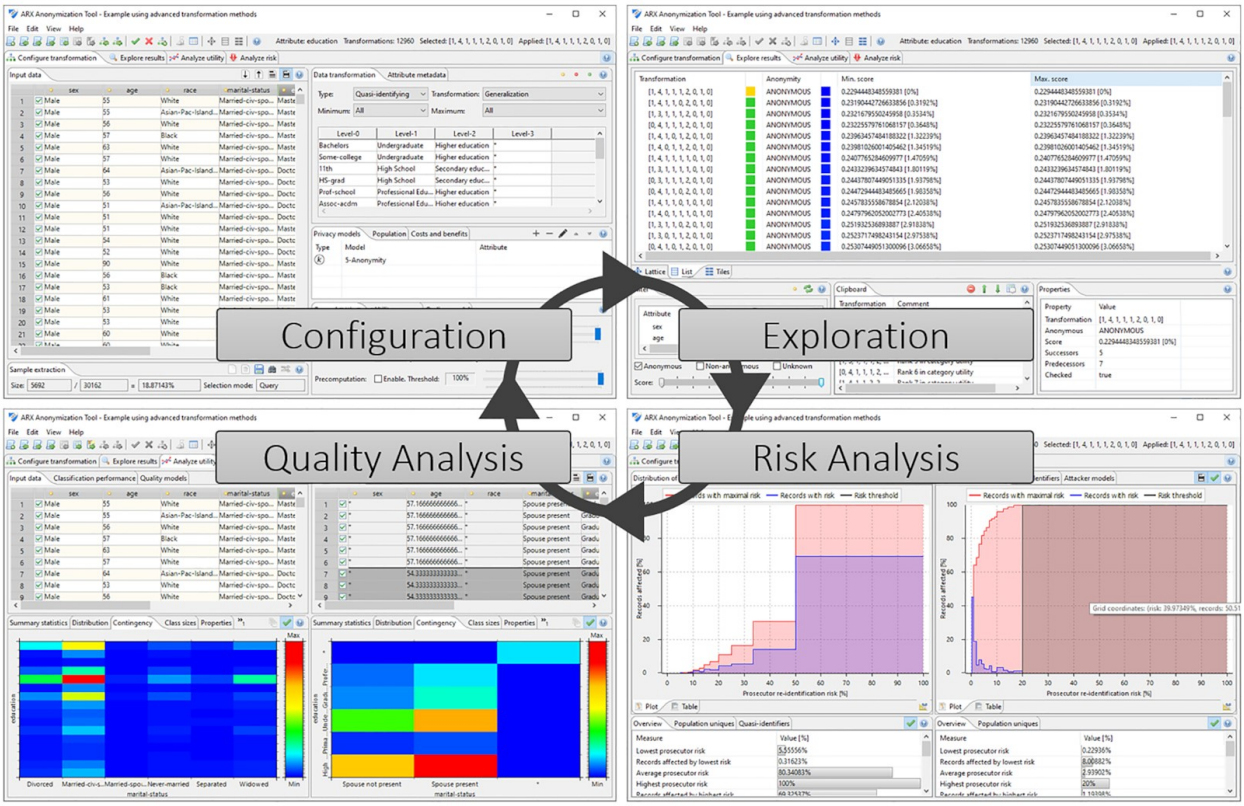
Matemáticamente, el espacio de solución es una red, cuyo tamaño crece exponencialmente de acuerdo con el número de atributos que deben protegerse. Como ARX también puede aplicar diferentes esquemas de generalización automáticamente a diferentes partes del conjunto de datos de entrada, el tamaño del espacio de la solución puede crecer aún más por un factor multiplicativo que representa el número de filas. ARX admite diferentes algoritmos para encontrar soluciones

15 Disponible en https://doi.org/10.1093/gigascience/giab068

óptimas dentro de espacios de solución de tamaño manejable, así como un algoritmo heurístico para espacios de búsqueda más grandes que intenta determinar un buen esquema de transformación sobre la base del mejor esfuerzo.

Además de su motor de anonimización, ARX también cuenta con una GUI multiplataforma. En la Ilustración 4 se muestra una descripción general de las diferentes perspectivas proporcionadas por la plataforma.

Ilustración 4. Perspectivas básicas de la interfaz gráfica de la herramienta de anonimización de datos ARX



Fuente: Gigascience, volumen 10, número 10, octubre de 2021[16]

- En la perspectiva de "configuración" es posible definir umbrales de riesgo para diferentes tipos de ataques, priorizar atributos por importancia, modelar el conocimiento previo de posibles atacantes y definir métodos y reglas de transformación.
- En la perspectiva de "exploración", Se visualizan estrategias de anonimización para los datos de entrada y se admite una categorización según la calidad de los datos de salida.
- En una perspectiva del "análisis de calidad" manual de los datos de salida, se proporcionan diferentes métodos para medir el contenido de información de los datos de salida, estadísticas descriptivas y métodos para comparar la utilidad de los datos de entrada y salida para diferentes escenarios de aplicación.

16 Disponible en https://doi.org/10.1093/gigascience/giab068

- En una perspectiva de "análisis de riesgos", es posible comparar visualmente los datos de entrada y salida utilizando diferentes modelos de riesgo. Sin embargo, en la interfaz de usuario es un desafío admitir conjuntos de datos de alta dimensión. Por ejemplo, varias perspectivas y vistas del software muestran listas de todos los atributos del conjunto de datos cargado, lo que puede volverse confuso y generar problemas de rendimiento en algunas plataformas con un número creciente de atributos.

Integración de algoritmos de anonimización para datos de alta dimensión

Los procedimientos de anonimización admitidos por ARX se basan en un operador básico que busca a través de la red de generalización. Este algoritmo comienza en el esquema de generalización inferior, que no aplica generalización a los datos. Luego "expande" este esquema de generalización aplicando todos los esquemas de generalización al conjunto de datos de entrada que se puede derivar al aumentar uno de los niveles de generalización. La calidad del conjunto de datos de salida resultante se calcula para todos estos esquemas y el proceso se repite expandiendo la generalización, lo que da como resultado el conjunto de datos con la calidad más alta. Este proceso luego se repite hasta que haya pasado un período de tiempo especificado por el usuario. Durante la ejecución del algoritmo, se almacena una lista de todos los esquemas de generalización que se han evaluado y, en cada iteración, se expande el esquema con mayor calidad de datos de salida que aún no se ha expandido.

Cabe señalar que este proceso solo es adecuado para procesar conjuntos de datos de dimensionalidad media (~15 atributos) por varias razones.

- ✓ Primero, el proceso de búsqueda puede quedar atrapado en mínimos locales porque no hay una diversificación significativa de las soluciones consideradas.
- ✓ En segundo lugar, el proceso favorece naturalmente los esquemas de transformación ubicados en la parte inferior del espacio de búsqueda (es decir, esquemas que aplican un bajo grado de generalización). Si bien esto tiene sentido para los procesos de anonimización que solo aplican la generalización, el método alcanza sus límites con las operaciones de transformación complejas admitidas en las versiones más recientes de ARX en las que se utilizan diferentes esquemas de transformación para transformar diferentes partes de un conjunto de datos. En este caso, a veces se puede determinar una mejor solución general si los valores atípicos se transforman con más fuerza.

Se optó por el algoritmo genético porque es una de las metaheurísticas basadas en poblaciones más conocidas. En comparación con los algoritmos basados en una sola solución, los enfoques basados en la población mantienen múltiples soluciones, lo que potencialmente da como resultado un alto grado de diversificación y un menor riesgo de quedarse atascado en los óptimos locales. El algoritmo genético implementado en ARX se basa en el trabajo de Wan et al., él utilizó el algoritmo para la anonimización de datos genómicos utilizando un modelo de privacidad de teoría de juegos, que ya se adaptó e integró con éxito en ARX en trabajos anteriores.

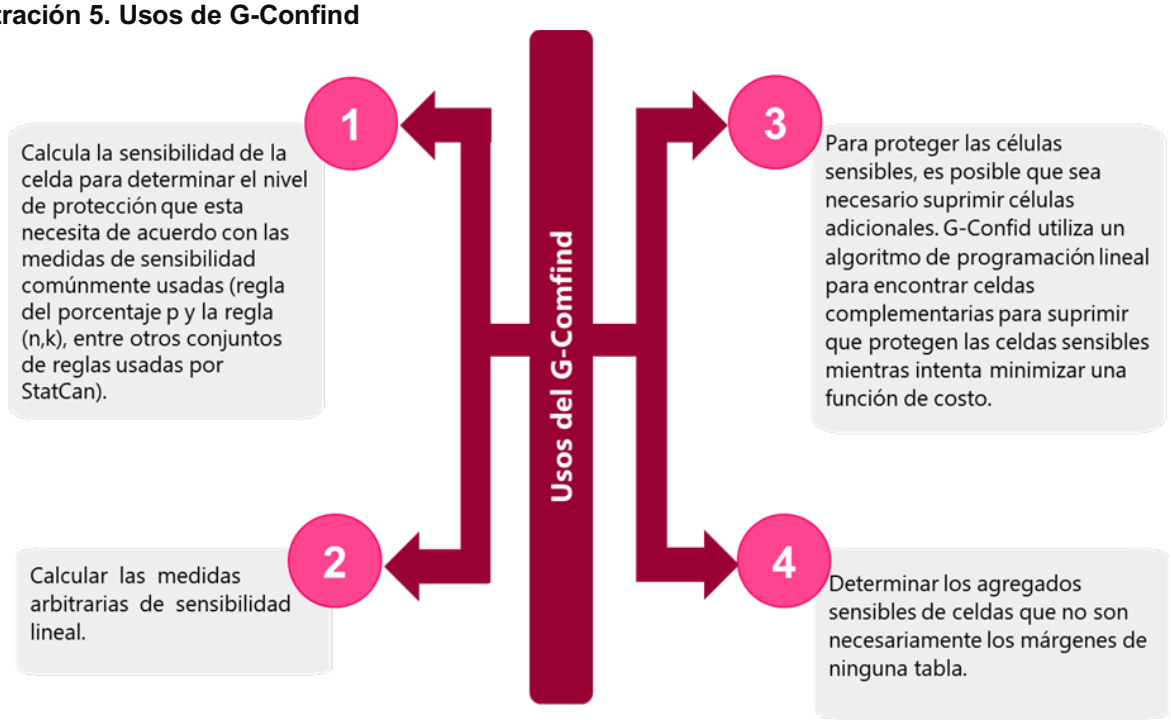
El trabajo presentado en este artículo, ha mejorado significativamente la capacidad de ARX para manejar datos de alta dimensión, tanto en la GUI (interfaz gráfica de usuario) como en la API (interfaz de programación de aplicaciones). Todas las funciones descritas en este artículo están

disponibles como software de código abierto y se incluyen en la última versión del software. En conclusión y con las adiciones realizadas por los investigadores, se ha mejorado significativamente la capacidad de ARX para manejar datos de alta dimensión en términos de rendimiento de procesamiento y facilidad de uso y, por lo tanto, se puede facilitar aún más el intercambio de datos.

1.3.6 Canadá

Statistics Canada desarrolló el software de control de divulgación automatizado G-Confind con el fin de proporcionar el nivel adecuado de protección para celdas confidenciales y minimizar la perdida de información[17]. G-Confind es un sistema generalizado que evita la divulgación de información confidencial en datos tabulares por medio del método de supresión de las celdas, en el cual se identifican y suprimen las celdas sensibles y complementarias a proteger, además, puede usarse para auditar patrones de supresión de celdas, encontrar agregados confidenciales y datos tabulares redondos, la **Ilustración 5** muestra los diferentes usos que se le pueden dar a G-Confind[18].

Ilustración 5. Usos de G-Confind



Fuente DANE a partir de StatCan 2022

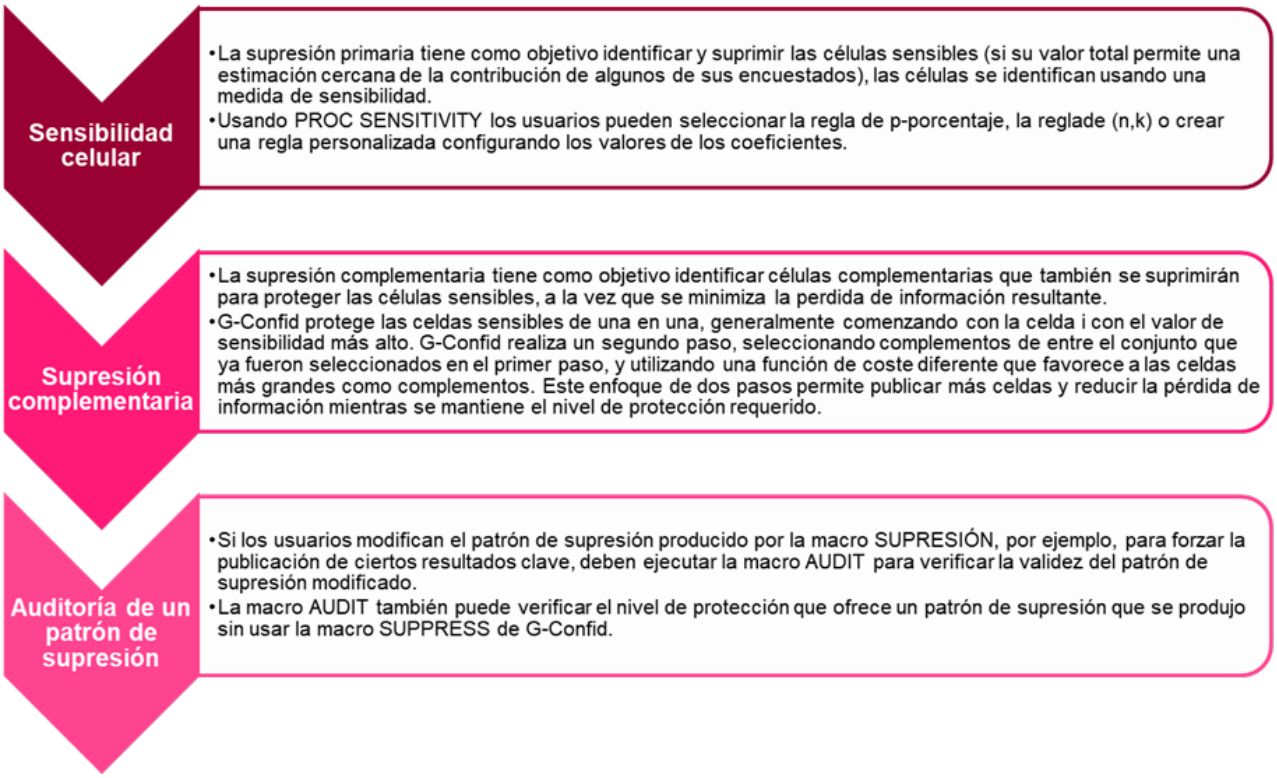
Este software está compuesto por un procesamiento y cinco macros SAS, y está respaldado por un equipo de programadores y metodólogos que ofrecen soporte técnico y metodológico, los usuarios de G-Confind pueden especificar los niveles de agregación en los que crea o valida, patrones de supresión de celdas, este software identifica las celdas de las tablas que requieren supresión primaria y luego optimiza la selección de celdas para la supresión de celdas complementarias, además, puede procesar grandes tablas multidimensionales, así como jerarquías que involucran

17 Disponible en https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/7_gconfid.pdf

18 Disponible en https://www150.statcan.gc.ca/n1/en/catalogue/10H0109

múltiples descomposiciones. G-Confind establece un conjunto de tres componentes SAS para usar con datos económicos tabulares en diferentes niveles de agregación, la Figura G los desarrolla:

Ilustración 6. Componentes SAS del G-Confind



Fuente DANE a partir de StatCan 2013

PROC SENSITIVITY: identificar celdas sensibles, seleccionar un conjunto óptimo de células para la supresión complementaria, este componente procesa los microdatos garantizando que se detecten y evalúen los complementos falsos (celda de complemento potencial que parece ofrecer más protección a una celda sensible de lo que realmente ofrece) a lo largo de líneas unidimensionales (filas, columnas, etc.), una vez se identifica la unión de celdas sensibles y no sensibles que comprenden agregados sensibles, estos se pueden proteger mediante la macro SUPPRESS. El usuario generalmente debe proporcionar cuatro entradas en este componente:

- Un archivo de microdatos.
- Una definición de la(s) jerarquía(s) para cada dimensión de la tabla.
- Los rangos de códigos asociados con el nivel más bajo de cada jerarquía (opcional).
- Las reglas utilizadas para identificar celdas sensibles.

La macro SUPPRESS: Esta macro lleva a cabo la supresión de celdas complementarias, protegiendo las celdas identificadas por PROC SENSITIVITY, y requiere dos archivos de sus salidas:

- El archivo de datos a nivel de celda que contiene la sensibilidad de las celdas (e incluye los agregados sensibles, si los hay).
- El archivo que describe las restricciones lineales que relacionan dos o más celdas a un subtotal o agregado.
- Los usuarios también pueden configurar los valores de costo a cada celda (CVar1 y CVar2), o aceptar los valores predeterminados que son los totales de celda. La asignación de valores de costo personalizados permite al usuario de G-Confid influir en el patrón de supresión.

La macro AUDIT: Auditar los patrones de supresión para verificar la divulgación exacta o parcial, mediante el cálculo de valores mínimos y máximos para cada celda suprimida o agregado sensible utilizando el solucionador LP, garantizando que cada celda sensible está protegida, de lo contrario esta macro indica si alguien que usa la tabla resultante puede descifrar el valor exacto de la celda sensible o un intervalo de los valores con gran precisión[19]:

- Además, existen dos macros auxiliares que brindan mucha información al usuario, aunque no son necesarias en el proceso de supresión de las celdas:
- La macro auxiliar AGGREGATE provee más información sobre uniones sensibles de células.
- La macro auxiliar REPORTCELLS proporciona una instantánea visual del patrón de supresión para facilitar la creación de tablas de salida de los datos económicos[20].

1.3.7 Finlandia

El marco legal para el proceso de anonimización que emplea la Oficina Nacional de Estadística de Finlandia se basa en la Ley de Estadísticas (280/2004) la cual fue sancionada por el Parlamento del país nórdico[21]. Esta ley establece disposiciones sobre los procedimientos y principios relacionados con la recopilación de datos, diseño y producción de estadísticas que deberán aplicar las autoridades estatales en su compilación de estadísticas. Asimismo, la Ley de Estadística de Finlandia está fundamentada en un conjunto de reglamentos, programas y normativas establecidos por los diferentes organismos de la Unión Europea, como se muestra en la **Ilustración 7**[22].

19 Disponible en https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/12_Canada.pdf

20 Disponible en https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_7_Wright.pdf

21 Disponible en https://www.stat.fi/tup/mikroaineistot/ohjeita_tutkijalle_en.html

22 Disponible en https://www.tilastokeskus.fi/meta/lait/2013_tilastolaki_en.pdf

Ilustración 7. Reglamentación que constituyen la Ley 280/2004

Fuente: DANE con base en Statistics Finland[23].

Statistics Finland ha implementado diferentes aplicativos, softwares y sistemas con la finalidad de garantizar el tratamiento y protección de los datos recolectados, procesados y producidos. Uno de los sistemas es FIONA, un software cerrado que ofrece un entorno seguro de datos para acceder a datos de investigación a nivel de unidad estadística. En caso de que una organización, entidad o centro educativo desee adquirir una licencia, deberá someterse a un proceso de selección y celebrar un acuerdo de confidencialidad, protección de datos e implementación de un marco de prácticas de seguridad de datos.

Los datos que se encuentran en FIONA pueden transferirse fuera del sistema, siempre y cuando cumplan con un proceso de selección, en caso de ser aprobado, el sistema garantiza la anonimización de la información, por lo que no se puede identificar ninguna persona o empresa a partir de los datos publicados. Asimismo, es importante declarar que el software contiene datos de otras autoridades gubernamentales de Finlandia y dispone de herramientas de análisis como Stata, R, Python, SAS y SPSS[24].

Teniendo en cuenta el caso de Finlandia, en Colombia se podría adaptar el aplicativo, FIONA, de manera tal que aquellos centros educativos, think tanks, corporaciones o compañías que quieran acceder a datos y microdatos recolectados, procesados por el DANE y demás entidades que componen, lo hagan por medio de un aplicativo que ofrece garantías de protección y anonimización, a la vez que dispone de herramientas de tratamiento y análisis de información como Stata, R o Python.

23 Disponible en: https://www.tilastokeskus.fi/meta/lait/2013_tilastolaki_en.pdf

24 Disponible en https://www.stat.fi/tup/mikroaineistot/etakaytto_en.html

1.3.8 Países Bajos

La Red de Excelencia en el Sistema Estadístico Europeo en el Campo del Control de la Divulgación Estadística (ESSNet SDC), configurada por Eurostat, publicó el Manual sobre el Control de la Divulgación Estadística[25], en este documento provee una guía técnica para balancear la necesidad de los INEs de producir estadísticas y la necesidad de proteger la confidencialidad de los encuestados. No obstante, el control de la divulgación estadística debería ser combinada con otras herramientas administrativas, legales y tecnológicas para definir una estrategia de difusión de datos apropiada basada en el enfoque de gestión de riesgos. Dado que la estrategia de difusión ofrece múltiples productos estadísticos, como por ejemplo datos tabulares, conjuntos de datos dinámicos, microdatos y productos de análisis estadísticos; y cubre un rango amplio de temas para diferentes tipos de usuarios, se requieren diferentes enfoques de SDC y la combinación de diferentes tipos de herramientas.

Los métodos SDC minimizan el riesgo de revelación a un nivel considerado aceptable mientras se libera la mayor cantidad de información como sea posible. Hay dos tipos de métodos, perturbativos y no perturbativos. Los métodos perturbativos falsifican los datos antes de la publicación introduciendo un elemento de error a propósito por razones confidenciales. Los métodos no perturbativos reducen la cantidad de información publicada a través de la supresión o agregación de datos. El handbook describe métodos para el SDC relacionados con dos tipos de productos tabulares, las tablas de magnitud y las tablas de frecuencia, así como para los microdatos.

En lo que concierne a la implementación de los métodos descritos, el handbook se centra en el software μ-Argus, el cual fue desarrollado inicialmente por Estadísticas de Países Bajos, con el objetivo de crear una herramienta para aplicar la metodología de SDC que pusiese ser usada por los INEs, y no ser una caja negra que genere un archivo seguro sin saber el trasfondo de la metodología SDC. Este software se ha convertido en un punto de partida, sobre el cual se han adherido otros métodos adicionales. A continuación, se puede observar un esquema general de las funcionalidades de este software.

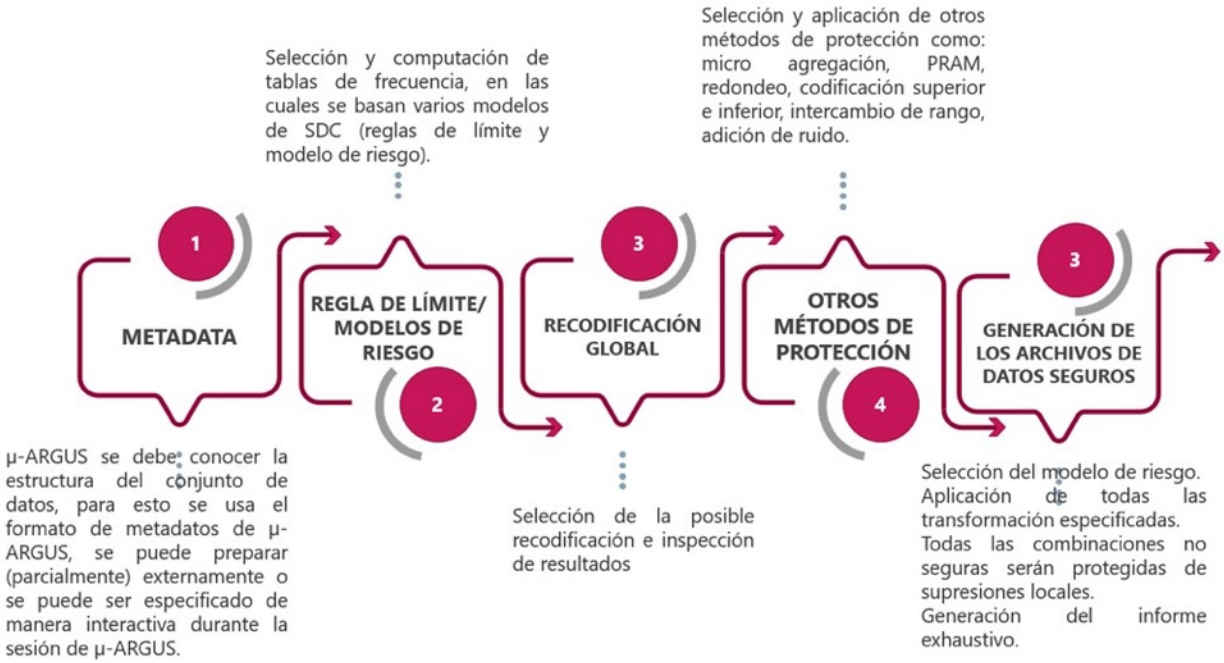
25 Disponible en https://research.cbs.nl/casc/SDC_Handbook.pdf

Ilustración 8. Visión general de µ-ARGUS

Fuente: DANE basado en el Manual sobre control de divulgación estadística del Sistema Estadístico Europeo.

µ-ARGUS es un programa interactivo flexible que guía a los usuarios en el proceso de protección de los datos. En una aplicación típica de µ-ARGUS, si el conjunto de microdatos está listo, se deberían seguir los siguientes pasos:

Ilustración 9. Pasos para el uso µ-ARGUS



Fuente: DANE basado en el Manual sobre control de divulgación estadística del Sistema Estadístico Europeo.

Cuando se sigue todo este esquema, se genera un archivo de microdatos seguro, µ-ARGUS es capaz de manejar conjuntos de datos muy grandes. Solo durante la primera fase, cuando el archivo de datos es explorado y las tablas de frecuencia son calculadas, se realizan algunos cálculos pesados. Esto puede tomar algún tiempo, dependiendo del tamaño del archivo de datos. Sin embargo, el trabajo real de SDC (es decir, aquellos que se nombran en la parte 4 del cuadro anterior), son ejecutados al nivel de la información preparada durante esta primera fase. En la fase final, cuando se generan los archivos de datos, se puede tomar un tiempo considerable. Esta arquitectura de µ-ARGUS tiene la ventaja que todo el trabajo real de SDC se hace de manera interactiva, por lo cual se cuenta con una respuesta de tiempo muy rápida. Así, inspeccionar los resultados de varias recodificaciones es fácil y simple.

Adicionalmente, Estadísticas de Países Bajos aplica un conjunto de métodos y reglas para lograr que los archivos de microdatos sean usados tanto por los investigadores como por otros usuarios públicos. Este enfoque está disponible en µ-ARGUS, y está basado en el enfoque de reglas de límite de ARGUS, en combinación con la recodificación global y la supresión local. Esta regla se concentra en la identificación de variables clave, dado que son el punto de partida de cualquier intrusión. En la Tabla 5 se listan las reglas para el SDC tanto de los microdatos usados tanto por los investigadores como por el público en general.

Tabla 5. Métodos y reglas usados por investigadores y público en general

<table>
<tr><th>Reglas para la anonimización de microdatos usados por investigadores</th><th>Reglas para la anonimización de microdatos usados por el público en general</th></tr>
<tr><td rowspan="2">1. Los identificadores directos no deberían ser liberados y, por lo tanto, deberían ser removidos del conjunto de datos.</td><td>1. Los microdatos tienen que ser al menos un año previo para ser publicados.</td></tr>
<tr><td>2. Los identificadores directos no deberían ser publicados. Tampoco deberían ser publicadas las variables regionales, nacionalidad, país de nacimiento, y origen étnico.</td></tr>
<tr><td rowspan="2">2. Los identificadores indirectos están subdivididos en variables extremadamente identificables, como por ejemplo género, etnia, etc. variables muy identificables, por ejemplo, y variables identificables. Cada combinación de valores de estos tres tipos de variables debería ocurrir al menos 100 veces en la población.</td><td>3. Solo una clase de variables regionales directas (p. ej. la clase del tamaño del lugar de residencia) pueden ser publicadas. Las combinaciones de valores de las variables regionales indirectas deberían ser suficientemente dispersas, p. ej. cada área que pueda ser distinguida debería contener, al menos, 200000 personas en la población objetivo y además debería consistir en municipalidades de al menos seis de las doce provincias en Los Países Bajos. El número de habitantes de una municipalidad en un área que pueda ser distinguida debería ser menor al 50 % de habitantes de esa área.</td></tr>
<tr><td>4. El número de variables de identificación en los microdatos es como máximo 15.</td></tr>
<tr><td rowspan="2">3. El nivel máximo de detalle por ocupación, empresa y nivel de educación está determinado por la variable regional directa más detallada, esta regla no reemplaza la regla 2, pero es una extensión práctica de la regla.</td><td>5. Variables sensibles no deberían ser publicadas.</td></tr>
<tr><td>6. Debería ser imposible derivar información de identificación adicional de los pesos de muestreo.</td></tr>
</table>

	7. Al menos, 200000 personas en la población deberían puntuar en cada valor de una variable de identificación.
4. Una región que pueda ser distinguida en los microdatos debería contener al menos 10.000 habitantes.	8. Al menos 1.000 personas de la población deberían puntuar en cada valor del cruce de dos variables identificadoras.
5. Si los microdatos se refieren a datos de panel, los datos regionales no deberían publicarse, esta regla impide la divulgación de información individual utilizando la característica panel de los microdatos.	9. Por cada hogar del que participe más de una persona en la encuesta, se debe cumplir que el número total de hogares que corresponde a cualquier combinación particular de valores de variables de hogar es al menos de 5 en los microdatos.
	10. Los registros de los microdatos deberían publicarse en orden aleatorio.

Fuente: DANE basado en el Manual sobre control de divulgación estadística del Sistema Estadístico Europeo.

Para el caso en que las reglas para el SDC de microdatos usados por los investigadores, sean violadas, Estadísticas de Países Bajos recomienda aplicar el método de recodificación global o el de supresión local para lograr un archivo seguro. Se debe tener en cuenta que estos dos métodos conllevan pérdida de información debido a que se provee información menos detallada o alguna información no se entrega en absoluto. Es importante encontrar un balance entre estos dos métodos con el fin de hacer que la pérdida de información, debida al SDC, sea lo más pequeña posible. Se recomienda empezar recodificando algunas variables globalmente hasta que el número de combinaciones inseguras que deban ser protegidas sea lo suficientemente bajas. Entonces las combinaciones que permanecen inseguras tienen que ser protegidas por supresiones locales.

1.3.9 Bélgica

En el Instituto Nacional de Bélgica no se evidenció algún software que se esté utilizando para la anonimización de datos. Sin embargo, realizando la búsqueda por país, se encontró una publicación para Bélgica de la empresa IBM que cuenta con las herramientas que facilita a las entidades anonimizar los datos; en publicación se ofrece **"la forma más fácil de anonimizar los datos del sector público antes de su publicación"**[26].

Actualmente, la tarea de anonimizar los datos es extremadamente laboriosa, con grandes secciones de texto bloqueadas, lo que dificulta que cualquier persona lea y comprenda el contenido, lo que a menudo crea sentimientos de desconfianza y frustración entre los ciudadanos que con frecuencia han esperado durante largos períodos de tiempo para recibir la información solicitada.

Una alternativa es usar una nueva solución de You-Get e IBM que identifica cualquier información personal que podría usarse para identificar a una persona y determina el contexto en el que se usa la información personal. De esta manera, el software Automatic Anonymization & Publication Workflow puede sustituir estos datos identificables por etiquetas genéricas, por ejemplo "Testigo 1",

26 Disponible en https://www.ibm.com/blogs/think/be-en/2021/07/01/the-easiest-way-to-anonymize-public-sector-data-before-publication/

"Ciudadano 2", o "Funcionario 3". El resultado es un texto legible, completamente anónimo para cumplir con las últimas normas de privacidad.

Si una autoridad pública prefiere un proceso manual, la solución puede acelerarlo, sugiriendo automáticamente qué información personal debe borrarse, dejando la mayor cantidad posible de información legible y comprensible para el destinatario.

Flujo de trabajo automático de publicación y anonimización
You-Get e IBM ofrecen una solución única para las autoridades públicas que buscan anonimizar documentos de forma segura y escalable. Con el respaldo de IBM Cloud Pak for Automation e IBM Cloud Pak for Data, la solución se integra con la infraestructura y los sistemas de TI existentes para permitir a las autoridades públicas anonimizar sus datos de manera eficiente mientras comparten información con sus ciudadanos de manera transparente, amigable para los ciudadanos y confiable.

1.3.10 Estados Unidos

El Instituto Nacional de Estándares y Tecnología (NIST) cuenta con el programa de ingeniería de privacidad[27], en el que involucran Herramientas de desidentificación. Estas herramientas son "una técnica o proceso aplicado a un conjunto de datos con el objetivo de prevenir o limitar ciertos tipos de riesgos para la privacidad de los individuos, los grupos protegidos y los establecimientos, al tiempo que se permite la producción de estadísticas agregadas, estas técnicas pueden ser introducción de ruido como la privacidad diferencial, el enmascaramiento de datos y la creación de conjuntos de datos sintéticos que se basan en modelos que preservan la privacidad". A continuación, se describen las herramientas utilizadas en la Tabla 6:

Tabla 6. Herramientas utilizadas en NIST

Herramienta	Palabras clave de desidentificación	Descripción	Organización que contribuye	Información adicional
Approximate Minima Perturbation (AMP)	privacidad diferencial, aprendizaje automático	Sirve para la optimización convexa diferencialmente privada, y una extensa evaluación empírica en conjuntos de datos reales de AMP y una serie de enfoques previos para resolver este problema.	Carnegie Mellon University; Universidad de Boston; Universidad de California, Berkeley; Universidad de California, Santa Cruz; Universidad de Pekín.	El repositorio de GitHub contiene implementaciones Python de AMP, descenso de gradiente estocástico ruidoso, Frank-Wolfe ruidoso, perturbación objetiva y dos variantes de perturbación de salida, así como una serie de puntos de referencia para generar resultados experimentales.

27 Disponible en https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id/tools

				https://github.com/sunblaze-ucb/dpml-benchmark Descripción: https://www.uvm.edu/~jnear/papers/TPDPCO.pdf
ARX Data Anonymization Tool	Privacidad diferencial, K-Anonimato, Anonimización, Aprendizaje automático	ARX es un completo software de código abierto para anonimizar datos personales confidenciales. Admite una amplia variedad de (1) modelos de riesgo y privacidad, (2) métodos para transformar datos y (3) métodos para analizar la utilidad de los datos de salida.	TUM - Universidad Técnica de Múnich	https://arx.deidentifier.org/
Chorus	privacidad diferencial	es una herramienta para responder consultas SQL con privacidad diferencial. Chorus funciona con una base de datos SQL estándar y escala a grandes conjuntos de datos descargando el trabajo pesado de responder consultas a la base de datos.	Universidad de Vermont, Universidad de California Berkeley	Chorus utiliza una combinación de reescritura de consultas y posprocesamiento, para implementar mecanismos de privacidad diferencial https://github.com/uvm-plaid/chorus Descripción: https://ieeexplore.ieee.org/document/9230409
Differential Privacy Synthetic Data Challenge Algorithms	privacidad diferencial, generación de datos sintéticos	Realizaron un Desafío de datos sintéticos de privacidad diferencial “Match #3”[28], con el fin de desarrollar algoritmos de código abierto como parte de un esfuerzo para promover la privacidad diferencial.		Desafiaron a los participantes a crear nuevos métodos o mejorar los existentes sobre desidentificación de datos, preservando la utilidad del conjunto de datos para el análisis.
DP_WGAN-UCLANESL		El equipo UCLANESL obtuvo el quinto lugar del desafío. La solución consiste en entrenar una red antagónica generativa wasserstein (w-GAN) que se entrena en el conjunto de datos privado real. El entrenamiento		GitHub https://github.com/nesl/nist_differential_privacy_synthetic_data_challenge https://github.com/usnistgov/PrivacyEngCollabSpace/tree/mast

28 Disponible en https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic

	diferencialmente privado se aplica desinfectando (recorte de normas y agregando ruido gaussiano) los gradientes del discriminador. Una vez que se entrena el modelo, se puede utilizar para generar un conjunto de datos sintéticos alimentando ruido aleatorio en el generador.		er/tools/de-identification/Differential-Privacy-Synthetic-Data-Challenge-Algorithms/DP_WGAN-UCLANESL
DPFieldGroups	Obtuvo el cuarto lugar del desafío, su solución consista en agrupar campos que se han encontrado que están altamente correlacionados. Para cada uno de estos grupos, se crea un histograma con el fin de contar el número de ocurrencias de cada posible combinación de valores de todos los campos del grupo. Para la privatización, el ruido laplaciano se agrega a cada contenedor con una escala proporcional al número de grupos/épsilon total. Los datos sintéticos se generan seleccionando un contenedor aleatorio para cada grupo con probabilidad ponderada por estos conteos de contenedores ruidosos. Los valores de campo correspondientes al bin seleccionado de cada grupo se escriben como una sola fila de datos sintéticos.		GitHub https://github.com/usnistgov/PrivacyEngCollabSpace/tree/master/tools/de-identification/Differential-Privacy-Synthetic-Data-Challenge-Algorithms/DPFieldGroups
DPSyn	Presentaron un algoritmo para sintetizar microdatos mientras se satisface la privacidad diferencial, y su creación de instancias en el conjunto de datos utilizado en la competencia, a saber, Muestra de microdatos de uso público (PUMS) de los datos del censo de EE. UU. de 1940.		GitHub https://github.com/usnistgov/PrivacyEngCollabSpace/tree/master/tools/de-identification/Differential-Privacy-Synthetic-Data-Challenge-Algorithms/DPSyn
rmckenna	Obtuvo el primer puesto del desafío y presentaron la idea de:		GitHub https://github.com/usnistgov/PrivacyEn

		(1) utilizar el mecanismo gaussiano para obtener respuestas ruidosas a un conjunto cuidadosamente seleccionado de consultas de recuento (marginales de 1, 2 y 3 vías) y (2) encontrar un conjunto de datos sintéticos que se aproxime a los datos verdaderos con respecto a esas consultas. El último paso se realiza con 3 vías, y el anterior utiliza ideas inspiradas en 1 y 2 vías. Más concretamente, esto se hace calculando la información mutua (en el conjunto de datos públicos) para cada par de atributos y seleccionando las consultas marginales que tienen una alta información mutua.		gCollabSpace/tree/master/tools/de-identification/Differential-Privacy-Synthetic-Data-Challenge-Algorithms/rmckenna
Differentially Private Stochastic Gradient Descent (DP-SGD)	privacidad diferencial, aprendizaje automático	Entrene modelos de aprendizaje automático con privacidad diferencial mediante el recorte y el ruido de gradientes durante el descenso de gradiente estocástico.	Google	Descripción: https://arxiv.org/abs/1607.00133 GitHub https://github.com/tensorflow/privacy
Diffprivlib	privacidad diferencial, aprendizaje automático, análisis de datos	s una biblioteca de Python de uso general para experimentar y crear herramientas para la privacidad diferencial. Diffprivlib incluye una serie de algoritmos para el aprendizaje automático y el análisis de datos con privacidad diferencial listos para usar en la conocida sintaxis de Scikit-learn y Numpy.	IBM Research	Descripción: https://arxiv.org/abs/1907.02444 GitHub https://github.com/IBM/differential-privacy-library
Duet	privacidad diferencial, verificación de algoritmos, aprendizaje automático	Duet es un lenguaje de programación que deriva (y verifica) automáticamente los límites de privacidad diferencial para los programas escritos en el lenguaje. Duet está diseñado para admitir algoritmos de aprendizaje automático modernos y variantes	Universidad de Vermont, Universidad de California en Berkeley, Universidad de Utah	GitHub : https://dl.acm.org/doi/10.1145/3360598 https://github.com/uvm-plaid/duet

		avanzadas de privacidad diferencial para agregar un ruido mínimo a los resultados del algoritmo para garantizar la privacidad.		
Ektelo	privacidad diferencial	Ektelo es un marco de programación y un sistema que ayuda a los programadores a desarrollar programas diferencialmente privados con gran utilidad. Ektelo se puede usar para crear programas para una variedad de tareas estadísticas que implican responder consultas de conteo sobre una tabla de dimensión arbitraria.	UMass Amherst, Duke University, Colgate University	Ektelo se describe en detalle en un documento SIGMOD 2018, titulado "EKTELO: un marco para definir computaciones diferencialmente privadas". https://dl.acm.org/doi/10.1145/3183713.3196921 GITHUB https://ektelo.github.io/
Google Differential Privacy Library	privacidad diferencial	proporciona un conjunto de componentes básicos que permiten a los desarrolladores crear aplicaciones privadas de forma diferencial en C++, Java y Go. Además, la biblioteca de DP de Google ofrece 'Privacy on Beam', una implementación integral de privacidad diferencial que ayuda a los desarrolladores a realizar operaciones de manera diferentemente privada.	Google	GitHub https://github.com/google/differential-privacy
GUPT: Privacy preserving data analysis made easy	privacidad diferencial, aprendizaje automático, consultas de base de datos	proporciona garantías de privacidad diferenciales a los algoritmos estadísticos/de aprendizaje automático al tratar el algoritmo subyacente como una caja negra y solo depende de las firmas de entrada/salida.	Universidad de California, Berkeley; Universidad de California, Santa Cruz; Universidad de Cornell	Implementa una variante del célebre marco de muestras y agregados de Nissim, Rashkhodnikova y Smith, 2007. GUPT se describe en detalle en un documento SIGMOD 2012, titulado "GUPT: análisis de datos de preservación de la privacidad simplificado". https://www.cs.umd.e

				du/~elaine/docs/gupt.pdf GitHub https://github.com/prashmohan/GUPT
PixelDP	privacidad diferencial, verificación de algoritmos, aprendizaje automático, ejemplos adversarios	Los ejemplos adversarios que engañan a los modelos de predicción son una nueva clase de ataques introducidos por las implementaciones de aprendizaje automático. PixelDP es la primera defensa certificada que ofrece garantías comprobables de robustez contra estos ataques y se adapta a grandes modelos y conjuntos de datos, como el conjunto de datos Inception on the ImageNet de Google. El diseño de PixelDP se basa en un uso novedoso de la privacidad diferencial en el momento de la predicción.	Universidad de Columbia	Descripción https://arxiv.org/pdf/1802.03471.pdf GitHub https://github.com/columbia/pixeldp
Privacy Protection Application (PPA)	de desidentificación: anonimato K, anonimización, fuga de información, imparcialidad algorítmica, consultas de base de datos, datos de ubicación	desidentifica las bases de datos que contienen datos de geolocalización secuenciales, a veces denominadas bases de datos de objetos en movimiento. Un registro de la ruta de viaje de un vehículo de propiedad personal es un ejemplo, pero la herramienta puede procesar otros tipos de secuencias de geolocalización. La aplicación tiene una interfaz gráfica de usuario y funciona en Linux, OS X y Windows. La supresión de ubicación es la estrategia de desidentificación utilizada, y las decisiones sobre qué ubicaciones suprimir se basan en la teoría de la información. Esta estrategia no modifica la precisión de la información de ubicación retenida. Uno de los objetivos es producir datos útiles para el análisis de la seguridad de los vehículos y el desarrollo de aplicaciones de transporte.		GitHub https://github.com/usdot-its-jpo-data-portal/privacy-protection-application/releases/tag/hmm-mm https://github.com/usdot-jpo-ode/jpo-cvdp https://github.com/usdot-its-jpo-data-portal/privacy-protection-application

Private Aggregation of Teacher Ensembles (PATE)	privacidad diferencial, aprendizaje automático	El marco PATE logra un aprendizaje privado diferencial al coordinar cuidadosamente la actividad de varios modelos ML diferentes.	Google	Descripción https://arxiv.org/abs/1610.05755 GitHub https://github.com/tensorflow/privacy/tree/master/research

Fuente: DANE basado en NIST

De las anteriores herramientas se destaca el ARX Data Anonymization Tool, por ser un software de código abierto para anonimizar datos personales confidenciales. Puede manejar grandes conjuntos de datos en hardware básico y cuenta con una interfaz gráfica de usuario intuitiva multiplataforma, proporciona varias visualizaciones, asistentes y una ayuda sensible al contexto. En la Ilustración 10 se presenta la descripción general de los métodos admitidos por ARX.

Ilustración 10. Métodos de anonimización admitidos

Fuente: DANE basado en ARX

1.3.11 Francia

El Instituto Nacional de Estadística y Estudios Económicos de Francia - INSEE - publicó el documento “Gestión de la privacidad de los datos individuales”[29] el cual se centra en las técnicas

29 Disponible en https://www.insee.fr/fr/statistiques/fichier/2535625/M201607.pdf

implementadas en µ-Argus[30] el software de protección de microdatos de las estadísticas oficiales (más usado en Europa) y el paquete sdcMicro R[31], la aplicación de los métodos sigue tres pasos:

1. **Medición del riesgo de reidentificación.** La estimación de los riesgos de reidentificación permiten definir umbrales máximos de riesgo a nivel de hogar y/o individual, donde los parámetros de reducción de riesgo dependen en gran medida de los usuarios potenciales, restricciones legislativas y grado de sensibilidad de los datos, el uso de métodos de protección de datos ayuda a reducir el riesgo al umbral deseado.
2. **Técnicas de protección para reducir el riesgo de reidentificación.** El documento presenta diferentes métodos de reducción de reidentificación que pueden ser usados para variables cuasi_identificadoras (permiten limitar la revelación de identidad, mientras que se busca proteger contra la revelación de atributos cuando los mecanismos de protección consideran variables no identificativas), y/o variables sensibles no identificadoras, la Tabla X ilustra los métodos disruptivos y no disruptivos implementados en el software µ-Argus.

Tabla 7. Métodos disruptivos y no disruptivos implementados en el software µ-Argus

Métodos no disruptivos: limitar las posibilidades de reidentificación actuando sobre las variables cuasi-identificadoras para que no haya más claves de identificación en riesgo.	
Método	Descripción
Grabaciones de variables	La recodificación de variables se realiza a nivel global, es decir, se recodifica la variable para todos los individuos del fichero. Se ofrece un algoritmo de grabación local en el paquete sdcMicro R. Para una variable categórica Vi, la recodificación global consiste en combinar categorías para formar otras nuevas. Obtenemos así una nueva variable Vi'. Si Vi es una variable continua, la recodificación consiste en una discretización de la variable en una variable categórica. Un caso especial de recodificación es la recodificación superior o inferior para variables ordenables (variables categóricas ordinales o variables continuas).
Eliminaciones locales – Minimización de eliminaciones	La realización de borrados locales consiste en borrar, para determinados individuos que poseen una clave de identificación de riesgo, una o más de las variables cuasi-identificadoras reemplazándolas por un valor faltante. La minimización de borrados locales se realiza con la restricción de un objetivo de reducción del riesgo, que podría ser por ejemplo el k-anonimato o la obtención de un umbral máximo de riesgo de reidentificación por clave de identificación por debajo de un determinado umbral. Esta técnica se puede utilizar para variables categóricas cuasi-identificadoras.
Métodos disruptivos: métodos de perturbación de datos, por lo que el archivo de datos de transmisión contiene variables modificadas en lugar de variables exactas, logrando un compromiso entre la protección del archivo y la pérdida de información.	

30 Disponible en https://research.cbs.nl/casc/mu.htm

31 Disponible en https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf

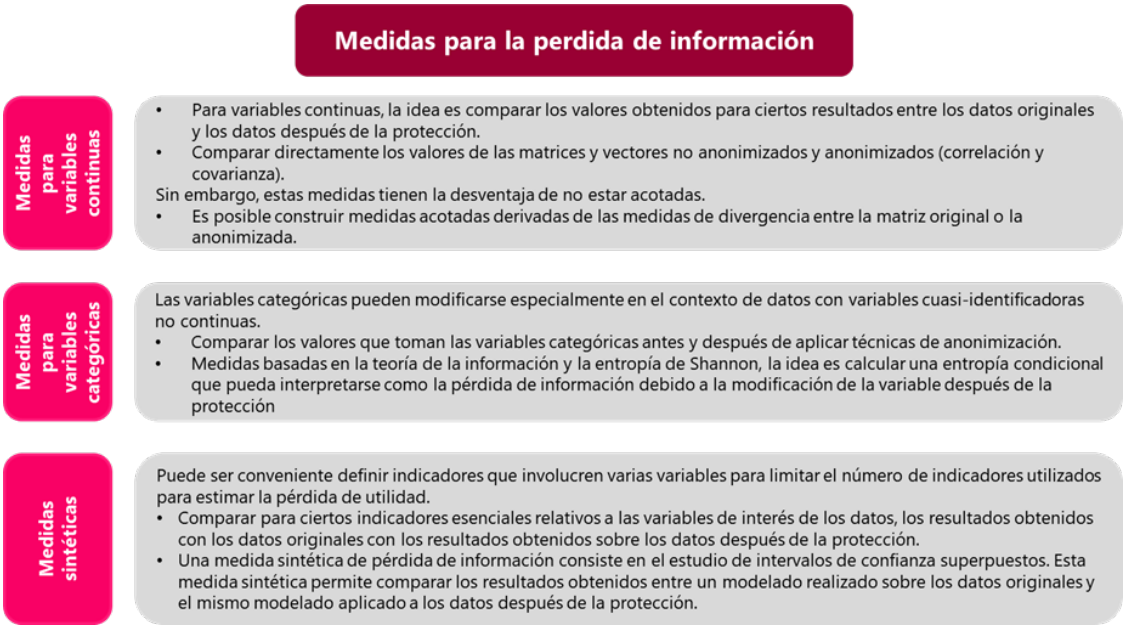
Perturbación por ruido aditivo	La perturbación por ruido aditivo es adecuada para variables continuas, pues: i) no se hacen suposiciones sobre los valores que posiblemente tomen las variables a perturbar, ii) el ruido añadido es, en general, continuo y se supone que tiene una expectativa cero y iii) La adición de ruido hace imposible la coincidencia exacta con archivos externos. • Ruidos independientes: con este método se conservan las medias y las covarianzas, pero no se conservan las varianzas y, en consecuencia, los coeficientes de correlación. • Ruidos correlacionados: con este método se conservan los coeficientes de correlación y de la expectativa, agregar ruidos correlacionados es preferible a agregar ruidos independientes porque se pueden obtener estimaciones no sesgadas para varias estadísticas importantes. • Adición de ruido y transformación lineal, este método asegura que la matriz de varianza-covarianza de las variables perturbadas es un estimador insesgado de la matriz de varianza-covarianza de las variables originales.
El método PRAM (Post-Randomization Method)	Esta técnica de perturbación adaptada a datos categóricos, es una perturbación aleatoria de datos, donde el usuario define completamente el mecanismo de la perturbación. Si el mecanismo de perturbación se difunde con el archivo perturbado, un usuario puede estimar sin sesgo las características del archivo de datos inicial utilizando los datos perturbados y corrigiendo la perturbación.
Técnicas de microagregación: la idea de la microagregación es formar en el archivo de datos inicial g grupos de tamaño al menos k, siendo la idea subyacente obtener un archivo k-anónimo. Dentro de un grupo, para cada variable, el valor inicial se reemplaza por el valor ”promedio” o “mediana” de la variable dentro del grupo. Los grupos g se forman de tal forma que los individuos que los constituyen tienen características similares.	Microagregación univariada: las primeras técnicas de microagregación desarrolladas para anonimizar archivos de datos consideran una sola dimensión. Se puede obtener una solución óptima al problema de minimización presentado anteriormente en un tiempo razonable. Se proponen dos técnicas principales para gestionar la perturbación de varias variables simultáneamente con algoritmos de microagregación univariante: • Ranking individual Se realiza una microagregación independiente para cada variable, es decir, la formación de los g grupos depende de la variable considerada, esta técnica proporciona poca protección, siendo la reidentificación relativamente sencilla cuando las microagregaciones se realizan de forma independiente. • Proyecciones sobre un eje: también puede proyectar las variables del archivo sobre un eje antes de lanzar la microagregación sobre esta variable que se supone sintetizará el archivo de datos. Esta técnica proporciona una protección significativa, pero crea una gran pérdida de información, en particular sobre los vínculos entre variables.
	Microagregación multivariante: en este caso general, el problema de minimizar la variabilidad dentro de los grupos de la partición es NP-difícil, no puede resolverse en un tiempo limitado. Por lo tanto, se han desarrollado varios algoritmos para aproximar la solución óptima. • Algoritmo MDAV (Microagregación Multivariante basada en Distancia Máxima al Vector Promedio) Este algoritmo busca una partición donde los grupos son de tamaño fijo k e idénticos para cada uno de los g grupos de la partición.

Métodos de redondeo	Los procedimientos de redondeo, los valores originales de las variables se reemplazan por valores redondeados. Para una variable Xi, las posibilidades de redondeo se definen en un conjunto de posibles redondeos: se puede, por ejemplo, utilizar una base de redondeo b y los redondeos siempre serán múltiplos de esta base. Normalmente, en un conjunto de datos multivariante, el redondeo se realiza de forma independiente para cada variable.

Fuente DANE a partir de INSEE 2016

3. **Medición de la pérdida de información luego de la anonimización de los datos**. La confidencialidad consiste en equilibrar la protección de datos individuales y la perdida de utilidad que generan los métodos de anonimización, este criterio puede permitir comparar varios métodos entre sí, sin embargo, la elección de las métricas usadas para medir la perdida de utilidad es compleja, pues requiere hacer suposiciones sobre el supuesto uso de los datos protegidos que se desea difundir, la Figura Y describe algunas medidas para variables continuas, categóricas y sintéticas.

Ilustración 11. Medidas para variables continuas, categóricas y sintéticas



Fuente DANE a partir de INSEE 2016

Por último, el INSEE publicó en 2018 el "Manual de análisis espacial"[32], el cual proporciona una visión general de los métodos que se pueden usar cuando se conoce la ubicación de las unidades estadísticas estudiadas, el objetivo del manual es dar respuesta a las preguntas: ¿qué hacer con estas nuevas fuentes de datos geolocalizados?, ¿en qué casos se debe tener en cuenta su dimensión espacial?, y ¿cómo aplicar los métodos de la estadística y la econometría espacial?

32 Disponible en https://www.insee.fr/fr/statistiques/fichier/version-html/3635442/imet131.pdf

Además, ilustra temas específicos como muestreo espacial, estimación en áreas pequeñas y confidencialidad de los datos espaciales, en este último tema en específico se aborda el problema de diferenciación geográfica, el cual se presenta cuando un intruso es capaz de combinar datos difundidos en diferentes geografías para reconstruir así estadísticas en un área más grande o inferir la ubicación a la que se refiere una observación. En sistemas de geografías anidadas, las pequeñas áreas que se pueden deducir por sustracción están ligadas a la jerarquía de las diferentes geografías, por lo que una vez identificadas las áreas a proteger por lo general se usa un software de confidencialidad como µ-Argus para gestionar el secreto secundario (para todas las áreas a proteger).

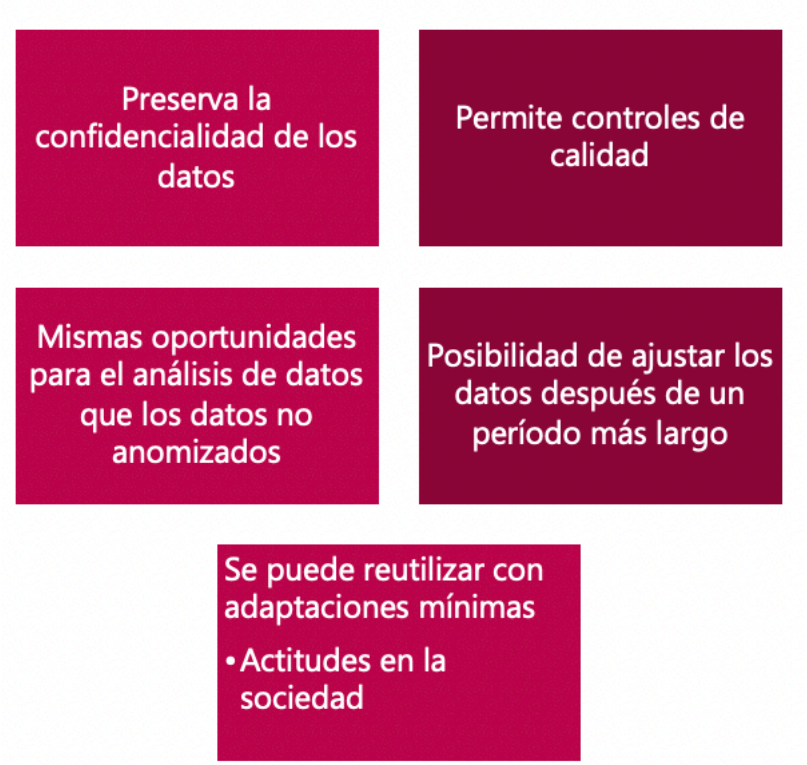
1.3.12 Noruega

Noruega, al igual que la mayoría de los países europeos, dispone de un marco normativo para la información y estadística, en el caso del país nórdico, este se encuentra concebido en la Ley de Estadística. Este precepto dispone de un conjunto de reglas, principios y procedimientos para el asertivo tratamiento, procesamiento, producción y divulgación de información estadística. Dentro de los elementos que esta normativa contiene, se contempla la anonimización de toda información que sea procesada por la ONE de este país, esto con la finalidad de no identificar, ni permitir, el rastreo de las unidades estadísticas[33].

Con el fin de anonimizar la información de las diferentes operaciones estadísticas y registros administrativos, Statistic Norway ha trabajado en diferentes estrategias, una de ellas es la generación de datos sintéticos con fines de anonimización. Un ejemplo implementación de esta metodología fue sustentada en octubre del 2019 por representantes de la ONE de Noruega, en el marco de la sesión de trabajo sobre confidencialidad en datos estadísticos liderado por la Comisión Económica de las Naciones Unidas para Europa (UNECE) y EUROSTAT[34]. Allí se mostraron los resultados y beneficios (Ilustración 12) de la aplicación de esta metodología para anonimizar la Encuesta noruega sobre condiciones de vida del año 2015.

[33] Disponible en: https://lovdata.no/info/information_in_english
[34] Disponible en: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S1.2_Norway_Heldal_Iancu_P.pdf

Ilustración 12. Beneficios de la anonimización por medio matching estadístico de datos sintéticos

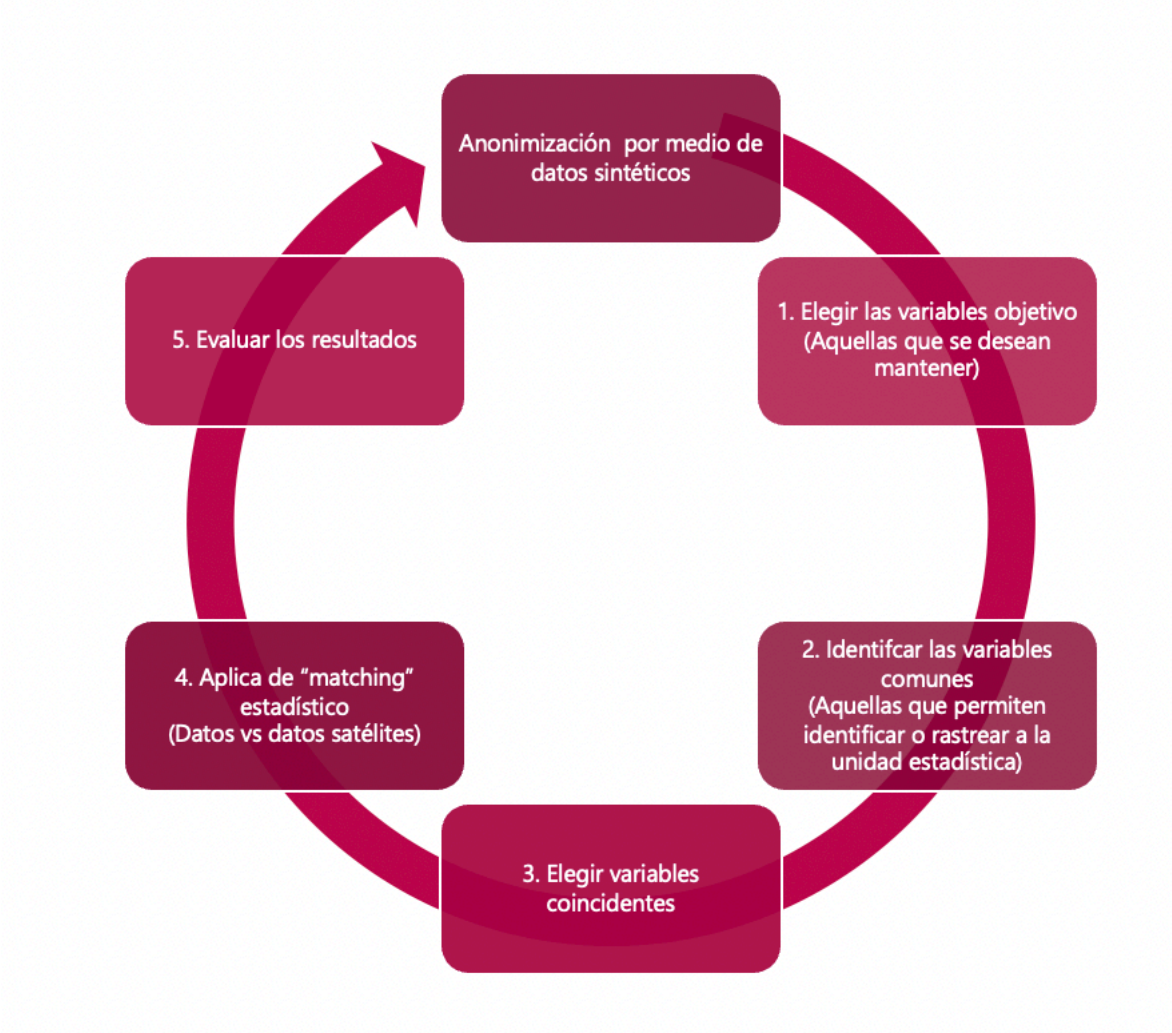


Fuente: DANE con base Statistics Norway

La generación de datos sintéticos con fines de anomización consiste en crear conjuntos de datos "satélite" que se puedan cruzar con los datos recolectados y eliminen aquellas variables que impiden que la información esté anonimizada[35]. En otras palabras, Statistic Norway propone anonimizar, por medio de *matching* estadístico, información que permita identificar o hacer seguimiento a las unidades estadísticas. El procedimiento para anonimizar la información contempla 5 pasos sustentados en la Ilustración 13.

[35] Disponible en: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S1_Norway_Heldal_Iancu_AD.pdf

Ilustración 13. Procedimiento para anonimización por medio de matching estadístico de datos sintéticos.



Fuente: DANE con base en Statistic Norway[36]

1.4 Conclusiones

La revisión de referentes internacionales que hemos realizado para el tema Implementación de instrumentos, software, aplicativos o sistemas que permitan la anonimización de bases de datos, nos permite dar respuesta a las preguntas de investigación planteadas y se concluye lo siguiente:

A. El paquete sdcMicro R a menudo se complementan con otras herramientas (SAS, STATA, R, SPSS o Excel) para lograr una efectiva protección de datos en tablas y archivos de microdatos. sdcMicro optimiza la anomización para grandes conjuntos de datos, el cual

[36] Disponible en: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S1.2_Norway_Heldal_Iancu_P.pdf

puede usarse por medio de R o por una interfaz sdcApp (GUI) para aquellos usuarios que no son expertos en R.

B. El software sdcApp no solamente permite aplicar técnicas de limitación de divulgación, sino que facilita todo el proceso de anonimización, esta interfaz permite subir datos al sistema, modificarlos y crear un objeto que defina el escenario de divulgación, una vez que se ha definido un problema de control de divulgación estadística (SDC), los usuarios pueden aplicar técnicas de anonimización a este objeto y obtener comentarios instantáneos sobre el impacto en el riesgo y la utilidad de los datos después de que se hayan aplicado los métodos SDC.
   Se destaca que sdcApp proporciona la lista más completa de métodos populares, además posee muchas más posibilidades de gestionar problemas complejos referentes a anonimización de lo que terminan ofreciendo otras herramientas, como Arx, PARAT, OpenAnonymizer, SECRETA, Amnesia, Cornell Anonymization Toolkit y TIAMAT.
   En este contexto, el grupo de expertos en anonimización de bases de datos del DANE, puede evaluar por medio de la aplicación de ejercicios que permitan la combinación de los softwares actuales utilizados en la entidad con sdcMicro R y sdcApp, cuál software se adapta mejor a la naturaleza propia de la información y operaciones estadísticas susceptibles al proceso.

C. Se presentan algunas recomendaciones de los referentes que cuentan con mayor experiencia en la aplicación de software de anonimización.
   a. Eurostat apoya la migración de herramientas de control de divulgación estadística hacia soluciones de código abierto, entre ellas están µ-Argus el software de protección de microdatos de las estadísticas oficiales más usado en Europa y el software ARX el cual es usado por Instituto Nacional de Estándares y Tecnología de Estados Unidos, este software permite manejar grandes conjuntos de datos en hardware básico y cuenta con una interfaz gráfica de usuario intuitiva multiplataforma que proporciona varias visualizaciones, asistentes y una ayuda sensible al contexto.
   b. Statistic Norway propone la generación de datos sintéticos con fines de anomización, por medio de matching estadístico, el cual ofrece ventajas como i) permitir controles de calidad, ii) las mismas oportunidades de análisis de datos que los no anonimizados, iii) posibilidad de ajustar los datos después de un periodo mas largo y iv) se puede reutilizar con mínimas adaptaciones.
   c. El grupo de Análisis y Liberación de Resultados de Registros Administrativos del INEGI, utilizó el software REDATAM, versión REDATAMSP+ Process, en el procesamiento de bases de datos para las encuestas de Salud, Cultura, Intentos de Suicidio y Suicidios, posteriormente lo utilizó en el procesamiento de Conteo de Población y Vivienda de 2005, además, plantea que la versión REDATAM7 permite una buena aplicación en encuestas especiales de hogares, procesando los datos mediante el módulo Process y editanto los resultados en hojas de cálculo. Desde el INEGI se recomienda ahondar en las bondades que permite REDATAM en el procesamiento y análisis de información estadística por medio de la generación de cuadros, indicadores, gráficas, mapas y formatos de intercambio como SIDRA, considerando que es un software muy robusto y no tiene costo de uso.

2.

Gobernanza y ética de los datos en torno a la interoperabilidad en el Sistema Estadístico Nacional

2. Recomendaciones que ofrecen los referentes internacionales respecto a la gobernanza y ética de los datos en torno a la interoperabilidad en el Sistema Estadístico Nacional

2.1 Resumen

Actualmente, en Colombia, los datos son sub utilizados, y existe un importante potencial sin explotar en todos los datos disponibles del Sistema Estadístico Nacional. Es posible identificar la producción de información estadística que usan diferentes tipos de datos (censos, encuestas, registros administrativos, observaciones, sensores terrestres, etc.) provenientes de un ecosistema de datos que tiene a su vez diversos niveles de habilidades y capacidades instaladas. En muchos países, esta situación es una falla en la integración de los sistemas estadísticos nacionales y les impide aprovechar las posibles nuevas fuentes de información para ayudar a subsanar las brechas de conocimiento y orientar las decisiones políticas. La capacidad de compartir datos de manera fácil, eficiente y ética es crucial para superar estos desafíos.

"La interoperabilidad es la habilidad de reunir datos de diversas fuentes de manera estandarizada y contextualizada. Sin embargo, se trata de más que solo la forma y estructura de los datos, también se trata de resolver problemas de forma conjunta. […] La interoperabilidad puede ayudar a reducir el tiempo, esfuerzo y gastos que conlleva la recopilación de datos; a eliminar la frustración y los riesgos asociados al manejo de datos incompletos e incoherentes; y a satisfacer la necesidad de contar con datos comparables internacionalmente, sostenibles y desagregados, para garantizar que nadie se quede atrás". (JUDS, 2016)[37].

Para visualizar una mejor integración y uso sistémico de datos para Colombia, se debe buscar tener información fácil de compartir, y es necesario trabajar para que los datos, es decir, los elementos básicos de la información, sean pensados también de una manera más accesible y garantizar su amplia reutilización. Enfrentarse al problema de la interoperabilidad de datos en el Sistema Estadística Nacional, significa explorar cómo se pueden organizar y estructurar mejor los conjuntos de datos, los sistemas de tecnología de la información, el intercambio y la gestión de datos internos, así como la coordinación interinstitucional en torno a las cuestiones relativas a ética de la interoperabilidad de los datos.

En línea con lo anterior y teniendo presente que existe un Marco de Interoperabilidad para Gobierno Digital[38] desarrollado por el Ministerio de Tecnologías de la Información y las Telecomunicaciones, se plantea la necesidad de contar con una revisión internacional de las recomendaciones que ofrecen los Institutos de Estadística, Ministerios y Organizaciones Privadas respecto a la gobernanza y ética de los datos en torno a la interoperabilidad en el Sistema Estadístico Nacional para entender cuáles son las fortalezas y debilidades del modelo propio desarrollado, dentro del contexto y el parámetro internacional.

37 Disponible en: http://devinit.org/wp-content/uploads/2018/02/The-frontiers-of-datainteroperability-for-sustainable-development.pdf

38 Disponible en: http://lenguaje.mintic.gov.co/sites/default/files/archivos/marco_de_interoperabilidad_para_gobierno_digital.pdf

3.1 Síntesis de hallazgos

A continuación se presenta una breve descripción de los principales hallazgos de la revisión de referentes internacionales, la gobernanza y ética de los datos en torno a la interoperabilidad en el Sistema Estadístico Nacional, se consultaron cuatro organismos internacionales, 2 países de América del Norte, tres países de Europa, un país de Suramérica y dos empresas del sector privado.

Tabla 8. Principales hallazgos sobre la Gobernanza y Ética de los Datos en torno a la interoperabilidad en el Sistema Estadístico Nacional

Referente	¿Qué recomendaciones ofrecen los INEs, Ministerios y Organizaciones privadas en términos de gobernanza y ética de los datos en torno a la interoperabilidad en el Sistema de estos en el Estadístico Nacional? ¿Qué ejercicios se vienen realizando?
Organizaciones internacionales	**OECD:** La OECD publicó el documento "Recomendación de la OECD sobre la mejora del acceso e intercambio de datos (EASD)" cuyo objetivo es determinar principios generales y orientación política para que los gobiernos maximicen los beneficios de mejorar los acuerdos de acceso e intercambio de todos los tipos de datos (privados, públicos, abiertos, personales y no personales), a la vez que se protegen los derechos de las personas y organizaciones, en esa misma línea, desarrolló el documento "Mejora del acceso y el intercambio de datos: reconciliación de riesgos y beneficios para la reutilización de datos en las sociedades" el cual propone diferentes enfoques para alentar, facilitar y mejorar el acceso e intercambio de datos, donde cada enfoque propone una estrategia para abordar los principales desafíos que se presentan en este proceso. **CEPAL:** La CEPAL generó un documento guía denominado "Gobernanza digital e interoperabilidad gubernamental", dirigido a actores de todos los sectores (gobierno, sociedad civil, sector académico y sector privado), donde se presentan las recomendaciones, fundamentos y estándares de la gobernanza digital y la interoperabilidad gubernamental, enfocándose en su aplicación práctica, se propone una hoja de ruta para la definición de institucionalidad y las características que abarquen los componentes de gobernanza e interoperabilidad, con énfasis en la identificación y puesta en marcha de iniciativas que surgen de las experiencias adquiridas por medio de la asistencia técnica realizada a Costa Rica. **BID:** El BID ofrece marcos de referencia sobre la gestión ética de los datos y la importancia de la confiabilidad de la información, expone los desafíos en el uso de datos, las buenas prácticas en la gestión ética y establece una propuesta de criterios para una gestión ética de datos en el sector público, basada en cinco etapas de creación de valor público: i) recolección, ii) almacenamiento, iii) análisis, iv) compartición y v) eliminación, además, diseñó una metodología que permite una visión amplia de los pasos a seguir en el proceso de interoperabilidad. **Unión Europea:** Durante el 2020, la Comisión Europea propuso un nuevo esquema normativo para la gobernanza de los datos, cuyo principal objetivo es facilitar el intercambio de datos en toda la Unión Europea y entre sectores, a la vez que se

Referente	¿Qué recomendaciones ofrecen los INEs, Ministerios y Organizaciones privadas en términos de gobernanza y ética de los datos en torno a la interoperabilidad en el Sistema de estos en el Estadístico Nacional? ¿Qué ejercicios se vienen realizando?
	aumenta la confianza en los datos, esperando así generar mayor provecho de los datos e información.
Canadá	Statistics Canada, cuenta con un centro de confianza de estadísticas de Canadá, encargado de encriptar y anonimizamos todos sus datos, comprometidos en proteger la privacidad y salvaguardar la confidencialidad de los datos que se les confían. Para liderar con integridad, Statistics Canada cuenta con una Secretaría de Ética de Datos y con consejos asesores, especialmente el Consejo Asesor de Ética y Modernización de Acceso a Microdatos. Statistics Canada adoptó el Marco de Necesidad y Proporcionalidad. Cada propuesta para un nuevo proyecto o adquisición de datos debe explicar por qué es importante, cuáles son los beneficios para los canadienses, quién necesita la información y abordar consideraciones éticas como privacidad, transparencia y se somete a una revisión ética por parte de la Secretaría de Ética de Datos.
España	En España, una de las entidades representativas en las actividades de interoperabilidad e intercambio de información es la Oficina del Dato, creada en 2020 y adscrita a la Secretaría de Estado de Digitalización e Inteligencia artificial, esta Oficina promueve el fortalecimiento de la infraestructura de datos en España con el fin de apoyar las políticas digitales del Gobierno y los programas para desarrollar la estrategia España Digital 2025. Dentro de sus actividades promueve los principios de los datos tales como: Encontrables, Accesibles, Interoperables y Reutilizables.
Nueva Zelanda	La gobernanza de datos y ética de datos está liderada por el Instituto Nacional de Estadística de Nueva Zelanda (Stats NZ) en cabeza del director del Instituto, quien en colaboración con el director digital del gobierno, el director de seguridad de la información del gobierno y el director de privacidad del gobierno plantean la estrategia y hojas de ruta de los datos. En la estructuración de la estrategia y hoja de ruta participan grupos consultivos en el que se destaca el Grupo Asesor de Ética de Datos que fomenta los principios de la privacidad, la seguridad y la confidencialidad. En el Stats NZ se destaca la estrategia de los procesos de Infraestructura Integrada de datos que integra datos de diferentes organismos tanto gubernamentales como no gubernamentales para ofrecer información a nivel de hogares. Dentro de las experiencias de interoperabilidad en el SEN se distingue el proyecto de interoperabilidad del Ministerio de Salud.
Reino Unido	El gobierno del Reino Unido cuenta con el Centro de Ética e Innovación de Datos, con el cual buscan ofrecer, probar y refinar enfoques confiables para el gobierno de datos y la inteligencia artificial; y su trabajo está basado en la Estrategia Nacional de Datos que busca aprovechar los datos para brindar servicios nuevos e innovadores, esta estrategia se compone de pilares, misiones y oportunidades para el aprovechamiento de los datos en todo el Reino Unido para generar intercambio de información.
BCG	Boston Consulting Group, junto a BCG Henderson Intitute, realiza una publicación nombrada “Gobernanza simple para un ecosistema de datos”, donde destacan una serie de factores clave a la hora de entender los datos que se comparten en un

Referente	¿Qué recomendaciones ofrecen los INEs, Ministerios y Organizaciones privadas en términos de gobernanza y ética de los datos en torno a la interoperabilidad en el Sistema de estos en el Estadístico Nacional? ¿Qué ejercicios se vienen realizando?
	ecosistema, resaltando que la buena gobernanza es una cuestión de cooperación. En el documento exponen los principales problemas a la hora de generar intercambio de datos y asimismo, nombran un conjunto de reglas que guían una buena práctica. Las primeras tres reglas ayudan a las organizaciones a crear las condiciones para la autonomía individual y empoderamiento. Las otras tres obligan a los usuarios a enfrentar la complejidad y cooperar con otros para que el desempeño general del ecosistema de datos se vuelve tan importante para ellos como su propia actuación individual.
Estados Unidos	En la Estrategia Federal de Datos presentan el marco de ética de datos, el cual orienta las actividades de los organismos en materia de datos, proporcionando la base para la adquisición, gestión y uso éticos de los datos. Adicionalmente, establece los beneficios de la ética de los datos y los siete principios éticos de los datos federales para ayudar a los usuarios de los datos a tomar decisiones de forma ética y promover la responsabilidad a lo largo del ciclo de vida de los datos.
KPMG	KPMG desarrolló los documentos: i) “Uso ético de los datos de clientes en una economía digital” el cual discute los desafíos éticos clave que enfrentan las instituciones financieras en la actualidad como custodios de los datos de sus clientes, y ii) en coautoría con UK Finance elaboró el documento “los principios éticos para AAAI en servicios financieros” cuyo objetivo es ayudar a mantener la confianza del público mitigando los riesgos potenciales de las tecnologías de IA en el sector financiero. Ambos documentos proponen cinco principios éticos que se pueden aplicar como punto de referencia para fomentar o mejorar los principios internos y de gobierno propios de la organización, permitiendo desarrollar productos, servicios y aplicaciones administrativas que se basen en AAAI con un manejo ético de los datos de sus clientes.
Chile	El ministerio de Ciencia lleva a cabo la publicación de un informe referente a las alternativas para una gobernanza eficiente y ética de los datos en Chile. Este informe estuvo liderado por la Comisión Asesora de Datos de Interés Público, donde se aborda la importancia de definir un modelo de administración de datos, se expone un modelo propuesto por la OCDE, igualmente se realiza la caracterización de los usuarios que tiene un rol en toda la gobernanza. Adicionalmente, en este documento se exponen algunas prácticas internacionales como la de Suecia, Noruega, Dinamarca, Nueva Zelanda, Reino Unido y Canadá; que sirven como referente para el desarrollo propio de chile como lo es la gobernanza de datos.

Fuente: DANE a partir de las revisiones de referentes internacionales

2.3 Revisión de referentes

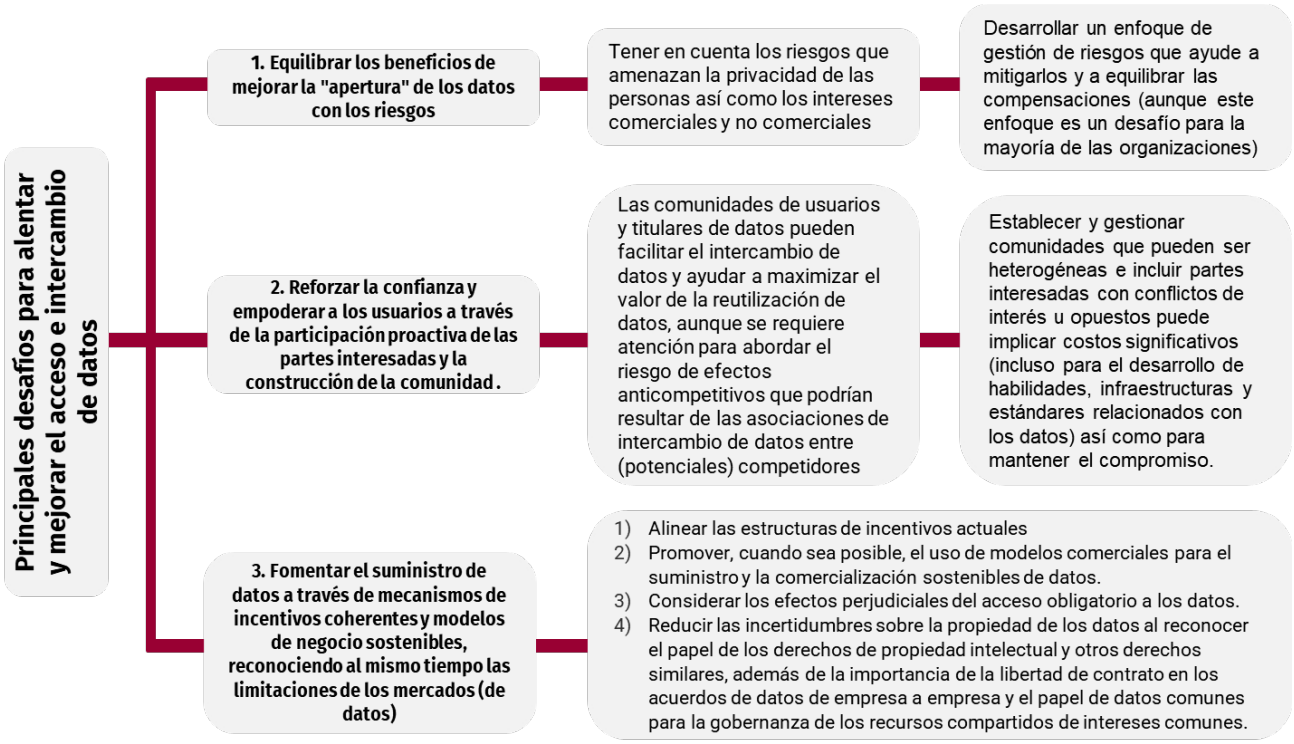
En esta sección se presentan, de forma sintetizada, la revisión de referentes internacionales.

2.1.1 OECD

En los últimos años, se ha incrementado la importancia de la innovación basada en datos gracias a la creciente demanda de estos en toda la sociedad (tanto del sector público como del privado), sumado con una mayor capacidad para recopilar, acceder, compartir y usar datos en formatos

digitales. El acceso e intercambio de datos ofrece múltiples beneficios, pues las innovaciones en diferentes sectores ayudan a resolver problemas económicos, desafíos sociales y ambientales (según la OECD se estima que el acceso e intercambio de datos del sector público generan beneficios sociales y económicos por valor de entre 0,1 % y el 1,5 % del PIB y puede llegar hasta el 4 % si se incluyen los datos del sector privado[39]), no obstante, a pesar de la creciente necesidad de datos, aún no se ha aprovechado la totalidad de su potencial, pues las personas, empresas y gobiernos a menudo se enfrentan a barreras de acceso, que son agravadas por la resistencia a compartirlos, la Ilustración 14 muestra los principales desafíos para alentar y mejorar el acceso e intercambio de datos.

Ilustración 14. Principales desafíos para alentar y mejorar el acceso e intercambio de datos



Fuente DANE a partir de OECD 2019

Para abordar estos desafíos multidimensionales los responsables de la formulación de políticas deben evitar buscar soluciones únicas para todos, pues según el informe "*Mejora del acceso y el intercambio de datos: reconciliación de riesgos y beneficios para la reutilización de datos en las sociedades, no existe un único nivel óptimo de apertura de datos; el valor del acceso y el intercambio de datos depende del tipo de datos y el contexto en el que se reutilizan, incluido el entorno social, económico y cultural en el que se llevan a cabo las actividades*[40])", dado esto los encargados de la formulación de políticas deben tener en cuenta:

39 Disponible en https://www.oecd-ilibrary.org/sites/276aaca8-en/index.html?itemId=/content/publication/276aaca8-en

40 Disponible en https://www.oecd-ilibrary.org/sites/276aaca8-en/1/1/1/index.html?itemId=/content/publication/276aaca8-en&_csp_=a1e9fa54d39998ecc1d83f19b8b0fc34&itemIGO=oecd&itemContentType=book

- El grado en el que los datos personales podrían volver a identificarse y la sensibilidad de estos.
- Los derechos e intereses detrás de las partes interesadas relevantes.
- La manera en que se generan los datos, para tener en cuenta las contribuciones de los diferentes interesados en la creación de los datos.

Además, pueden considerarse varios enfoques para alentar, facilitar y mejorar el acceso e intercambio de datos, cada enfoque representa una estrategia para abordar los principales desafíos, entre los enfoques más destacados están: i) acuerdos contractuales: enfoque basado en el mercado para mejorar el acceso e intercambio de datos en un contexto empresa a empresa, ii) datos abiertos: el enfoque con mayor apertura de datos y el más usado por los gobiernos, pero no siempre es el más apropiado para el acceso e intercambio de datos (EASD), pues los intereses públicos y privados pueden requerir un enfoque más restringido para el acceso e intercambio de datos, iii) portabilidad de los datos: este enfoque ofrece un acceso restringido a los involucrados en la creación, recopilación de datos y a los interesados en estos (los usuarios tienen más control sobre los datos, por lo que pueden exponerse a nuevos riesgos), iv) acuerdos restringidos de intercambio de datos: enfoque usado en datos demasiado confidenciales (intereses comerciales, atención médica, entre otros).

Asimismo, la mejora del acceso e intercambio de datos (EASD) conlleva a múltiples riesgos de confidencialidad y privacidad para las personas y organizaciones, por lo que los beneficios de esta mejora deben equilibrarse con los costos, intereses y derechos (públicos y privados) de las partes involucradas en especial cuando trata de datos confidenciales, pues si no se protegen los incentivos para aportar datos e invertir en innovación basada en estos pueden verse debilitados (la evidencia confirma que los riesgos de violación de confidencialidad han llevado a los usuarios a tomar una actitud reacia para compartir sus datos).

Sin embargo, el desarrollo tecnológico ahora supera con creces el establecimiento de estándares legales, lo que crea brechas regulatorias[41], para ello se han puesto en marcha algunas iniciativas de arreglos institucionales para equilibrar los riesgos y beneficios de un mayor acceso e intercambio de datos como por ejemplo la *Legislación australiana sobre intercambio y publicación de datos DS&R*[42] propone una serie de arreglos institucionales para balancear los riesgos y mejorar la confianza en el intercambio y reutilización de datos; además, la OECD publicó la *Recomendación del Consejo sobre la Gobernanza de Datos de Salud*[43] la cual consiste en 12 principios de alto nivel, que establecen las condiciones para fomentar una mayor armonización entre países de los marcos de gobernanza de datos, para que así más países puedan hacer uso de los datos de salud para la mejora de la calidad médica, investigación y estadísticas, además, es indispensable incentivar la creación de comunidades de partes interesadas para el intercambio y reutilización de datos, pues la participación de la comunidad puede ayudar a asignar responsabilidades y definir los niveles de riesgo aceptables. La manera en la que se estructura y gobierna la comunidad refleja la

41 Disponible en https://www.oecd-ilibrary.org/sites/947717bc-en/index.html?itemId=/content/publication/947717bc-en&_csp_=592b0175998ee36de7bd22e602c0d73e&itemIGO=oecd&itemContentType=book

42 Disponible en https://www.pmc.gov.au/resource-centre/public-data/issues-paper-data-sharing-release-legislation

43 Disponible en https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0433?_ga=2.202600404.458767222.1652297696-1611880960.1643989201

heterogeneidad y los intereses opuestos de las partes que los formuladores de políticas deben gestionar al momento de desarrollar marcos de gobernanza de datos[44].

Dado lo anterior, deben establecerse marcos de gobernanza de datos que promuevan un entorno en el que el acceso e intercambio de datos sean confiables y respondan a políticas públicas y objetivos sociales específicos basados en la ética, el estado de derecho, la protección de los derechos humanos y de privacidad, poniendo a las personas y comunidades en el centro de las decisiones sobre los datos que les conciernen a los que acceden, comparten o usan tanto del sector público como del privado, por lo que es indispensable la cooperación y confianza entre las partes interesadas para la creación de valor compartido en el ecosistema de datos.

Por esta razón, el pasado 6 octubre del 2021 la OECD adoptó la "Recomendación de la OECD sobre la mejora del acceso e intercambio de datos[45] – EASD" (el cual es el primer conjunto de principios y orientación política acordado internacionalmente sobre la maximización de beneficios intersectoriales de todos los tipos de datos) cuyo objetivo es determinar principios generales y orientación política para que los gobiernos maximicen los beneficios de mejorar los acuerdos de acceso e intercambio de todos los tipos de datos (privados, públicos, abiertos, personales y no personales), a la vez que se protegen los derechos de las personas y organizaciones, la Tabla 9 Ilustra los principios de la recomendación con sus respectivas acciones:

Tabla 9. Principios y orientación política para la maximización de beneficios de todos los tipos de datos

Principio	Descripción	Acciones
Reforzar la confianza en todo el ecosistema de datos	Empoderar e involucrar de manera proactiva a todas las partes interesadas relevantes junto con esfuerzos más amplios para aumentar la confiabilidad del ecosistema de datos antes y durante el establecimiento y la implementación de medidas políticas para mejorar el acceso y el intercambio de datos.	• Promover la representación inclusiva e involucrar a las partes interesadas relevantes en el ecosistema de datos, durante el diseño, implementación y el monitoreo de los marcos de gobernanza de datos. • Fomentar asociaciones de intercambio de datos neutrales a la competencia, incluidas las asociaciones público-privadas (PPP), donde cree valor adicional para la sociedad, evitando conflictos de intereses o socavar los acuerdos gubernamentales de datos abiertos o los intereses públicos. • Fomentar la adopción de prácticas responsables de gobierno de datos a lo largo del ciclo de valor de estos, para que cumplan con las normas y obligaciones legales aplicables, reconocidas y ampliamente aceptadas, incluidos los códigos de conducta, los principios éticos, la privacidad y regulación de protección de datos. • Empoderar a individuos, grupos sociales y organizaciones a través de mecanismos e instituciones apropiados, como terceros de confianza, que aumentan su agencia y control sobre los datos que han aportado o que se relacionan con ellos, y les

44 Disponible en Disponible en https://www.oecd-ilibrary.org/sites/276aaca8-en/1/2/5/index.html?itemId=/content/publication/276aaca8-en&_csp_=a1e9fa54d39998ecc1d83f19b8b0fc34&itemIGO=oecd&itemContentType=book#endnotea4z3

45 Disponible en https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463

		permite reconocer y generar valor a partir de los datos de manera responsable y efectiva.
	Adoptar un enfoque estratégico de todo el gobierno para el acceso y el intercambio de datos para garantizar que los arreglos para el acceso y el intercambio de datos ayuden a cumplir de manera efectiva y eficiente los objetivos sociales, políticos y legales específicos que son de interés público.	• Priorizar los acuerdos de acceso e intercambio de datos, teniendo en cuenta las leyes y regulaciones aplicables, trabajando junto con las partes interesadas clave para definir claramente el propósito de estos arreglos e identificar los datos relevantes para estos propósitos, teniendo en cuenta sus beneficios, costos y posibles riesgos. • Adoptar y revisar regularmente marcos de gobernanza de datos coherentes, flexibles y escalables, incluidas las estrategias nacionales de datos, que integran cuestiones transversales de gobernanza económica, social, cultural, técnica y legal, para fomentar el acceso y el intercambio de datos dentro y entre la sociedad sectores público, privado, y jurisdicciones. • Adoptar entornos legales y normativos ágiles y neutrales desde el punto de vista tecnológico que promuevan el acceso y el intercambio responsable de datos y permitan la innovación normativa, al mismo tiempo que brindan la seguridad y la protección jurídicas necesarias con la participación de todas las autoridades de aplicación, organismos de supervisión y grupos de partes interesadas independientes pertinentes.
	Maximizar los beneficios del acceso e intercambio de datos, mientras se protegen los derechos de las personas y las organizaciones y se tienen en cuenta otros intereses y objetivos legítimos, junto con esfuerzos más amplios para promover y habilitar una cultura de responsabilidad para el gobierno de datos.	• Alentar acuerdos de acceso e intercambio de datos que aseguren que los datos sean lo más abierto posible para maximizar sus beneficios y tan cerrados como sea necesario para proteger los intereses públicos y privados legítimos. • Tomar las medidas necesarias y proporcionadas para proteger estos intereses públicos y privados legítimos como condición para el acceso y el intercambio de datos, garantizando que las partes interesadas estén plenamente informadas sobre sus derechos, responsabilidades y obligaciones respectivas en caso de violaciones de la privacidad, los derechos de propiedad intelectual, las leyes de competencia u otros derechos y obligaciones. • Garantizar que las partes interesadas rindan cuentas al asumir la responsabilidad, de acuerdo con sus roles, por la calidad de los datos que comparten y por la implementación sistemática de medidas de gestión de riesgos a lo largo del ciclo de valor de los datos. • Fomentar la adopción de acuerdos condicionados de acceso e intercambio de datos con el uso de entornos y métodos tecnológicos y organizacionales, incluidos mecanismos de control de acceso a datos y tecnologías de mejora de la privacidad, a través de los cuales se puede acceder y compartir datos de manera segura entre usuarios aprobados.
Estimular la inversión en	Proporcionar mecanismos de	• Fomentar mercados competitivos para los datos a través de políticas y regulaciones de competencia sólidas que aborden la

datos e incentivar el acceso y el intercambio de datos	incentivos coherentes y promover las condiciones para el desarrollo y la adopción de modelos comerciales y mercados sostenibles para el acceso y el intercambio de datos.	posible explotación del dominio del mercado y otras medidas apropiadas. • Promover, cuando corresponda, mecanismos de autorregulación o corregulación que brinden flexibilidad jurídica, al tiempo que garantizan que todas las partes interesadas relevantes tengan certeza sobre las leyes y reglamentos aplicables. • Apoyar inversiones a largo plazo en acuerdos de acceso e intercambio de datos para garantizar su sostenibilidad, incluso en acuerdos de datos abiertos. • Promover mecanismos de incentivos apropiados que permitan la distribución justa de los beneficios de los acuerdos de acceso e intercambio de datos. • Apoyar el desarrollo y la ampliación de nuevos modelos comerciales y áreas de aplicación para el acceso y el intercambio de datos a través de una combinación de políticas para la innovación.
Fomentar el acceso, el intercambio y el uso de datos efectivos y responsables en toda la sociedad	Mejorar aún más las condiciones para el acceso transfronterizo de datos y el intercambio con confianza.	• Evaluar y en la medida de lo posible, minimizar las restricciones al acceso y el intercambio de datos transfronterizos, en particular para fines de interés público mundial. • Garantizar que las medidas que condicionan el acceso y el intercambio de datos transfronterizos no sean discriminatorias, transparentes, necesarias y proporcionadas al nivel de riesgo, teniendo en cuenta, entre otros, la sensibilidad de los datos, el propósito y el contexto del acceso a los datos a compartir y usar. • Promover el diálogo continuo y la cooperación internacional sobre formas de fomentar el acceso y el intercambio de datos entre jurisdicciones, así como la interoperabilidad y el reconocimiento mutuo de los acuerdos de acceso e intercambio de datos, teniendo en cuenta los requisitos legales aplicables y los estándares globales.
	Fomentar la capacidad de búsqueda, accesibilidad, interoperabilidad y reutilización de datos entre organizaciones, incluso dentro y entre los sectores público y privado.	• Esforzarse por garantizar que los datos se proporcionen junto con los metadatos, la documentación, los modelos de datos y los algoritmos requeridos de manera transparente y oportuna, respaldados por mecanismos de control de acceso a datos adecuados, incluidas las interfaces de programación de aplicaciones (API). • Evaluar y promover el desarrollo y la adopción de especificaciones interoperables para el acceso, el intercambio y el uso efectivo de datos, incluidos estándares comunes para formatos y modelos de datos, así como implementaciones de código abierto.
	Mejorar la capacidad de todas las partes interesadas para usar los datos de manera efectiva y responsable a lo largo	• Fomentar la conciencia sobre los beneficios y riesgos del acceso, el intercambio y el uso de datos para fomentar la gobernanza responsable de los datos a lo largo del ciclo de valor de los datos al entablar diálogos con todos los grupos de partes interesadas y asociaciones relevantes, difundiendo buenas

	del ciclo de valor de los datos.	prácticas sobre el acceso, el intercambio y el uso de datos que ayuden a superar las barreras para acceder, compartir y aumentar la capacidad de las personas y organizaciones para administrar, acceder, compartir y usar datos de manera responsable • Promover el desarrollo de las habilidades y competencias relacionadas con los datos necesarias, incluso por parte de los trabajadores y servidores públicos, para aprovechar los beneficios del acceso, el intercambio y el uso de datos a lo largo del ciclo de valor de los datos de manera coherente con el enfoque estratégico para el acceso y el intercambio de datos. • Facilitar el acceso y la adopción de infraestructuras fundamentales, sostenibles, abiertas, escalables, seguras y protegidas necesarias a lo largo del ciclo de valor de los datos, incluso para la conectividad, el almacenamiento y la informática, mediante la promoción de prácticas de gestión de riesgos de seguridad digital a lo largo del ciclo de valor de los datos.

Fuente DANE a partir de OECD 2021

Finalmente, la OECD publicó en el 2021 el documento "Portabilidad de datos, interoperabilidad y competencia de plataformas digitales[46], el cual describe el papel que desempeñan las medidas de interoperabilidad y portabilidad de datos entre las plataformas digitales, destacando el papel de los mecanismos de implementación en la efectividad de estas medidas y reconociendo la importancia de una autoridad de competencia o de un tercero independiente para establecer estándares de interoperabilidad y adjudicar disputas.

2.1.2 Comisión Económica para América Latina – CEPAL

La Comisión Económica para América Latina y el Caribe – CEPAL, resalta la importancia de la gobernanza digital y la interoperabilidad gubernamental en el desarrollo de los países, es por ello que realizó actividades de asistencia técnica sobre gobernanza digital e interoperabilidad gubernamental[47], prestada al Ministerio de Ciencia, Tecnología y Telecomunicaciones (MICITT) de Costa Rica. Creó una hoja de ruta y recomendaciones adicionales sobre cómo robustecer los procesos en las instituciones públicas para ofrecer mayor interoperabilidad entre las instituciones y sus sistemas de información, permitiendo el desarrollo y elaboración de trámites, mejorando el acceso a la información y fortalecimiento la transparencia de los procesos.

Desde CEPAL, se aborda la gobernanza digital por medio de dos conceptos, el primero la define como la articulación y concreción de políticas de interés público con los diversos actores involucrados (Estado, sociedad civil y sector privado), con la finalidad de alcanzar competencias y cooperación para crear valor público y la optimización de los recursos de los involucrados, mediante el uso de tecnologías digitales; la segunda la define como la organización y reglas presentes en un

46 Disponible en https://www.oecd.org/daf/competition/data-portability-interoperability-and-digital-platform-competition-2021.pdf

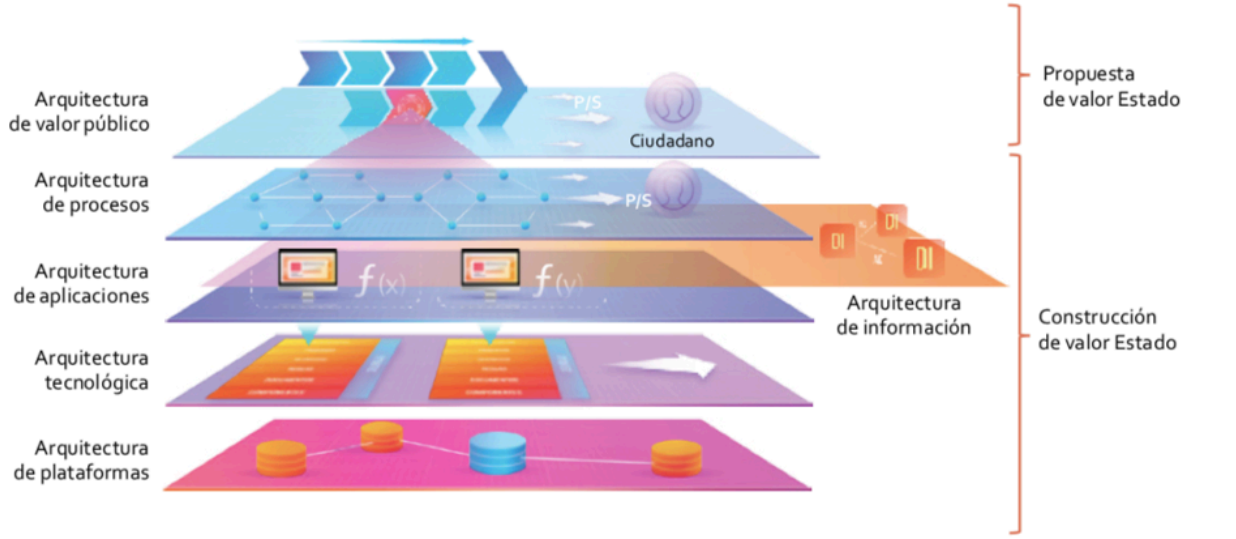
47 Disponible en https://repositorio.cepal.org/bitstream/handle/11362/47018/1/S2100258_es.pdf

gobierno para conducir su política y estrategia de digitalización con el objetivo de mejorar su gestión y brindar servicios a los ciudadanos y las empresas.

Implementación de la gobernanza digital

El éxito de una institución pública, consiste en gestionar sus recursos de información cumpliendo el principio de legalidad, respetando los marcos jurídicos, las disposiciones legales y regulatorias relacionadas con el acceso a los datos, la identificación de los ciudadanos interesados, mejorando los datos abiertos y el acceso digital, en este contexto, la gobernanza digital permite desarrollar mecanismos, procesos e instituciones a través de los cuales los actores articulen sus intereses y mejoren sus procesos, por medio de un enfoque holístico del Estado y su capacidad de generar valor público, que permita visibilizar los puntos de contacto con la ciudadanía, las empresas y las organizaciones. Para alcanzar esta visión integrada se recomienda adoptar un modelo desarrollado por The Open Group en 2020, denominado de arquitectura institucional a nivel del estado, como se muestra a continuación:

Ilustración 15. Modelo de gobernanza digital con base en la arquitectura institucional



Fuente CEPAL – Gobernanza digital e interoperabilidad gubernamental, una guía para su implementación[48].

Cada institución debe identificar cuál es su papel en relación con la ciudadanía, las organizaciones y otras instituciones, lo cual permitirá tener un panorama integrado de cómo se relaciona con otras instituciones en la generación de valor público. En este contexto, cualquier Estado que quiera abordar la interoperabilidad desde una perspectiva holística debe tomar como punto de partida la gobernanza digital, ya que esta generará la política, la estrategia, los recursos y el apoyo institucional.

Interoperabilidad

48 Disponible en https://repositorio.cepal.org/bitstream/handle/11362/47018/1/S2100258_es.pdf

Según la Comisión Europea, la interoperabilidad es la capacidad de que las organizaciones interactúen con vistas a alcanzar objetivos comunes que sean mutuamente beneficiosos y que hayan sido acordados de manera previa y conjunta. Para ello recurren a la puesta en común de información y conocimientos a través de los procesos institucionales que apoyan, mediante el intercambio de servicios, datos o documentos entre sus respectivos sistemas de tecnologías de la información y las comunicaciones (TIC).

La interoperabilidad en la gobernanza digital, requiere un manejo adecuado de políticas, personas, procesos y tecnologías que, abordado de la mano de conceptos y métodos probados, pueda generar valor, sostenibilidad y beneficios colaterales a la gestión integral de las políticas, es un requisito para hacer posible la comunicación digital y el intercambio de información entre las administraciones públicas, y entre estas y las empresas privadas y los organismos no gubernamentales que deban interactuar con el Estado, con la finalidad de lograr un mercado digital único.

Gestión de cambio para la interoperabilidad

Se hace necesario identificar las barreras que no permiten el pleno desarrollo de la interoperabilidad, por ejemplo: i) las barreras de competencias, cuando las instituciones no cuentan con personal con las competencias para impulsar el proceso, ii) barreras tecnológicas, cuando las tecnologías de información de las instituciones y organizaciones no son compatibles para procesar el intercambio de datos, iii) barreras conceptuales, cuando se tienen diferentes conceptualizaciones e interpretaciones en las partes que interoperan, etc. Debe hacerse cargo de dichas barreras, analizarlas y emprender acciones para evitarlas. Asimismo, se deben identificar los facilitadores del cambio, los rasgos, características, personas o situaciones de la organización que pueden permitir, acelerar o instalar el cambio que se desea, el objetivo es hacer participar al personal de la institución en el proceso de transformación, definiendo las mejores soluciones para materializar las iniciativas y asimilar las mejoras resultantes del proceso.

Los cambios derivados de las iniciativas de interoperabilidad generan distracciones que resultan en pérdida de productividad de los equipos de trabajo, por ello, se proponen 3 planes:

1. Plan de contención: Identificar los factores que obstaculizan y facilitan el cambio y los efectos que tendrá la implementación de las iniciativas de interoperabilidad en la organización y las personas.
2. Plan de formación/Entrenamiento: Las personas son el principal agente de cambio, el ámbito de la transferencia de conocimiento se debe hacer cargo del conocimiento técnico y también de los componentes adaptativos que implica el cambio.
3. Plan de comunicación y difusión: Se debe formular e implementar un plan de comunicación y difusión del proyecto derivado en un modelo de gestión de cambio que se aplique a todos los proyectos involucrados en las iniciativas de interoperabilidad.

Modelo de gobernanza digital en Costa Rica

El modelo de gobernanza digital propuesto para Costa Rica, aborda recomendaciones del Consejo de la OCDE, sobre estrategias de gobierno digital en sentido de:

- Asegurar el liderazgo y el compromiso político con la estrategia mediante una combinación de esfuerzos encaminados a promover la coordinación y la colaboración interministerial, establecer prioridades y facilitar la participación y coordinación de los organismos pertinentes en todos los niveles de gobierno.
- Garantizar el uso coherente de las tecnologías digitales en todas las áreas de política y los niveles de gobierno mediante la integración de la estrategia de gobierno digital en las reformas generales de la administración pública, la participación de todos los niveles de gobierno en el desarrollo de la estrategia de gobierno digital, la identificación de la complementariedad, la alineación y el refuerzo mutuo entre la estrategia de gobierno digital y otras estrategias sectoriales relevantes, etc.
- Establecer marcos organizativos y de gobernanza eficaces para coordinar la implementación de la estrategia digital en los niveles de gobierno y entre ellos.
- Desarrollar casos de negocio claros para sustentar el financiamiento y la implementación focalizada de proyectos de tecnologías digitales, mediante la identificación de los beneficios económicos, sociales y políticos esperados con el fin de justificar las inversiones públicas y la participación de todos los actores.
- Fortalecer las capacidades institucionales para gestionar y monitorear la implementación de proyectos.
- Adquirir tecnologías digitales basadas en la evaluación de los activos existentes, incluidas las habilidades digitales, los perfiles laborales, las tecnologías, los contratos y los acuerdos interinstitucionales, para aumentar la eficiencia, apoyar la innovación y sustentar mejor los objetivos establecidos en la agenda general de modernización del sector público.

La Contraloría General de la República de Costa Rica define los énfasis en fortalecimiento de la gobernanza digital en ocho pilares, y propone un abordaje integral de la transformación digital que incorpore elementos propios de la gestión pública al proceso de cambio:

1. Fortalecimiento del modelo de gobernanza digital.
2. Transformación hacia una gestión pública eficiente.
3. Definición clara de roles y compromisos.
4. Fortalecimiento de la participación ciudadana.
5. Transparencia a través de gobierno abierto.
6. Política pública con base en evidencia científica.
7. Consolidación del sistema de innovación pública.
8. Institucionalidad adaptable.

2.1.3 Banco Interamericano de Desarrollo – BID

Para el Banco Interamericano de Desarrollo, BID, en el sector público, la disponibilidad de los datos y la mejora de las capacidades técnicas para utilizarlos, puede servir para mejorar la eficiencia en la prestación de servicios sociales, políticos, de transporte, como para promover la transparencia, la participación ciudadana y la rendición de cuentas.

La Organización Europea para la Cooperación y el Desarrollo Económico, con la colaboración del Consejo de Europa, publicó una guía denominada “La gestión ética de los datos”[49] para resguardar la privacidad y los flujos transfronterizos de datos personales a partir de los siguientes principios:

- Establecer límites claros para la obtención de datos.
- Determinar la relevancia de los datos para el uso previsto.
- Definir con claridad el uso que se dará a los datos antes de solicitarlos.
- No utilizar los datos para usos distintos al determinado originalmente sin el consentimiento de las personas afectadas.
- Asegurarse de proteger los datos contra el acceso ilícito o piratería.
- Asegurar que los avances, prácticas y políticas sobre el uso de datos sean abiertos y transparentes.
- Garantizar que las personas cuyos datos se han recolectado tengan acceso a los mismos y puedan solicitar modificaciones o su eliminación definitiva.

El Reglamento General de Protección de Datos – RGPD, se elaboró como respuesta al reto que plantea el uso intensivo de datos en la sociedad y se ha constituido como una especie de estándar global en términos de legalidad, buenas prácticas, proporcionalidad, limitaciones, transparencia y responsabilidad.

La transformación digital hace referencia al modo en que las tecnologías cambian las reglas de participación y la forma en que las personas trabajan, interactúan y piensan. La digitalización fortalece el uso de herramientas digitales para automatizar o almacenar información en formato digital sin rediseñar los procesos existentes por medio de tecnologías de información.

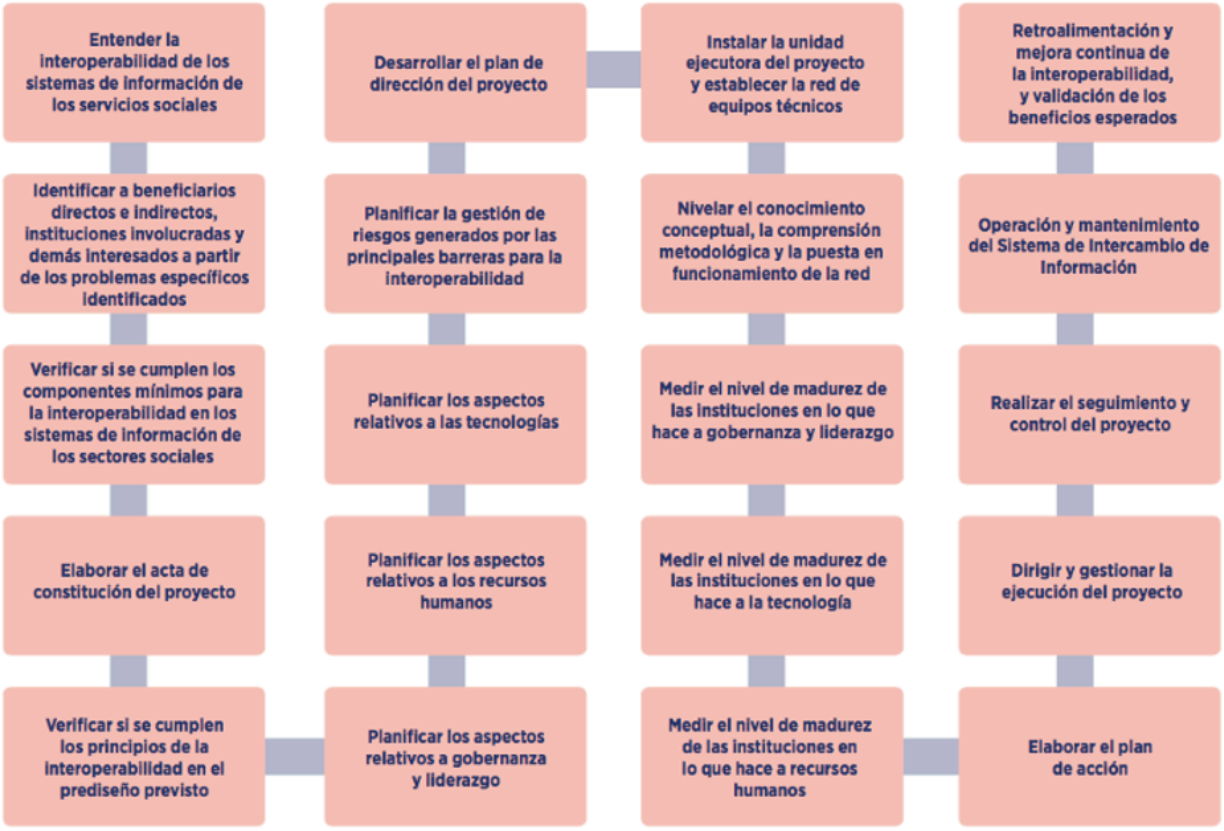
El proceso de transformación se establece cuando el gobierno logra interconectar sus entidades, los mecanismos de intercambio de información común permiten que los usuarios de la entidad cuenten con acceso a la información de manera continua en múltiples canales, lo cual mejora la calidad del servicio que reciben, se reducen los costos para las entidades y el ciudadano, se promueve la transparencia, etc.

En el desarrollo del proceso de interoperabilidad, se propone establecer dos marcos:

i) Marco conceptual que determine por qué y para qué se debe contar con sistemas de intercambio de información interoperables en el sector social. Esto permitirá tener una visión amplia de elementos políticos, normativos, de gestión, técnicos y funcionales en el ámbito nacional.
ii) Marco metodológico que indique cómo poner en práctica el marco conceptual y cómo desarrollar las plataformas o sistemas interoperables.

49 Disponible en https://publications.iadb.org/publications/spanish/document/La_Gestión_Ética_de_los_Datos.pdf

Ilustración 16. Guía de uso de la metodología



Fuente BID – La gestión ética de los datos[50].

2.1.4 Unión Europea

Se prevé que la cantidad de datos generados por los organismos públicos, las empresas y los ciudadanos de la Unión Europea se quintuplique entre 2018 y 2025, ante tal fenómeno, los órganos de esta organización han venido trabajando en múltiples políticas, regulaciones, procedimientos y mecanismos para el tratamiento, procesamiento y divulgación de estos. Durante el 2020, la Comisión Europea propuso un nuevo esquema normativo para la gobernanza de los datos, cuyo principal objetivo es facilitar el intercambio de datos en toda la Unión Europea y entre sectores, a la vez que se aumenta la confianza en los datos, esperando así generar mayor provecho de los datos e información[51].

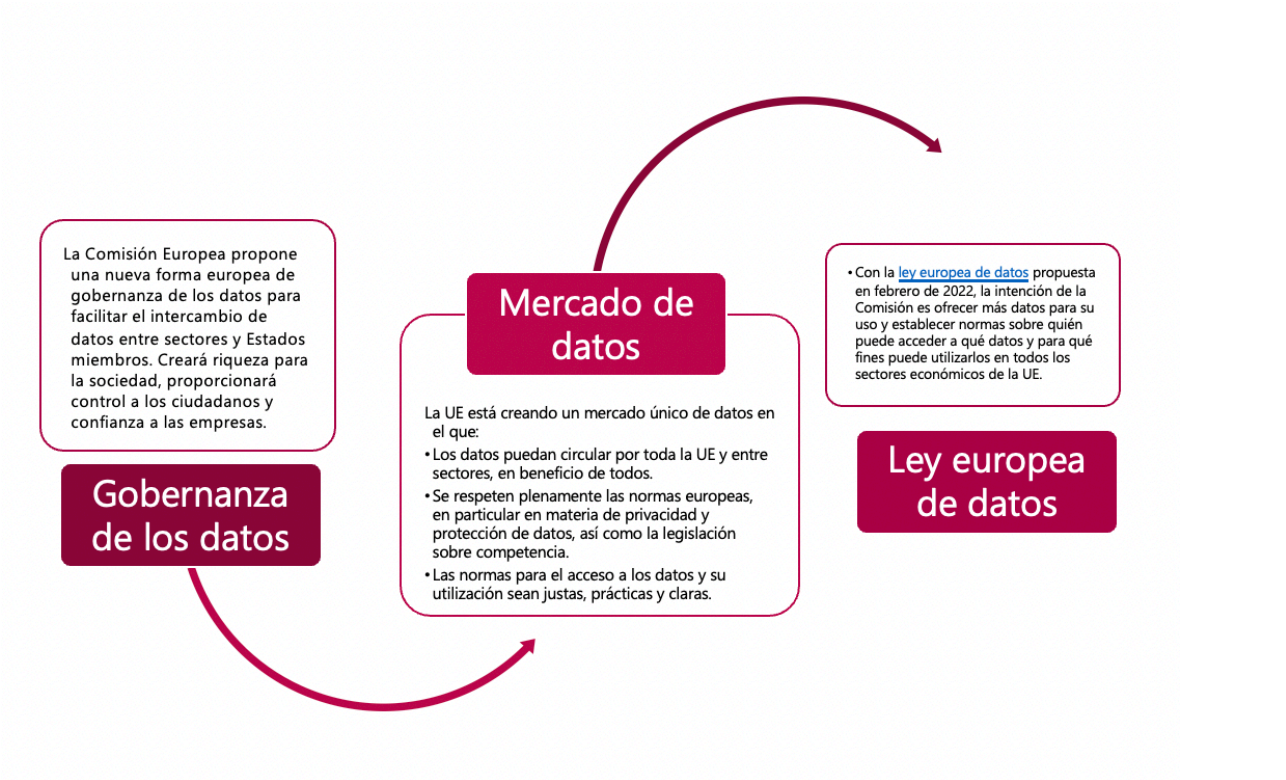
Es importante resaltar que este esquema normativo hace parte de la *"estrategia de datos europea"* (Ilustración 17 Estrategia de datos europea) la población de la Unión Europea en una sociedad impulsada por los datos, la creación de un mercado único de datos que permitirá a estos fluir libremente entre sectores, en beneficios de las empresas, los investigadores y las administraciones públicas[52].

50 Disponible en https://publications.iadb.org/publications/spanish/document/La_Gestión_Ética_de_los_Datos.pdf

51 Disponible en: https://ec.europa.eu/commission/presscorner/detail/es/ip_20_2102

52 Disponible en: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_es

Ilustración 17 Estrategia de datos europea



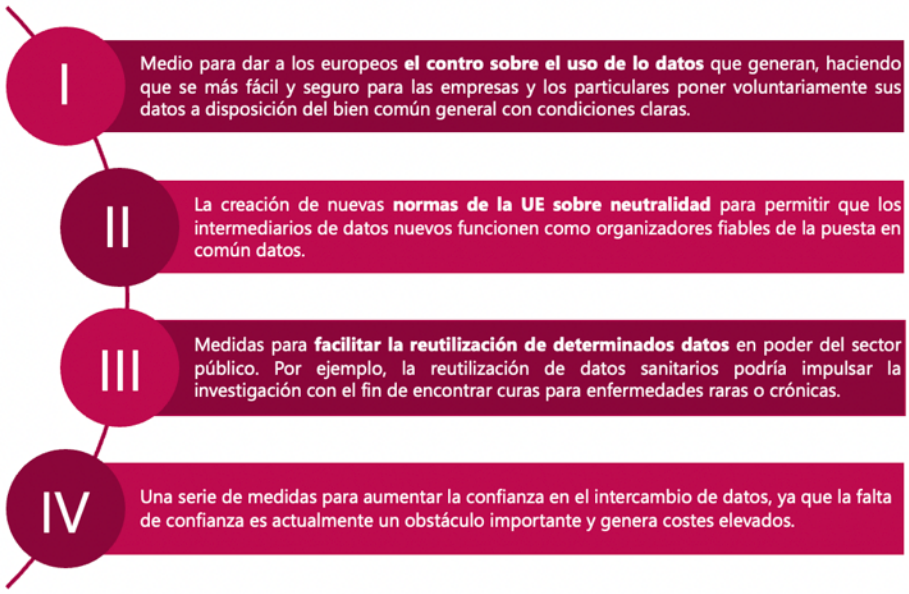
Fuente: DANE con base en la Unión Europea[53]

La Comisión considera que una gobernanza de datos con altos estándares de calidad, herramientas eficientes y un marco normativo y regulatorio adaptado permitirá crear riqueza para la sociedad, aumentar el control y la confianza de los ciudadanos y las empresas en los datos. Adicionalmente, consideran que, con este conjunto de medidas, se puede establecer un modelo alternativo a las prácticas de tratamiento de datos de las principales plataformas tecnológicas y apoyar la "*estrategia de datos europea".* En tanto al modelo, este regula el intercambio de datos entre empresas a cambio de una remuneración en cualquier forma, así como el uso de datos fines altruistas, de igual manera, busca hacer que los datos del sector público sean reusables y finalmente, que los datos se puedan usar siempre y cuando tenga ayuda de un "intermediario de intercambio de datos personales" que conozca a fondo el Reglamento de Protección de Datos[54].

Finalmente, el reglamento sentó las bases para una forma de gobernanza de los datos que esté en consonancia con los valores y principios de la Unión Europea. Una estructura que proteja los datos, a la vez que, a los consumidores, respetando siempre las normas de competencia. La neutralidad y transparencia de datos son pilares fundamentales de este reglamento, el cual está proyectado tal y como es expuesto la ilustración 18.

[53] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_es
[54] Disponible en: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

Ilustración 18 Proyección reglamento de protección de datos



Fuente: DANE con base en la Comisión Europea[55].

2.1.5 Canadá

Statistics Canada, cuenta con un centro de confianza de estadísticas de Canadá[56], encargado de encriptar y anonimizar todos sus datos, ya que están comprometidos con proteger la privacidad y salvaguardar la confidencialidad de los datos que se les confían. Para liderar con integridad, Statistics Canada cuenta con una Secretaría de Ética de Datos y con consejos asesores, especialmente el Consejo Asesor de Ética y Modernización de Acceso a Microdatos.

El papel de la Secretaría de Ética de Datos para la adquisición de datos no tradicionales[57]
Históricamente, la recopilación de información por parte de Statistics Canada ha seguido un enfoque tradicional centrado en encuestas. Se crea una encuesta, una nota en la parte superior de la encuesta les dice a los posibles encuestados cómo se usarían sus respuestas y los posibles encuestados pueden decidir si quieren participar. En el pasado, Statistics Canada se basaba principalmente en encuestas para generar estadísticas oficiales.
En respuesta a la disminución de las tasas de respuesta, el aumento de los costos de las encuestas y el aumento de la demanda de datos oportunos para informar la toma de decisiones, Statistics Canada ha explorado nuevos métodos para recopilar información gracias a los avances tecnológicos.

Es por eso que Statistics Canada cambió su enfoque de un enfoque de encuesta primero a un enfoque que prioriza el uso de fuentes de datos no tradicionales o alternativas, que incluyen datos como datos de observación de la tierra, datos de escáner y datos administrativos (es decir, certificados de nacimiento y defunción los registros de salud y educación, y los registros

[55] Disponible en: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767
56 Disponible en https://www.statcan.gc.ca/en/trust
57 Disponible en https://www.statcan.gc.ca/en/trust/integrity

relacionados con el flujo de bienes y personas a través de las fronteras son ejemplos de datos administrativos).

El uso de este tipo de datos ha sido el núcleo de las recientes iniciativas de modernización en Statistics Canada. Junto con el uso de fuentes de datos no tradicionales, también se han producido avances tecnológicos. Por ejemplo, el aprendizaje automático, la inteligencia artificial y la capacidad de raspado web son técnicas que se pueden utilizar para obtener y comprender datos. Pero con esta innovación han surgido nuevas cuestiones éticas. El hecho de que sea posible hacer algo no significa que debamos hacerlo, incluso con un mandato legal para hacerlo.

En primer lugar, el uso de datos alternativos a menudo significa utilizar la información para un propósito diferente al que se pretendía originalmente cuando se recopilaron los datos por primera vez. Esto podría verse como una superación de los límites de privacidad cuando los datos se refieren a información personal. Para utilizar estos datos, Statistics Canada debe asegurarse de que el nivel de información que se busca sea proporcional a los beneficios esperados de la información que se recopila.

En segundo lugar, al acceder a tanta información específica de diferentes partes del país, por ejemplo, datos relacionados con la salud, la educación o la criminalidad de una comunidad, pueden surgir descripciones detalladas de las comunidades. La agencia quiere asegurarse de que esta información dé voz a las comunidades y no estigmatice ni difunda estereotipos dañinos. Esto es especialmente importante cuando se recopila información sobre comunidades vulnerables o marginadas.

Para enfrentar estos desafíos, en 2019, Statistics Canada adoptó el Marco de Necesidad y Proporcionalidad, bajo este marco, cada propuesta para un nuevo proyecto o adquisición de datos debe explicar por qué es importante, cuáles son los beneficios para los canadienses, quién necesita la información y abordar consideraciones éticas como privacidad, transparencia y equidad.

Este marco también garantiza que la agencia solo adquiera la información que necesita, siempre tenga un propósito específico para la información y que cada proyecto tenga en cuenta los mejores intereses de la gente de Canadá, lo que garantiza una cultura de ética por diseño. La confianza es la base de nuestro trabajo en Statistics Canada y nuestra Secretaría de Ética de Datos ayuda a garantizar que cualquier posible inquietud ética se aborde a medida que surja.

Cada propuesta para un nuevo proyecto o adquisición de datos se somete a una revisión ética por parte de la Secretaría de Ética de Datos y se prueba en seis principios rectores:

1) Beneficios para los canadienses
2) Privacidad y seguridad
3) Transparencia y rendición de cuentas
4) Confianza y sostenibilidad
5) Calidad de los datos
6) Justicia y no hacer daño

Todas las nuevas solicitudes obligatorias de información y los detalles pertinentes se publican en el sitio web antes de que se realice la solicitud para garantizar la transparencia y la apertura con los canadienses. Solo después de esta rigurosa revisión ética, estas actividades se llevan a cabo.

La producción de estadísticas e investigaciones valiosas para beneficiar a los canadienses comienza trabajando en estrecha colaboración con el proveedor de datos. El pueblo de Canadá puede confiar en que la información recopilada de ellos y sobre ellos se hace para ellos, y que estas actividades se llevan a cabo con integridad y los más altos estándares éticos.

Consejo Asesor de Ética y Modernización del Acceso a Microdatos[58]
El Consejo proporciona a Statistics Canada la orientación adecuada sobre acceso a datos, privacidad y gobierno de datos para mantener y respaldar las necesidades de datos de los canadienses. El conocimiento y la experiencia que aportan los miembros del Consejo Asesor beneficiarán a la agencia a medida que Statistics Canada trabaja para facilitar el acceso a microdatos anónimos para investigadores, mejorar la seguridad de los datos y los protocolos de gestión de riesgos. El consejo se reunirá dos veces al año y los informes se pondrán a disposición del público.

2.1.6 España

En 2020, España crea la *Oficina de Datos*[59] como órgano administrativo que depende directamente de la Secretaría de Estado de Digitalización e Inteligencia Artificial. La Oficina tiene como líneas estratégicas el diseño, coordinación y seguimiento del modelo de referencia arquitectónico para fomentar la recolección, la gestión e intercambio de datos públicos, lo que abarca aspectos de tecnología, estándares, mejores prácticas, gobernanza, codificación, seguridad y privacidad de la información.

Con la Oficina de Datos[60] se espera responder en diferentes ámbitos, tales como:

- El desarrollo y seguimiento de políticas públicas basadas en datos
- El Gobierno Abierto y Transparencia
- La producción de datos para datos.gob.es. Así como, la innovación empresarial con base en la gestión y el uso de los datos
- Habilitar la colaboración privada en pro del interés público y la creación de espacios industriales sectoriales para impulsar la Economía del Dato

Lo anterior, como base para el robustecimiento de la infraestructura de datos con el propósito de apoyar las políticas digitales del Gobierno, además de, las medidas y programas contempladas en España Digital 2025[61].

En este sentido, se proponen los principios rectores para el dato y su infraestructura, entre los que se encuentran: la disponibilidad, la facilidad de ser encontrados, la accesibilidad, la seguridad/

58 Disponible en https://www.statcan.gc.ca/en/about/relevant/acemma#wb-sec

59 Disponible en https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-10008

60 Disponible en https://portal.mineco.gob.es/es-es/digitalizacionIA/oficina-del-dato/Paginas/oficina-del-dato-se-digitalizacion-ia.aspx

61 Disponible en https://portal.mineco.gob.es/es-es/digitalizacionIA/es-digital-2025/Paginas/es-digital-2025.aspx

privacidad, la interoperabilidad, la interpretabilidad, la transparencia y confianza, así como los datos de calidad y que generen valor final garantizando la transversalidad, dinamismo y alta frecuencia de actualización.

Por otro lado, el Ministerio de Asuntos Económicos y Transformación Digital[62] se refiere a los principios de intercambio de información tomando como base los principios FAIR para el manejo y administración de datos científicos. Estos principios ofrecen un conjunto de cualidades precisas y medibles que una publicación de datos debería seguir para que los datos sean Encontrables, Accesibles, Interoperables y Reutilizables (del inglés FAIR – Findable, Accessible, Interoperable, and Reusable), como se detallan en la Tabla 10.

Tabla 10. Principios de intercambio de información (FAIR)

PRINCIPIO	CARACTERÍSTICA
Encontrables	Los datos y metadatos pueden ser encontrados por la comunidad después de su publicación, mediante herramientas de búsqueda. Asignarles un identificador único y persistente a los datos y los metadatos Describir los datos con metadatos de manera prolija Registrar/Indexar los datos y los metadatos en un recurso de búsqueda En los metadatos se debe especificar el identificador de los datos que se describen.
Accesibles	Los datos y metadatos están accesibles y por ello pueden ser descargados por otros investigadores utilizando sus identificadores. Los datos y los metadatos pueden ser recuperados por sus identificadores mediante protocolos estandarizados de comunicación Los protocolos tienen que ser abiertos, gratuitos e implementados universalmente El protocolo debe permitir procedimientos para la autentificación y la autorización (por si fuera necesario). Los metadatos deben de estar accesibles, incluso cuando los datos ya no estuvieran disponibles.
Interoperable	Tanto los datos como los metadatos deben de estar descritos siguiendo las reglas de la comunidad, utilizando estándares abiertos, para permitir su intercambio y su reutilización. Los datos y los metadatos deben de usar un lenguaje formal, accesible, compartible y ampliamente aplicable para representar el conocimiento Los datos y los metadatos usan vocabularios que sigan los principios FAIR Los datos y los metadatos incluyen referencias cualificadas a otros datos o metadatos
Reutilizables	Los datos y los metadatos pueden ser reutilizados por otros investigadores, al quedar clara su procedencia y las condiciones de reutilización. Los datos y los metadatos contienen una multitud de atributos precisos y relevantes Los datos y los metadatos se publican con una licencia clara y accesible sobre su uso y reutilización Los datos y los metadatos se asocian con información sobre su procedencia Los datos y los metadatos siguen los estándares relevantes que usa la comunidad del dominio concreto

Fuente: tomado de Ministerio de Asuntos Económicos y Transformación Digital. Datos.gob.es

62 Disponible en https://portal.mineco.gob.es/es-es/digitalizacionIA/oficina-del-dato/Paginas/oficina-del-dato-se-digitalizacion-ia.aspx

2.1.7 Reino Unido

El gobierno del Reino Unido cuenta con el Centro de Ética e Innovación de Datos (CDEI)[63], este es un organismo experto sobre el uso confiable de datos y tecnologías basadas en datos, incluida la inteligencia artificial. Este centro está respaldado por un Consejo Asesor conformado por expertos líderes en el mundo para ofrecer, probar y refinar enfoques confiables para el gobierno de datos e IA, trabajando con organizaciones de todo el Reino Unido. Su trabajo se alinea con tres misiones establecidas en la Estrategia Nacional de Datos[64], allí buscan aprovechar los datos para brindar servicios nuevos e innovadores e impulsar un enfoque de los datos para beneficiarse con el uso de los datos de manera responsable. Esta estrategia se construyó con la finalidad de conocer la opinión de diferentes entidades y actores fundamentales para el gobierno del Reino Unido, realizando ajustes de acuerdo con los comentarios recibidos. En la Ilustración 19 se presenta el marco de la Estrategia Nacional de Datos, la cual se compone de pilares, misiones y oportunidades para el aprovechamiento de los datos.

Ilustración 19 Marco de la Estrategia Nacional de Datos

Fuente: DANE, basado en GOV.UK

Con los pilares desean apoyar el esfuerzo global de interoperabilidad, para facilitar la combinación y el cruce de referencias de diferentes fuentes de datos; revisar la publicación de datos abiertos y los procesos de toma de decisiones para garantizar su consistencia; y apoyar el desarrollo de métricas interoperables para medir el impacto de los datos publicados.

Además, a través de este marco evidencian cinco oportunidades significativas para que los datos transformen positivamente el Reino Unido, como se evidencia en la Tabla 11.

63 Disponible en https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation/about

64 Disponible en https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy#missions

Tabla 11. Oportunidades significativas del uso de los datos

Oportunidad	Descripción
Impulsar la productividad y el comercio	Los datos son conocimiento. Teniendo acceso a gran cantidad de datos y combinando la capacidad de analizarlos mediante técnicas modernas; se obtiene mayor comprensión de lo que funciona y lo que no, tanto en venta de productos y servicios, como en hacer más eficientes los procesos.
Apoyando nuevos negocios y empleos	Las competencias en el uso de datos son cada vez más importantes para todos los aspectos de la vida, especialmente para el entorno laboral.
Aumentar la velocidad, la eficiencia y el alcance de la investigación científica	Los nuevos avances científicos impulsados por los datos tienen aplicaciones potencialmente revolucionarias en toda la economía, como el seguimiento de los riesgos para la salud pública y la ayuda a la descarbonización mediante redes energéticas más inteligentes, el mantenimiento predictivo de las infraestructuras o una mejor gestión del tráfico.
Impulsar una mejor prestación de políticas y servicios públicos	Los datos pueden revolucionar el sector público, creando servicios mejores, más baratos y con mayor capacidad de respuesta.
Por una sociedad más justa para todos	Los datos tienen un gran potencial para empoderar a las personas y a la sociedad civil, brindando beneficios que van más allá de la economía. Las organizaciones de la sociedad civil pueden estar mejor equipadas para llegar a las personas más necesitadas, en el momento en que más lo necesitan, con la mejora de los datos.

Fuente: Tomado de GOV.UK

Para el Reino Unido, la Interoperabilidad de datos es la capacidad de los servicios y productos de datos para interactuar y compartir datos. Cubre dos aspectos (1) los protocolos digitales que permiten el intercambio de datos y (2) los estándares de datos utilizados para preservar la compatibilidad mientras se procesan los datos. El aumento de la interoperabilidad entre los regímenes de protección de datos generalmente indica la voluntad de eliminar las barreras a los flujos de datos.

Adicionalmente, el gobierno ha buscado mejorar la calidad de los datos con la estrategia de calidad del Servicio de Estadística del Gobierno 2019-21 que tiene como objetivo “mejorar la calidad estadística en todo el Servicio de Estadística del Gobierno (GSS)”[65] y una es Estrategia de armonización de GSS66, que establece acciones realistas para mejorar la comparabilidad y la coherencia entre las estadísticas oficiales.

65 Disponible en https://gss.civilservice.gov.uk/policy-store/government-statistical-service-gss-quality-strategy/

66 Disponible en https://gss.civilservice.gov.uk/policy-store/gss-harmonisation-team-workplan/

2.1.8 BCG

Boston Consulting Group junto con el Instituto BCG Henderson destacan en una publicación a finales del año 2020 que el intercambio de datos se ve facilitado por los ecosistemas de datos que comprenden múltiples partes dentro y fuera de la industria de una organización[67]. Los autores resaltan que estos ecosistemas pueden llegar a superar algunas barreras importantes para el intercambio de datos, como el valor poco claro de los datos en el punto de generación y la necesidad de inteligencia colectiva para identificar y unir a los diferentes usuarios de la gobernanza con oportunidades para la creación de valor. En esta publicación se hace énfasis en entender que los datos que se comparten en un ecosistema es fundamentalmente una cuestión de cooperación, con reglas que guían buen comportamiento y establecen los términos del compromiso.

> *En general, los ecosistemas de datos son vehículos importantes para alinear organizaciones o instituciones en torno a objetivos comunes mientras les da la agilidad necesaria para innovar.*

Desde BCG se expone el marco Smart Simplicity, el cual está diseñado para ayudar a las empresas a dar sentido a la complejidad organizativa y fomentar la cooperación. Ya que básicamente, la investigación generada por BCG y el instituto de Henderson indica que los problemas con el intercambio de datos en general giran en torno a cuatro temas:

- **Confianza y Privacidad:** es generada por el temor de que los datos sean mal manejados, mal usados o mal compartidos. Tecnología deficiente, gobernanza débil y las violaciones de datos reales pueden conducir a datos que se utilizan para fines que fueron no acordados por el autor de los datos y otros en el ecosistema.
- **Costos de transacción:** estos son la base de cada intercambio de datos y estos pueden llegar a ser tanto por temas tecnológicos como de procedimiento. Los impedimentos tecnológicos incluyen mala conectividad, estándares no coincidentes y restricciones sobre interoperabilidad. Barreras de proceso puede implicar habilidades no coincidentes, complejidad organizativa o reglas ambiguas.
- **Preocupaciones Competitivas:** Uno de los panoramas que no se mapeaba en ese entonces es que las empresas u organizaciones sienten el riesgo de renunciar a una ventaja competitiva propia de ellas, con lo son los datos estratégicos. Los usuarios que contribuyen a un ecosistema se preocupan de que la información sensible pueda llegar a ser lanzada a los rivales Los nuevos usuarios digitales también puede preocuparse por la presencia de los gigantes digitales. Todos los participantes pueden preocuparse de que los orquestadores del ecosistema capturen una parte desproporcionada del valor.
- **Oportunidad financiera perdida:** existe otra parte del panorama que aún no se ha mapeado, el cual es la posibilidad del hecho de compartir datos puede provocar que las oportunidades financieras sean pasados por alto. En el documento se expone el siguiente ejemplo, los proveedores y los clientes pueden llegar a trabajar juntos para coordinar logística, racionalizar el inventario e incluso diseñar en conjunto algunos productos, pero los

67 Disponible en https://web-assets.bcg.com/67/a3/cccd5d6648a98ad0a7e5c4066e0f/bcg-simple-governance-for-data-ecosystems-nov-2020.pdf

beneficios y las inversiones pueden no acumularse uniformemente a lo largo de la cadena de suministro.

Reglas de Simplicidad Inteligente de Gobernanza de datos

Una vez identifican los grandes temas donde se presentan la mayoría de los problemas, BCG basado en el enfoque Smart Simplicit se enfoca en promover una cooperación dentro de organizaciones más complejas. Básicamente, en lugar de crear reglas una por una para apuntar a comportamientos específicos, Smart Simplicity explora el contexto de un sistema de comportamientos analizando las motivaciones de individuos.

Posteriormente, usa seis reglas simples para cambiar el contexto. Las primeras tres reglas ayudan a las organizaciones a crear las condiciones para la autonomía individual y empoderamiento. Los otros tres obligan a las personas a enfrentar la complejidad y cooperar con otros para que el desempeño general de la organización, en este caso, el ecosistema de datos se vuelve tan importante para ellos como su propia actuación individual. Igualmente, se resalta que hay reglas adicionales que los orquestadores de los ecosistemas deben considerar con respecto a la propiedad de los datos, acceso y uso, sin embargo, se expone que estas variarán según el ecosistema.

Tabla 12. Reglas de simplicidad inteligente de gobernanza de datos

Regla 1.
Entender lo que los usuarios realmente hacen. Para desenredar las barreras competitivas y de confianza dentro de un ecosistema, los orquestadores necesitan comprender objetivos, recursos y limitaciones de sus participantes. De esta manera, la comprensión ayuda a los orquestadores a diseñar medidas de gobernanza que abordan las motivaciones de los participantes del ecosistema y no solo sus comportamientos.
Regla 2.
Reforzar los integradores. En los ecosistemas de datos, los orquestadores y habilitadores por lo general desempeñan un papel de integradores, los cuales son los que reúnen a otros e impulsan los procesos. Si se comparten los mismos formatos de datos, usando los mismos protocolos, y siguiendo la misma arquitectura de referencia, se puede reducir las barreras para el intercambio de datos, puede facilitar el intercambio y posteriormente, se puede reducir la fricción.
Regla 3.
Incrementar la cantidad total de potencia. Empoderar a los usuarios para tomar decisiones sin quitarle el poder a los demás es una gran manera de asegurarse de que todos los participantes se sientan que tienen un gran posición en el éxito del ecosistema. Los orquestadores tienen varias maneras de hacer esto, como lo son los controles de acceso, uso compartido de las reglas y la gestión de derechos, con esto se pueden crear confianza en que los datos de los contribuyentes se usan de la manera correcta, incluso cuando se muda de las manos del orquestador. 1. Proporcionar transparencia en las fuentes y usos de datos, y dar a los contribuyentes un papel a la hora de decidir de cómo se utilizarán sus datos.

2. Ser explícito sobre el posicionamiento estratégico del ecosistema en el mercado e implementar acuerdos de no competencia entre los participantes, esto puede reducir las barreras competitivas y dar tranquilidad a los contribuyentes. 3. Respecto al intercambio de datos confidenciales, se pueden generar herramientas analíticas de preservación de la privacidad que protegen los datos subyacentes mientras otros los analizan, así se puede dar a los participantes más confianza en que sus datos están seguros. 4. Los ecosistemas de datos necesitan estándares claros de comportamiento aceptable e inaceptable. Captura de datos sin consentimiento, intercambio de datos con la competencia y la reventa no autorizada a terceros debe estar claramente fuera de los límites. Los orquestadores pueden admitir la supervisión y aplicación de estas normas.
Regla 4.
Aumentar la reciprocidad. El éxito de cada usuario o participante en el ecosistema depende del éxito de los demás. Por lo tanto, los orquestadores deben definir claramente el propósito común de los participantes y su interés mutuo en compartir datos. Cuanto más claro sea el propósito, más fácilmente los contribuyentes individuales del ecosistema podrán moverse en la dirección deseada y evitar malas prácticas.
Regla 5.
Expande la sombra del futuro. Un buen modelo de creación y distribución de valor muestra a las personas u organizaciones cómo se promueve el éxito propio contribuyendo al éxito de otros. Asi mismo se ve reflejado con la distribución justa del valor resultante, ya sea económico o de otro tipo.
Regla 6.
Recompensa a los que cooperan. Hay al menos dos formas en que los orquestadores pueden crear un sistema de reconocimiento para contribuyentes de datos. Una es asignar un valor financiero a los datos compartidos, la otra, es vincular alguna forma de remuneración no financiera.

Fuente: Propia BCG

2.1.9 Estados Unidos

En el Plan de Acción 2020 de la Estrategia Federal de Datos (FDS)[68] se comprometieron a completar la Acción 13, la cual consiste en Desarrollar un catálogo de habilidades de datos seleccionados (Catálogo)[69] y se puede utilizar para ayudar a las agencias a desarrollar competencias para administrar datos como un activo estratégico y tomar decisiones basadas en datos; y la Acción 14, consiste en Desarrollar un marco de ética de datos (Marco)[70], que puede ser utilizado por líderes federales y usuarios de datos a medida que toman decisiones éticas al adquirir, administrar y usar datos para respaldar la misión de su agencia.

68 Disponible en https://strategy.data.gov/action-plan/

69 Disponible en https://resources.data.gov/assets/documents/fds-data-skills-catalog.pdf

70 Disponible en https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf

El marco de ética de datos orienta las actividades de los organismos en materia de datos, proporcionando la base para la adquisición, gestión y uso éticos de los datos. Este marco aporta los beneficios presentados en la Ilustración 20.

Ilustración 20. Beneficios de la Ética de los Datos

Fuente: DANE, Basado en el Marco de ética de datos

Los principios éticos de los datos federales ayudan a los usuarios de los datos federales a tomar decisiones de forma ética y promover la responsabilidad a lo largo del ciclo de vida de los datos, dado que estos se adquieren, procesan, difunden, utilizan, almacenan y eliminan. El mismo conjunto de datos puede ser utilizado en diferentes momentos para diferentes propósitos. En el Documento Marco[71] se encuentran todas las recomendaciones para los responsables de las organizaciones federales y para los usuarios de datos federales. En la Tabla 13 se presentan los siete principios establecidos en el marco de ética de los datos.

Tabla 13. Principios del marco de ética de los datos

Principio	Descripción
Respetar las leyes, los reglamentos, las prácticas profesionales y las normas éticas aplicables	Los responsables de los datos y los usuarios de los mismos deben adherirse a todas las autoridades legales aplicables, debido a que existen leyes que reflejan y refuerzan la ética.
Respetar al público, a las personas y a las comunidades	Las actividades relacionadas con los datos tienen el objetivo general de beneficiar al bien público. El uso responsable de los datos comienza con una cuidadosa consideración de los impactos potenciales y diferenciales.
Respetar la privacidad y la confidencialidad	La privacidad y la Confidencialidad deben protegerse siempre respetando la dignidad, Los derechos y la libertad de los interesados, el objetivo esencial es minimizar las posibles consecuencias negativas mediante medidas, como la evaluación exhaustiva de los riesgos y evitando la divulgación.

71 Disponible en https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf

Principio	Descripción
	Las actividades de datos que implican la privacidad de las personas deben ajustarse a los principios de prácticas aceptables de información (PPAI): • Acceso y modificación • Rendición de cuentas • Autoridad • Minimización • Calidad e integridad • Participación individual • Especificación de la finalidad y limitación del uso • Seguridad • Transparencia
Actuar con honestidad, integridad y humildad	Los líderes federales y los usuarios de datos, deben mostrar honestidad e integridad en su trabajo con los datos independientemente del cargo, la responsabilidad es de los datos y el papel de la organización; no deben realizar o aprobar comportamientos poco éticos con los datos.
Responsabilizarse a sí mismo y a los demás	La rendición de cuentas requiere que cualquier persona que adquiera, gestione o utilice datos sea consciente de las partes interesadas y se responsabilice ante ellas. La responsabilidad incluye el manejo responsable de la información clasificada y controlada, El cumplimiento de los acuerdos de uso de los datos con los proveedores de datos como la minimización de la recopilación de datos, la información a las personas y organizaciones sobre los posibles usos de los datos.
Promover la transparencia	Los individuos, las organizaciones y las comunidades se benefician cuando el proceso de toma decisiones éticas es lo más transparente posible para las partes interesadas. La transparencia depende de una comunicación clara de todos los aspectos de las actividades de datos y de un compromiso adecuado con las partes interesadas de los datos.
Mantenerse informado de los avances en los campos de la gestión de datos y la ciencia de los datos	Los dirigentes y federales y los usuarios de los datos deben mantenerse al corriente de las innovaciones y de cómo utilizar esos métodos de manera ética, estos avances pueden estar dados por tecnologías avanzadas, análisis y métodos computacionales que deben ser supervisados y evaluados. • Inteligencia artificial • Aprendizaje automático • Redes neuronales • Automatización de procesos robóticos • Internet de las cosas • Blockchain

Fuente: Tomado del Marco de ética de datos

2.1.10 Nueva Zelanda

El Director del Instituto Nacional de Estadística de Nueva Zelanda (Stats NZ) también tiene el rol de jefe administrador de datos del gobierno[72]. Este rol facilita y permite un enfoque de uso de los datos para todo el gobierno. Además de desarrollar las políticas e infraestructura, proporciona apoyo y orientación para que las agencias puedan usar los datos de manera efectiva, al tiempo que mantiene la confianza de los neozelandeses. Tiene 4 tareas principales relacionadas con:

- Establecer la dirección estratégica para la gestión de datos del gobierno.
- Liderar la respuesta del sector estatal a los problemas de datos nuevos y emergentes.

72 Disponible en https://www.stats.govt.nz/about-us/data-leadership/

- Co-desarrollar un Marco de administración de datos para permitir a las agencias administrar datos como un activo estratégico y comparar la madurez de los datos.
- Liderar el compromiso del gobierno para acelerar la liberación de datos abiertos.

El rol de jefe administrador de datos del gobierno funciona en coordinación[73] con el Director Digital del Gobierno, el Director de Seguridad de la Información del Gobierno y el Director de Privacidad del Gobierno. Su trabajo está soportado en la estrategia y hoja de ruta de datos[74] del gobierno, que tienen cuatro áreas de enfoque asociadas a los datos, la capacidad, la infraestructura y el liderazgo. La Estrategia y la Hoja de Ruta están respaldadas por el compromiso de mantener y mejorar la confianza pública.

Dentro de la estrategia y hoja de ruta de datos del gobierno es de vital importancia la participación de los grupos consultivos y de gobernanza[75]. En la **Tabla 14** se indican sus particularidades:

Tabla 14. Grupos consultivos y de gobernanza de la estrategia de datos del gobierno

Grupos Consultivos y de gobernanza	Características
Grupo de información	Es un grupo interinstitucional presidido por líderes en el Servicio Público. El grupo tiene como objetivo aumentar la eficacia del Servicio Público mediante el fortalecimiento del liderazgo del sistema de datos del gobierno.
Grupo asesor de ética de datos[76]	Es convocado por el Jefe de Administración de Datos del Gobierno y el Director Ejecutivo de Stats NZ. El Grupo permite a las agencias gubernamentales probar ideas, políticas y propuestas relacionadas con los usos nuevos y emergentes de los datos. También proporcionará asesoramiento sobre tendencias, problemas, áreas de preocupación y áreas de innovación de las que tenga conocimiento.
Comité de aplicación y examen de normas encomendadas	Es un grupo interinstitucional formado por miembros representativos del sector público. Este comité aconseja sobre si una norma debe ser obligatoria para su uso por los Departamentos de Servicio Público y las Agencias Departamentales.

Fuente: DANE, basado en información del gobierno de Nueva Zelanda

En este marco también se han desarrollado principios clave para apoyar el análisis de datos seguros y efectivos, incluida la toma de decisiones algorítmicas, y se están elaborando lineamientos

73 Disponible en https://www.digital.govt.nz/digital-government/leadership/government-functional-leads/

74 Disponible en https://data.govt.nz/docs/data-strategy-and-roadmap-for-new-zealand-2021/

75 Disponible en https://data.govt.nz/leadership/advisory-governance/

76 Disponible en https://data.govt.nz/leadership/advisory-governance/data-ethics-advisory-group/

para ayudar a los organismos gubernamentales a adoptar un enfoque de mejores prácticas para las actividades de análisis de datos relacionadas con la privacidad, la seguridad y la confidencialidad[77].

Enfoque del Gobierno de datos en Nueva Zelanda

La administración efectiva de datos depende de un buen gobierno de los datos, en este sentido, Stats NZ se asoció con las entidades del gobierno para diseñar un nuevo Marco Operativo de Gobierno de Datos que está diseñado para respaldar la gestión exitosa de activos de datos, abordar las brechas de gobernanza actuales y promover la gestión del ciclo de vida de los datos y los modelos de procesos comerciales que se apoyan mutuamente.

El Marco de Gobierno de Datos Operativos[78] respalda las mejores prácticas consistentes en áreas como la gestión de datos y ayuda a las organizaciones a darse cuenta del valor de sus datos al:

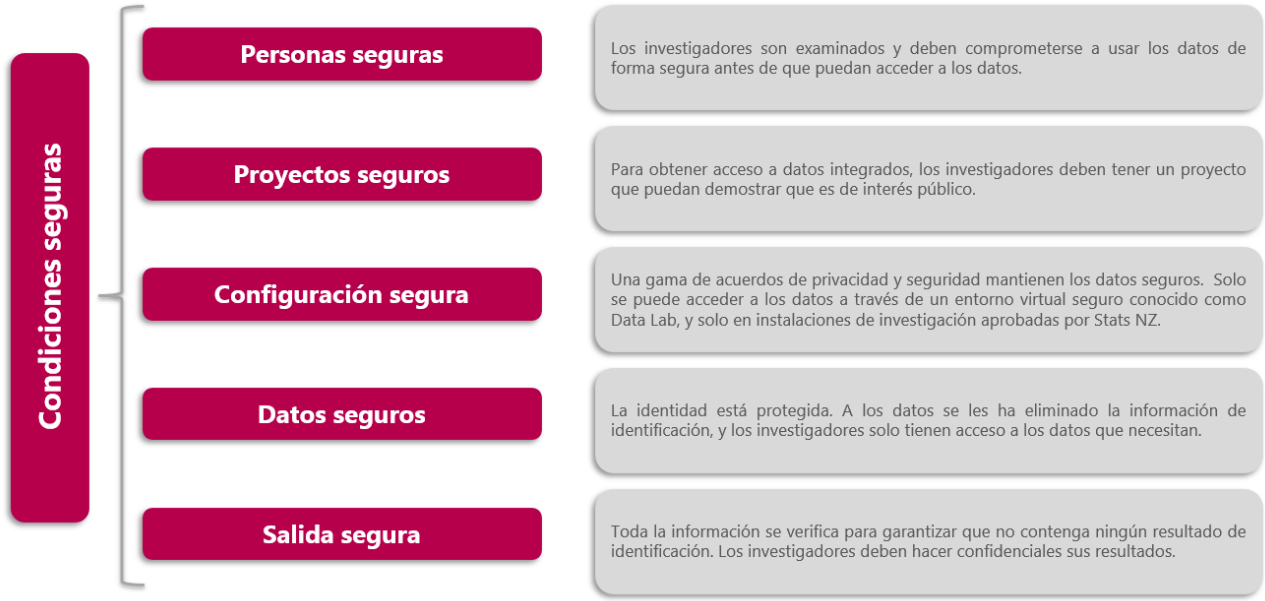
- Supervisar y mejorar la calidad de los datos
- Promover la rendición de cuentas y la transparencia
- Garantizar que las buenas prácticas de datos se incorporen en un enfoque de gestión de datos, desde el principio y "por diseño"
- Apoyar modelos de negocio ágiles, y
- Facilitar el monitoreo interno de los procesos de negocio.

Por otro lado, Stats NZ menciona dentro de sus actividades de interoperabilidad e intercambio de datos la infraestructura integrada de datos (IDI, por sus siglas en inglés). Esta infraestructura soporta una gran base de datos de investigación que contiene información sobre personas y hogares que provienen de las agencias gubernamentales, no gubernamentales y de las encuestas de Stats NZ. La IDI complementa la base de datos longitudinales de negocios (LBD, por sus siglas en inglés) que se vinculan a través de los datos fiscales. Los datos integrados se mantienen seguros a través de 5 condiciones seguras: personas seguras, proyectos seguros, configuraciones seguras, datos seguros y salidas seguras.

77 Disponible en https://data.govt.nz/toolkit/privacy-and-security/

78 Disponible en https://www.data.govt.nz/toolkit/data-governance/odgf/

Ilustración 21. Cinco condiciones seguras para los datos integrados



Fuente: DANE a partir de https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/ Stats NZ

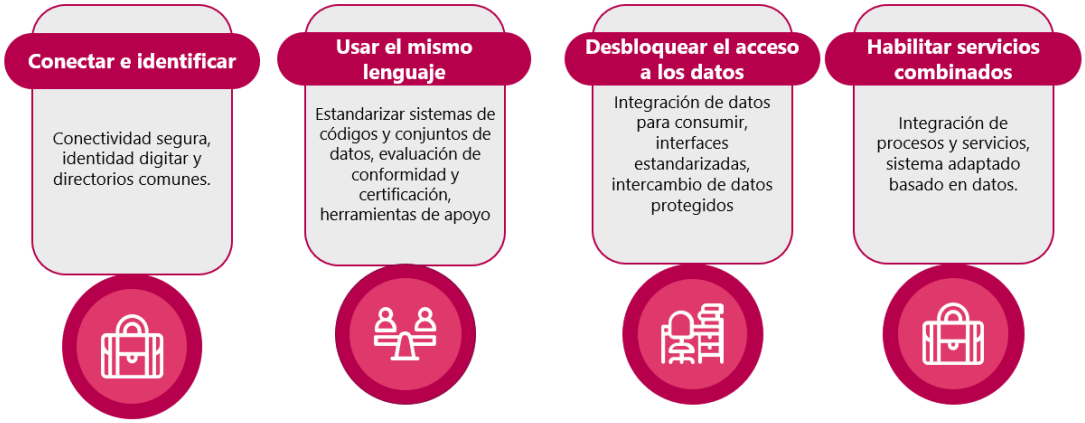
Ejercicios de interoperabilidad e intercambio de datos en el marco del SEN

Un caso para destacar en el sistema estadístico de Nueva Zelanda es el proyecto de interoperabilidad[79] del Ministerio de Salud. Este proyecto tiene como objetivo garantizar que los médicos, los pacientes y sus cuidadores puedan acceder a los datos y la información "en cualquier lugar, en cualquier momento y de cualquier manera", al tiempo que se garantiza que se cumplan los requisitos pertinentes de normas, privacidad y seguridad. La hoja de ruta de este proyecto tiene cuatro temas entrelazados que se muestran en la Ilustración 22.

Para generar estos procesos de interoperabilidad requiere de la participación de otros organismos de estandarización y normalización, entre ellos el Stats NZ, que produce especificaciones de requisitos de contenido de datos para administrar la información de los neozelandeses y el Consejo de iniciativa conjunta para la informática de la salud global.

79 Disponible en https://www.health.govt.nz/our-work/digital-health/digital-health-sector-architecture-standards-and-governance/health-interoperability

Ilustración 22. Hoja de ruta del proyecto de interoperabilidad del Ministerio de Salud



Fuente DANE a partir de información del Stats NZ.

2.1.11 Chile

En marzo del año 2022, la Comisión Asesora de Datos de interés Públicos del Ministerio de Ciencia realiza una entrega de un informe con alternativas para una gobernanza eficiente y ética de los datos en Chile[80]. Paralelamente, desde la academia declaraban que "La integración de datos de manera segura es indispensable para tener un Estado al servicio de las personas y para saber si las políticas públicas funcionan"[81]. Básicamente, este informe destaca la importancia de definir un modelo de administración de datos, mientras propone posibles modelos de implementación, uno descentralizado y otro centralizado. El informe hace mención y resalta un conjunto de Buenas prácticas para una adecuada gobernanza de datos y posteriormente, realiza una descripción de los usuarios como un componente clave de gobernanza.

> *Una buena gobernanza de datos contribuye a mantener una visión común; a mejorar su implementación coherente a todo nivel, y a fortalecer las bases institucionales para estas tareas. (Comisión 2022)*

Desde una perspectiva pública, la comisión destaca que la gobernanza de datos acerca de las personas o territorios, en poder de ministerios u otras agencias de los estados, ha cobrado importancia en los últimos años y es un factor importante para diseñar políticas públicas en los ámbitos socioeconómicos, de salud, y ambientales. La comisión enfatiza que el Ministerio de Ciencia puede complementar la trayectoria de modernización del sector público, con una visión del Estado como productor y responsable de disponibilizar datos para informar, tomar decisiones y crear conocimiento. En el mismo reporte, la comisión nombra que para que los datos contribuyan a informar, crear conocimiento y a tomar decisiones, una adecuada gobernanza debería promover la eficiencia y eficacia del Estado en al menos cinco aspectos fundamentales:

80 Disponible en https://datos.minciencia.gob.cl/documents/Datos_Interes_Publico.pdf

81 https://minciencia.gob.cl/noticias/comision-asesora-de-datos-de-interes-publicos-minciencia-entrega-informe-con-alternativas-para-una-gobernanza-eficiente-y-etica-de-los-datos-en-chile/

1) Los sistemas de observación estatales deben producir datos de observación del entorno y fenómenos sociales, asimismo, los registros producidos en la digitalización del Estado idealmente deben ser de buena calidad.
2) Debe existir mecanismos para que los usuarios de los datos puedan orientar y retroalimentar las prioridades sobre lo que estos sistemas observen, o para determinar cuáles de los registros administrativos deben convertirse en datos.
3) Debe existir acceso organizado y oportuno a los datos.
4) Los análisis que se elaboren con base en dichos datos deben tener sentido y contexto.
5) Los datos como sus análisis deben poder ser usados y reusados en el tiempo, para los ejercicios de chequeo tanto del conocimiento que se produce como su reproducibilidad.

Nombran una serie de ejemplos de uso de datos para la información, toma de decisiones y creación de conocimiento[82].

Observa, Observatorio Nacional del Sistema de CTCI
Es una plataforma que de manera clara y abierta, y en un solo lugar, permite visualizar y descargar datos e información sobre las capacidades y producción de Chile en Ciencia, Tecnología, Conocimiento e Innovación, mostrando aspectos tales como el desempeño de la innovación en empresas, la inversión en I+D en universidades, centros de investigación, privados y el Estado e investigadores, y las brechas de género en estas materias, entre otras temáticas que hacen posible comparar a nuestro país con las naciones de la OCDE. La información proviene de bases de datos construidas por el Minciencias y con apoyo de instituciones como el INE, a partir de estándares internacionales del manual de Frascati y Manual Oslo de la OCDE. Entre estos instrumentos se encuentran la Encuesta de I+D, la Encuesta Nacional de Innovación, la Encuesta de Trayectoria de Personas con Doctorado, la Encuesta de Presupuesto Público Destinado a I+D (GBARD), y el Registro de Empresas de Base Científico-Tecnológica - EBCT. Asi mismo, se presentan datos estandarizados provenientes de las agencias que ejecutan el presupuesto público en materias de CTCI, tales como la Agencia Nacional de Investigación y Desarrollo - ANID, la Fundación para la Innovación Agraria - FIA y Corfo. Finalmente, la plataforma también ofrece registros administrativos de producción tecnológica y capital humano provenientes de INAPI, y del Servicio de Información de Educación Superior - SIES, del Ministerio de Educación.
Centro Nacional en Sistemas de Información en Salud
El Centro Nacional en Sistemas de Información en Salud, CENS, desarrolla estrategias y actividades que permiten madurar la interoperabilidad de datos en salud, cerrando brechas en conocimiento y competencias en estándares, terminologías y buenas prácticas internacionales, apoyando criterios de gobernanza para asegurar la calidad de datos y sistemas de información hospitalaria. CENS ofrece soluciones y acompaña instituciones para enfrentar los desafíos en la implementación óptima de las tecnologías de información en salud en las áreas de interoperabilidad, calidad de datos, capital humano e innovación. Productos en el ámbito de bienes públicos y servicios incluyen (i) Guías de Buenas Prácticas y Recomendaciones para el uso de Telemedicina y Telesalud; (ii) Marco de Competencias Nacionales y Perfiles Laborales; (iii) Historia Clínica Compartida, con el objetivo de generar un registro único de encuentros clínicos para configurar la historia sanitaria de pacientes a nivel nacional; (iv) Registro Nacional de Cáncer, con el objetivo de tributar datos de las atenciones clínicas a un gran repositorio que agrupe la información de los pacientes; o (v) el Bien Público

82 Se destacan dos de los casos que se nombran en el Reporte generado por la comisión.

Regional Transformación Digital en Salud para Mitigar los Efectos de COVID-19 en América Latina y en el Caribe, financiado por el Banco Interamericano de Desarrollo (BID), trabajando en el Smart Vaccine Certificate según lineamientos de la Organización Mundial de la Salud (OMS) y Organización Panamericana de la Salud (OPS), mejorar intercambio de datos para la vigilancia epidemiológica y la salud pública a nivel nacional y regional y crear lineamientos y directrices para el desarrollo sostenible de la Telesalud.

Fuente DANE a partir Comisión Asesora de Datos de interés Públicos del Ministerio de Ciencia

Modelo "The Path to Becoming a Data-Driven Public Sector"

La comisión hace énfasis en el modelo propuesto por la OCDE en The Path to Becoming a Data-Driven Public Sector, el cual muestra todos los aspectos organizativos, políticos y técnicos para una gobernanza de datos exitosa e identifica una variedad de elementos y herramientas de gobierno de datos que organiza en seis grupos diferentes, agrupados en tres niveles.

Ilustración 23. Aspectos para una gobernanza de datos exitosa

Fuente: OCDE "The path to Becoming a Data-Driven Public Sector"

La comisión destaca que las recomendaciones OCDE representan una buena síntesis de lo que se ha promovido en la trayectoria de Modernización del Estado en Chile y deben ser abordadas también por una gobernanza que apunte a fomentar el ejercicio de informar, tomar decisiones y/o crear conocimiento con datos de interés público. A continuación, se presenta un conjunto de Buenas prácticas para una adecuada gobernanza de datos, ya que la comisión junto con los lineamientos de la OCDE declara que una buena gobernanza de datos debe abordar elementos tales como:

Nivel estratégico: se definen liderazgos, expectativas, las funciones y los objetivos que guían la formulación de políticas y/o estrategias de datos; y que se beneficia de procesos abiertos y participativos, integrando así los aportes de los actores dentro y fuera del sector público para una mayor apropiación social de las estrategias.

- Debe censar, coordinar, integrar y en lo posible habilitar el acceso a la ciudadanía a los sistemas de observación que usa o a los que tiene acceso directo o indirecto el Estado.
- Dar acceso, en lo posible, a las observaciones, bases de datos y al conocimiento que estos sistemas han promovido, contextualizando adecuadamente.
- Debe considerar también que, en ciertas situaciones, el Estado puede recolectar los datos producidos de manera transversal en el país, provenientes tanto de fuentes privadas como públicas.
- Debe considerar la interoperabilidad de los datos entre sistemas de observación. Un mismo tipo de dato puede ser integrado en diferentes contextos para generar conocimiento en ámbitos diversos. Si el dato es adquirido o generado por una entidad en particular, puede ser usado por varias, en diferentes sistemas.
- Debe considerar el rango dinámico y multidimensionalidad de los datos, distintas escalas de espacio y tiempo.
- Debe considerar los límites de usos de datos personales, reconociendo la trayectoria y estándares de proyectos y normativas existentes al respecto.
- Debe considerar las plataformas y herramientas de exploración y explotación de los datos que disponibiliza y a qué nivel se involucra con ellas; desde modelos en los que estas no son parte de su flujo, a unos dónde estas se coordinan para sinergias o incluso se desarrollen propias. Con el fin de definir su rol en el uso-reúso de los datos y el valor público que generan.
- Debe considerarse la apertura permanente a incluir nuevos sistemas de observación y sus datos, priorizados con base en las demandas por conocimiento que el país requiera y las comunidades de usuarios releven.

Nivel táctico: Se orientan las capacidades de implementación, marcos legales y regulatorios, velando por la coherencia de políticas públicas, estrategias y/o iniciativas basadas en datos. Se beneficia de las capacidades y competencias del sector público, una correcta definición de perfiles de trabajo, la buena comunicación, coordinación y colaboración y, también, de reglamentos, estrategias y políticas que orienten.

- Los reglamentos, políticas y herramientas para ofrecer acceso a la ciudadanía a los sistemas de observación que usa o a los que tiene acceso directo o indirecto el Estado, pueden afrontar sus preguntas de investigación con estos sistemas.
- Los reglamentos, políticas y herramientas para captura y acceso a datos que los sistemas de observación producen y disponibilizar data, incluyendo protocolos para anonimizar o pseudo-anonimizar y responder con éxito a la necesidad de compartir datos y al respeto a los derechos de sus titulares.
- Reglamentos, políticas y herramientas para mejorar la calidad de datos con validaciones, limpieza o enriquecimiento de datos.
- Reglamentos, políticas y herramientas para establecer estándares de datos y plataformas para acceder a ellos.

Nivel de despliegue: Aborda la integración de lo definido en los 2 niveles anteriores al ciclo de los datos, la infraestructura y la arquitectura de datos. Este nivel abarca diferentes aspectos técnicos y políticos del ciclo de valor de los datos (desde la producción y apertura de datos hasta la reutilización), el rol y la interacción de diferentes actores en cada etapa y la interconexión de los datos fluyen a través de las etapas. La gobernanza de datos públicos debe abordar la pluralidad de contextos y objetivos que se presenten.

- Debe identificar y definir grupos de actores interesados en participar en los procesos de información, toma de decisiones y creación de conocimiento para establecer mecanismos de interacción, retroalimentación y cooperación permanente entre esas comunidades y los responsables de la Gobernanza.
- Debe definir los mecanismos de consumo, retroalimentación y de aseguramiento de calidad de los datos, definiendo adecuadamente roles y responsabilidades para cada proceso definido.

Usuarios y comunidades como un componente de la gobernanza

En el reporte se destaca que los usuarios y comunidades que se relacionan con los datos terminan siendo un componente clave en la gobernanza. A continuación, se incluye una descripción de los usuarios según su rol en el ciclo de los datos. Se hace énfasis en que una adecuada gobernanza de datos para la comunicación, toma de decisiones y creación de conocimiento debe considerar los siguientes perfiles de usuarios y sus intereses primarios:

- **Capturadores:** los datos son generados vía detección o creación por sistemas o instituciones, y existen equipos, estructuras y roles dedicadas a que esto ocurra.
- **Creadores:** algunos datos son producidos a través de procesos de creación, producción o investigación. El principal interés de este tipo de usuario es contar con mecanismos de verificación de que los datos creados tengan relación con la realidad, esto incluye su divulgación en algunos casos.
- **Recopiladores / almacenadores:** son entidades, equipos y roles que cumplen el rol de agrupar y almacenar datos que provienen de diversas fuentes, y de asegurar que ese dato sea fiel al dato originalmente producido, creado y/o capturado.
- **Curadores:** son los equipos, estructuras y roles encargados de depurar los datos, de limpiarlos, identificar errores o anomalías según marcos de comportamiento de aquellos datos, y aplicar modelos matemáticos que permitan corregir y asegurar su exactitud y que sean fehacientes.
- **Transformadores:** son estructuras o roles que trabajan grandes cantidades de datos y los transforman en entidades matemáticas nuevas, aplicando modelos probabilísticos, generando nueva información a partir de los datos base. Este tipo de usuario busca que los datos que utiliza sean de buena calidad y en tener acceso oportuno.
- **Analistas:** toman los datos y son capaces de producir nueva información, como generar indicadores o visualizaciones sobre aspectos específicos que pueden ser comprendidos a partir de los datos. Este grupo tiene como principal interés en que los datos que utiliza sean de buena calidad y en tener acceso oportuno.

- **Explotadores:** son equipos y roles que usan de forma exploratoria o innovadora los datos, para generar algún valor.
- **Contralores:** son equipos o roles a cargo de la trazabilidad. Velan porque las transformaciones que pueda experimentar un dato a lo largo de su ciclo de vida sea una transformación coherente con el dato original, así como también que el dato respete los marcos éticos y legales en todo su ciclo, incluido el respeto a sus términos de uso. El principal interés de este grupo de usuarios es tener acceso a auditar datos y sus usos.
- **Comunicadores:** equipos y roles se centran principalmente en la difusión y en el lenguaje adecuado a distintas audiencias, en función de los datos y de lo que se produce con ellos, considerando a las audiencias y sus requisitos, como los medios de comunicaciones que se utilizan.

2.1.12 KPMG

En los últimos años, la tasa de transformación digital se ha acelerado como resultado de la pandemia producida por el COVID-19, se ha amplificado la necesidad del acceso en línea a productos y servicios, lo que a la vez ha vuelto a los clientes más conscientes de la recolección, conservación y uso de sus datos por parte de las organizaciones, ante esto, las organizaciones deben demostrar que actúan en pro de los intereses de sus clientes y de manera ética, no obstante, las leyes actuales de protección de datos y del consumidor no tienen en cuenta el impacto de la innovación en análisis avanzado e inteligencia artificial (AAAI) sobre las personas y la sociedad, pues las regulaciones actuales son independientes de la tecnología. Así mismo, los avances de los sistemas AAAI toman decisiones que afectan en gran medida a los clientes con poca o ninguna supervisión humana, lo que en ocasiones deja resultados negativos, pues puede que estos sistemas reflejen el sesgo humano o exacerben el histórico[83].

La presión para aumentar la inversión en la automatización de procesos de los clientes, impulsada por el deseo de reducir costos operativos, mejorar su crecimiento (las empresas que invierten en análisis de datos han crecido tres veces más rápido que aquellas que están menos orientadas al análisis) y el retorno sobre el capital, ha generado una mayor comprensión de las responsabilidades asociadas con la propiedad de datos y el uso de tecnologías autónomas por parte de las instituciones de servicios financieros, pues si estas instituciones pierden su condición de custodios de confianza de datos de los clientes, es probable que pierdan su licencia para operar; el uso ético de los datos de los clientes ha sido un enfoque nuevo para el área financiera, un ejemplo de este son los “Principios de reciprocidad[84]” que se desarrollaron como base para compartir datos de clientes con proveedores externos con el fin de mejorar las verificaciones de crédito.

Aunque las regulaciones de protección de datos son un buen comienzo, un enfoque más global va más allá del cumplimiento directo, pues como se mencionó anteriormente es probable que las regulaciones no sigan el rápido ritmo del desarrollo tecnológico, es necesario un enfoque sólido que abarque la generación, almacenamiento y uso de todos los datos de los clientes de manera segura garantizando resultados justos y confiables. Dado lo anterior, es fundamental establecer un marco

83 Disponible en https://home.kpmg/uk/en/blogs/home/posts/2020/12/the-rise-of-aaai-how-a-data-ethics-framework-drives-value.html

84 Disponible en https://ico.org.uk/media/action-weve-taken/audits-and-advisory-visits/1042574/outcomes-report-credit-reference-agencies.pdf

ético que cuente principios formativos sobre el uso de datos y permita la integración de este en toda la organización, desde los aspectos legales y financieros, hasta el desarrollo comercial, satisfaciendo las necesidades de todas las partes interesadas[85].

En busca de ayudar a las empresas de servicios financieros a gestionar esta labor, KPMG (organización global de firmas profesionales que brinda servicios de auditoría, impuestos y asesoría[86]) ofrece sus servicios para ayudar a definir e integrar un marco de ética de datos que cubra todo el ciclo de vida de AAAI en tres fases: i) realizar evaluaciones sobre el cumplimiento actual de normas y políticas internas y externas para determinar el nivel de exposición al riesgo de AAAI que se está usando, ii) ayudar a respaldar el diseño y desarrollo de un marco ético que cuente con principios, gobernanza y un modelo operativo que respalde el diseño, desarrollo y pruebas de AAAI éticos y explicables, iii) una vez adoptado la ética de datos, es importante mantener el cumplimiento ético que permita, "desde tener un proceso establecido para probar y monitorear AAAI a lo largo de su ciclo de vida, hasta la implementación de procesos de escalamiento apropiados y administración de excepciones[87]".

Como apoyo adicional, KPMG desarrolló los documentos: "Uso ético de los datos de clientes en una economía digital[88]" el cual discute los desafíos éticos clave que enfrentan las instituciones financieras en la actualidad, y en coautoría con UK Finance elaboró el documento "los principios éticos para AAAI en servicios financieros[89]" cuyo objetivo es ayudar a mantener la confianza del público mitigando los riesgos potenciales de estas tecnologías. Ambos documentos proponen cinco principios éticos, la Tabla 15 los ilustra, estos se pueden aplicar como punto de referencia para fomentar o mejorar los principios internos y de gobierno propios de la organización (pues están diseñados para ser lo suficientemente flexibles para adaptarse según corresponda), permitiendo desarrollar productos, servicios y aplicaciones administrativas que se basen en AAAI con un manejo ético de los datos de los clientes.

85 Disponible en https://assets.kpmg/content/dam/kpmg/uk/pdf/2019/04/ethical-use-of-customer-data.pdf

86 Disponible en https://home.kpmg/xx/en/home/about/who-we-are/governance.html

87 Disponible en https://home.kpmg/uk/en/home/insights/2020/11/data-ethics-how-transparency-translates-into-trust.html

88 Disponible en https://assets.kpmg/content/dam/kpmg/uk/pdf/2019/04/ethical-use-of-customer-data.pdf

89 Disponible en https://assets.kpmg/content/dam/kpmg/uk/pdf/2020/12/ethical-principles-for-advanced-analytics-and-artificial-intelligence-in-financial-services-december-2020.pdf

Tabla 15. Principios éticos para la analítica avanzada y el uso ético de los datos

PRINCIPIOS ÉTICOS PARA ANALÍTICA AVANZADA E INTELIGENCIA ARTIFICIAL EN SERVICIOS FINANCIEROS			PRINCIPIOS PARA EL USO ÉTICO DE LOS DATOS DE CLIENTES EN UNA ECONOMÍA DIGITAL		
Principio	**Descripción**	**Medidas**	**Principio**	**Descripción**	**Medidas**
1: Explicabilidad y Transparencia	Sea transparente sobre cómo usamos AAAI y brinde explicaciones apropiadas sobre las decisiones	ü Entender la AAAI y los resultados que producen ü Desarrollar y usar la AAAI con entradas y salidas que son auditables en las etapas apropiadas a lo largo de su ciclo de vida ü Brindar información clara y explicaciones apropiadas para la audiencia, el contexto y el propósito	**1: Respetar la agencia humana**	Las instituciones financieras deben respetar la capacidad de los seres humanos para tomar sus propias decisiones libres	• Las instituciones financieras no deben engañar ni manipular a los clientes para que actúen en contra de sus propios intereses, ni restringir indebidamente el acceso de los clientes a la información. • A menos que exista un interés público superior para no hacerlo, los clientes deben poder saber o verificar cuándo está usando inteligencia artificial o decisiones automatizadas, y debe haber un nivel apropiado de control humano sobre estos sistemas, incluida una vía adecuada para que los clientes cuestionen decisiones automatizadas importantes.
2: Integridad de AAAI	Adoptar controles apropiados para la integridad, el abastecimiento y el intercambio de AAAI y sus datos asociados a lo largo del ciclo de vida de AAAI	• Emplear prácticas de protección de datos, seguridad y resiliencia apropiadas para el caso de uso de AAAI y su nivel de riesgo • Revisar y adaptar el gobierno para que satisfaga las necesidades cambiantes y tenga líneas claras de responsabilidad a lo largo del ciclo de vida de AAAI • Mantener una gestión de riesgos sólida y adecuada durante todo el ciclo de vida de AAAI	**2: Salvaguardar la igualdad y la equidad**	Las instituciones financieras deben tratar a sus clientes de manera justa y respetar sus derechos básicos	• Identificar los impactos negativos potenciales del procesamiento en los clientes y compararlos cuidadosamente con los beneficios previstos para garantizar la proporcionalidad (como la pérdida financiera, impactos sobre la privacidad, etc.) • Identificar y evaluar los riesgos de sesgo injusto y discriminación que pueden ocurrir a través de los datos mismos o del sesgo humano dentro de la fuerza laboral que programa el algoritmo de IA (incluye la eliminación de cualquier característica personal cuyo uso no pueda justificarse objetivamente).

3: Equidad y Alineación con los Derechos Humanos	Diseñar y usar la AAAI para que produzca resultados justos	• Definir la equidad dentro del contexto y el propósito de la solución AAAI, y considerar el impacto en la sociedad en general • El objetivo es comprender y mitigar el sesgo injusto en el desarrollo y uso de AAAI • Respetar los derechos humanos en el desarrollo y uso de AAAI	**3: Ofrecer transparencia**	Las instituciones financieras deben procesar los resultados de los datos dentro de los límites de una "caja de cristal"	• La transparencia apoyará la inteligibilidad, explicabilidad y verificabilidad de los datos y cualquier acción tomada sobre la base de los datos. Las empresas deben poder comprender las decisiones que están tomando y poder explicar estas decisiones a los clientes, auditores y reguladores de manera adecuada. • Cuando sea difícil explicar exactamente cómo se ha llegado a una decisión, las empresas deben tratar de minimizar el riesgo de resultados injustos o inesperados. Las pruebas exhaustivas y rigurosas, asegurándose de que el algoritmo funcione como se supone que debe hacerlo, pueden formar la base de la confianza. • La transparencia con el cliente ayudará a impulsar la alfabetización de datos y ayudará en la creación de una cultura de confianza.
4: Contestabilidad y empoderamiento humano	Apoyar el empoderamiento de los sujetos de la AAAI, respetando su toma de decisiones	• Adoptar medidas que permitan a los sujetos de la AAAI impugnar decisiones tomadas por los sistemas de la AAAI • Contar con procesos que permitan una revisión, supervisión y control humano efectivo de la toma de decisiones de la AAAI • No restringir injustamente la toma de decisiones de los sujetos AAAI • Brindar opciones adecuadas a los sujetos AAAI,	**4: Enfoque de toda la organización del patrocinador**	Las instituciones financieras deben impulsar la ética de los datos desde arriba y asegurarse de que se adopte en todas las funciones comerciales al incorporarla en sus marcos de gobierno existentes bajo el régimen de conducta de los servicios financieros	• Establecer un gobierno empresarial adecuado impulsado por marcos para proporcionar la transparencia necesaria, las comunicaciones, la cultura y la propiedad empresarial y de datos definida. • Los líderes senior deben ser evangélicos en su apoyo al uso ético de datos, pero también deben garantizar comportamientos sólidos en el día a día.

		cuando corresponda			
5: Responsabilidad y Rendición de Cuentas	Ser responsable y rendir cuentas sobre la AAAI	• Probar y monitorear el AAAI para garantizar que se cumple con los estándares regulatorios y éticos • Mantener políticas que describan el alcance de la responsabilidad a lo largo del ciclo de vida de AAAI • Realizar la diligencia debida antes de comprometerse con socios y proveedores • Identificar y equilibrar los requisitos competitivos para AAAI • Brindar recursos y capacitación, adecuados a las funciones de las personas, para que entiendan AAAI, sus beneficios y riesgos asociados. • Fomentar la participación inclusiva en el desarrollo de AAAI a lo largo de su ciclo de vida	**5: Establecer responsabilidad**	Las instituciones financieras deben establecer una cadena de mando sobre ética de datos, con principios claros de responsabilidad	• Definir la rendición de cuentas y garantizar que sea entendida y acordada por todos partes a lo largo de la cadena de suministro, ayudará a resolver cualquier ambigüedad, incluso en torno a la responsabilidad, si surgen problemas o infracciones éticas. Deben aplicarse a terceros los mismos principios, gobernanza y rendición de cuentas que se aplican internamente. • Desarrollar procesos y marcos para probar, monitorear y controlar la responsabilidad potencial en torno al uso ético de los datos. Los empleados deben comprender las limitaciones de la IA y los algoritmos.

Fuente DANE a partir de KPMG 2019 y 2020.

Finalmente, KPMG proporciona el documento “Desmitificando la cadena de bloques para las ciencias de la vida[90]” en el que propone la tecnología Blockchain como la clave para proporcionar un flujo libre y seguro de datos entre individuos, organizaciones y terceros, en información altamente confidencial (información de salud) o de propiedad intelectual.

2.4 Conclusiones

A partir de la revisión de referentes, se generan las siguientes conclusiones sobre la gobernanza y ética de los datos en torno a la interoperabilidad en el Sistema Estadístico Nacional, en torno a las preguntas planteadas:

- Para la mayoría de los países existe un órgano administrativo encargado de liderar las estrategias en torno al intercambio de datos públicos y las buenas prácticas en la gobernanza de datos. Por ejemplo, en países como España se destaca la creación de una Oficina de Datos para apoyar las políticas digitales del gobierno, mientras que, en Nueva Zelanda es el director del Instituto Nacional de Estadística el que tiene el rol de jefe administrador de datos del gobierno, y se distingue por su robusto sistema de gestión e instancias de coordinación que soportan las políticas, las estrategias y la construcción de la hoja de ruta de datos del país. Instancias como: el Grupo de Información, el Grupo Asesor de Ética de Datos y el Comité de Aplicación y Evaluación de Normas son de vital importancia para aplicación de las políticas de interoperabilidad e intercambio de datos. En el caso de Colombia, el ente regulador del Marco de Interoperabilidad es el Ministerio de Tecnologías de la Información y las Comunicaciones, en línea con lo anterior, la CEPAL recomienda que cada institución que hace parte de la gobernanza digital y la interoperabilidad gubernamental, debe identificar cuál es su papel en relación con la ciudadanía, las organizaciones y otras instituciones, permitiendo tener un panorama integrado que genere valor público.
- Desde Statistics Canada, se recomienda la implementación de un Marco de Necesidad y Proporcionalidad, donde cada propuesta para un nuevo proyecto o adquisición de datos debe explicar su importancia, cuáles son los beneficios para la ciudadanía, quién necesita la información y abordar consideraciones éticas como privacidad y transparencia; cada nuevo proyecto debe someterse a una revisión ética por parte de la Secretaría de Ética del Dato. Dicha característica es susceptible de revisión por parte del Sistema de Ética del DANE, con el objetivo de revisar las premisas de transparencia en el ámbito de la gobernanza de datos que se aplican actualmente.

- La adaptación de los códigos de buenas prácticas y la visibilización de nuevas fuentes de información destacan la necesidad de mejorar la confianza en el intercambio y la reutilización de información, por eso, se hace indispensable revisar nuevas fuentes de aprovechamiento estadístico.

90 Disponible en https://assets.kpmg/content/dam/kpmg/br/pdf/2018/05/br-demystifying-block-chain-for-life-sciences.pdf

3.

Marco conceptual de los métodos de recolección de información

3. Marco conceptual de los métodos de recolección de información

3.1 Resumen

Esta revisión de referentes internacionales se remite al creciente interés por estandarizar los conceptos asociados a los métodos de recolección de información, con el fin de lograr una clasificación efectiva que permita reconocer las características de las operaciones estadísticas desarrolladas por el DANE, desde el momento de su recolección hasta su difusión, con el objetivo de determinar si los métodos que se aplican actualmente son adecuados para las operaciones estadísticas desarrolladas o se pueden aplicar métodos adicionales.

3.2 Síntesis de hallazgos

A continuación, en la Tabla 16 se presenta una breve descripción de los principales hallazgos de la revisión de referentes internacionales sobre el marco conceptual de los métodos de recolección de información.

Tabla 16. Marco conceptual de los métodos de recolección de información

Término	Referente	Definición propuesta
Computer Assisted Interviewing (CAI) / Entrevista asistida por computadora	**Consejo de investigación económica y social de Reino Unido / Economic and social research Council**	Es la forma en que se pueden usar las computadoras en el desarrollo y administración de cuestionarios de encuestas, también se conoce como recopilación de información de encuestas asistidas por computadora (CASIC, por sus siglas en inglés). Los entrevistadores llevan un computador portátil desde el cual leen las preguntas e ingresa las respuestas a las preguntas de la encuesta, posteriormente, los datos se transmiten a la central por medio de un módem. Uno de los programas de software más utilizado en estas técnicas es Blaise, que fue desarrollado por el Instituto de Estadísticas de Noruega.
CATI (Computer Assisted Telephone Interview)	**Banco Mundial**	"Entrevista Telefónica Asistida por Computadora: el entrevistador utiliza un instrumento basado en computadora en lugar de un instrumento en papel. El instrumento informático muestra preguntas en la pantalla y el entrevistador se las lee al encuestado por teléfono e ingresa las respuestas del encuestado directamente en la computadora. Requiere de una infraestructura en casa para realizar entrevistas descentralizadas como: electricidad, buena cobertura telefónica, conexión a internet adecuada y estable, teléfono, auriculares y dispositivo informático para cargar el instrumento de encuesta CATI. La información de la muestra debe estar pre cargada e incluir: números de teléfono para llegar a las unidades, identificadores del hogar, idioma preferido del hogar, entrevistador que realizó encuestas anteriores si se trata de una encuesta de panel. "

	INEGI - MÉXICO	Esta modalidad consiste en realizar la entrevista vía telefónica, es decir, se cuenta con un entrevistador que hace las preguntas al informante de un lado de la línea del teléfono y del otro lado les da respuesta el entrevistado. Las entrevistas asistidas por computadora se empezaron a aplicar en centros telefónicos con varios entrevistadores, dando lugar a las entrevistas telefónicas asistidas por computadora. Tal vez la ventaja más importante de las CATI es que con este tipo de instrumento se puede tener la habilidad de evitar saltos equivocados que se llegan a dar en cuestionarios de papel muy complejos, lográndose así que las tasas de no-respuesta sean generalmente más bajas que con las PAPI, debido a que con el teléfono los pases se hacen de manera automática.
CAWI (Computer Aided Web Interviewing)	**FAO - The Food and Agriculture Organization**	"Autoentrevista asistida por ordenador (CASI) o entrevista web asistida por ordenador (CAWI) con cuestionario electrónico en línea, cuenta con los siguientes aspectos: • La oficina del censo envía un aviso a los encuestados con instrucciones sobre: a) cómo acceder a los cuestionarios web con un código de acceso seguro; b) un centro de llamadas de ayuda; y c) instrucciones sobre cómo completarlo. • El código de acceso seguro es necesario para autenticar a los encuestados, lo que les permite acceder a la aplicación y notificar a la operación de recogida sobre el terreno una vez que los encuestados han transmitido el cuestionario. • El cuestionario CASI/CAWI suele incluir información de ayuda para la navegación, menús desplegables y ediciones en línea. Las ediciones pueden simplificarse para reducir la frustración de los encuestados por los errores y la carga de respuesta asociada. • En los cuestionarios CASI/CAWI se incorporan patrones de salto para que solo se presenten a los encuestados las preguntas relacionadas con el tipo de operaciones de explotación."

	Departamento de Salud Pública de Polonia	La entrevista web asistida por computadora (CAWI) es un método de investigación que es el resultado de la evolución de los siguientes métodos populares utilizados anteriormente: entrevistas con lápiz y papel (PAPI) y entrevistas telefónicas/personales asistidas por computadora (CATI/CAPI). Con la creciente popularización del uso de Internet, el CAWI se está convirtiendo en una herramienta de investigación con un gran grupo de seguidores. El método CAWI implica la creación de un cuestionario de investigación, que se mostrará en el sitio web de tal manera que esté disponible en línea para que lo completen los encuestados (Mider, 2013). Las preguntas y respuestas del cuestionario están estandarizadas y previamente predefinidas (Fowler et al., 1990). La literatura a menudo plantea el problema de la denominación incorrecta del método CAWI (como una entrevista), que, como una modificación de los métodos clásicos (PAPI, CATI) de investigación por cuestionario, sugiere la presencia de una persona adicional que realiza la entrevista, funcionando como un intermediario entre el encuestado y el cuestionario. De hecho, el CAWI es más un estudio de encuesta en el que los encuestados completan un cuestionario sin la participación de la persona que realiza el estudio (Mider, 2013). Este problema se refleja directamente en el estándar metodológico del cuestionario o entrevista (la diferencia en el control del proceso de investigación). Otro problema con el método CAWI se refiere a la selección de una muestra de investigación que permitiría la generalización de los resultados estadísticos a toda la población (Bethlehem, 2009; Kraut et al., 2004). Esto parece ser difícil debido a dificultad para determinar la representatividad de los encuestados.
CAPI (Computer Assisted Personal Interviewing)	**Naciones Unidas**	"Entrevista personal asistida por computadora (CAPI): se refiere a un método de recopilación de datos cara a cara donde el entrevistador no usa cuestionarios de papel, sino que dispone de una computadora, tableta o teléfono para administrar el cuestionario al encuestado, haciendo que la recopilación de datos sea más fácil y rápida, los beneficios que ofrece son: • Mejorar la puntualidad de la recopilación de datos. • Garantizar la calidad y la comparabilidad de los datos • Permitir la recopilación de nuevos tipos de información/datos • Solución rentable y sostenible para los INES."
	Banco Mundial	El Banco Mundial define a las entrevistas personales asistidas por computadora (CAPI) como un método de recopilación de datos cara a cara en el que el entrevistador usa una tableta, un teléfono móvil o una computadora para registrar las respuestas dadas durante la entrevista. CAPI como los otros métodos de recopilación puede presentar tanto ventajas como desventajas a la hora de considerar si encaja bien según el tipo y diseño de cuestionario. Adicionalmente, con la apuesta tecnológica de estos métodos, el Banco Mundial plantea una lista de criterios necesarios a la hora de escoger un software CAPI.

Fuente: DANE a partir de las revisiones de referentes

A continuación en la tablaTabla 17 se presentan los conceptos estandarizados por el DANE.

Tabla 17. Términos estandarizados para Colombia

TÉRMINO INICIAL	PROPUESTA FINAL DE ESTANDARIZACIÓN
Entrevista asistida por computador (CAI) Computer Assisted Interviewing	Método de recolección de datos por computador para el desarrollo y la administración de encuestas.
Entrevista asistida por Teléfono (CATI) Computer Assisted Telephone Interview	Método de recolección de datos por teléfono en el que se cuenta con un entrevistador que hace las preguntas al informante de un lado de la línea del teléfono y las respuestas son ingresadas directamente al sistema.
Entrevista web asistida por computador (CAWI) Computer Aided Web Interviewing	Método de recolección de datos en el que el encuestado de manera autónoma realiza en la web la entrevista mediante un cuestionario electrónico.
Entrevista personal asistida por computador - CAPI - (Computer Assisted Personal Interviewing)	Método de recolección de datos donde el entrevistador de manera presencial dispone de un computador o de un dispositivo móvil de captura para gestionar el cuestionario del encuestado.

Fuente: DANE a partir de las revisiones de referentes

En el DANE se documentan 112 operaciones estadísticas, de las cuales 45 utilizan al menos uno de los métodos de recolección mencionados en la Tabla 17. A continuación en la Tabla 18, se relacionan las operacioes estadísticas, el tipo de operación, el método de recolección y la clasificación con los métodos estandarizados.

Tabla 18 Clasificación de las Operaciones Estadísticas vigentes del DANE según métodos de recolección

NOMBRE FINAL	SIGLA	TIPO DE OPERACIÓN	METODO DE RECOLECCIÓN DE DATOS	METODO DE RECOLECCIÓN DE DATOS CAI
Censo de Edificaciones	CEED	Censo	Dispositivo móvil de captura - DMC	CAPI
Censo de Habitantes de la Calle 2019 (22 municipios)	CHC	Censo	Dispositivo móvil de captura - DMC	CAPI
Censo Económico	CE	Censo	Dispositivo móvil de captura - DMC	CAPI
Censo Nacional Minero	CMN	Censo	DMC + Formulario electronico	CAPI + CAWI
Encuesta Ambiental Industrial	EAI	Muestreo probabilistico	Formulario electronico	CAWI
Encuesta Anual de Comercio	EAC	Censo	Formulario electronico	CAWI
Encuesta Anual de Inversión Directa	EAID	Muestreo no probabilistico	Formulario electronico	CAWI

Encuesta Anual de Servicios	EAS	Censo	Formulario electronico	CAWI
Encuesta Anual Manufacturera	EAM	Censo	Formulario electronico	CAWI
Encuesta de Consumo Cultural	ECC	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Encuesta de Convivencia y Seguridad Ciudadana	ECSC	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Encuesta de Cultura Política	ECP	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Encuesta de Desarrollo e Innovación Tecnológica en la Industria Manufacturera	EDIT	Censo	Formulario electronico	CAWI
Encuesta de Desarrollo e Innovación Tecnológica en los sectores de Servicios y comercio	EDITS	Censo	Formulario electronico	CAWI
Encuesta de Gasto Interno en Turismo	EGIT	Muestreo probabilistico	Telefono	CATI
Encuesta de Micronegocios	EMICRON	Muestreo probabilistico	Telefono + DMC	CATI + CAPI
Encuesta de Pulso Empresarial	EPE	Censo	Formulario electronico	CAWI
Encuesta de Pulso Social	EPS	Muestreo probabilistico	Telefono	CATI
Encuesta de Sacrificio de Ganado	ESAG	Censo	Formulario electronico	CAWI
Encuesta de Transporte Urbano de Pasajeros	ETUP	Censo	Formulario electronico	CAWI
Encuesta Longitudinal de Colombia	ELCO	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Encuesta Mensual de Alojamiento	EMA	Muestreo probabilistico	Formulario electronico	CAWI
Encuesta Mensual de Comercio	EMC	Muestreo probabilistico	Formulario electronico	CAWI
Encuesta Mensual de Comercio al Exterior	EMCES	Muestreo	Formulario electronico	CAWI
Encuesta Mensual de Servicios	EMS	Censo	Formulario electronico	CAWI
Encuesta Mensual Manufacturera con Enfoque Territorial	EMMET	Muestreo no probabilistico	Formulario electronico	CAWI
Encuesta Multipropósito	EM	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Encuesta Nacional Agropecuaria	ENA	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Encuesta Nacional de Arroz Mecanizado	ENAM	Muestreo probabilistico	Formulario electronico	CAWI
Encuesta Nacional de Calidad de Vida	ECV	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Encuesta Nacional de TIC Empresas	ENTIC_E	Muestreo	Formulario electronico	CAWI
Encuesta Nacional de TIC Hogares	ENTIC_H	Muestreo	Dispositivo móvil de captura - DMC	CAPI

Encuesta Nacional de Uso del Tiempo	ENUT	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Encuesta sobre Ambiente y Desempeño Institucional Departamental	EDID	Censo	Formulario electronico	CAWI
Encuesta sobre Ambiente y Desempeño Institucional Nacional	EDI	Muestreo probabilistico	Formulario electronico	CAWI
GEIH Marco 2018	P_GEIH	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Gran Encuesta Integrada de Hogares	GEIH	Muestreo probabilistico	Dispositivo móvil de captura - DMC	CAPI
Investigación de Educación Formal	EDUC	Censo	Formulario electronico	CAWI
Muestra Trimestral de Agencias de Viajes	MTA	Muestreo probabilistico	Formulario electronico	CAWI
Muestra Trimestral de Servicios Bogotá	EMSB	Muestreo no probabilistico	Formulario electronico	CAWI
Precio de Venta al Público de Licores, Vinos y aperitivos similares	PVPLVA	Muestreo	Formulario electronico	CAWI
Programa de Paridad de Poder Adquisitivo	PPA	Muestreo	DMC + Formulario electronico + Papel	CAPI + CAWI
Sistema de Información de Precios y Abastecimiento del Sector Agopecuario, Componente de Precios Mayoristas	SIPSA_P	Muestreo no probabilistico	Dispositivo móvil de captura - DMC	CAPI
Sistema de Información de Precios y Abastecimiento del Sector Agropecuario Componente Abastecimiento de Alimentos	SIPSA_A	Muestreo no probabilistico	Dispositivo móvil de captura - DMC	CAPI
Sistema de Información de Precios y Abastecimiento del Sector Agropecuario Componente de Insumos y Factores Asociados a la Producción Agropecuaria	SIPSA_I	Muestreo no probabilistico	Dispositivo móvil de captura - DMC	CAPI

Fuente: DANE

3.3 Conclusiones

Actualmente en el DANE los principales métodos de recolección de información utilizados son:

- CATI: Entrevista telefónica asistida por computador.
- CAWI: Entrevista web asistida por computador.
- CAPI: Entrevista personal asistida por computador.

Sin embargo, es pertinente explorar los métodos i) CASI, autoentrevista asistida por computadora y ii) AudioCASI, autoentrevista asistida mediante audio, pueden ser utilizados

particularmente en encuestas con temática delicada, por ejemplo delitos y ofensas o comportamiento y actitudes sexuales, donde el entrevistado ingresa por sí mismo sus respuestas, se considera que aumenta la validez de las respuestas, ya que es más probable que los encuestados den respuestas veraces al no tener un contacto directo con el entrevistador.

La preparación del Reporte de esta edición participamos los siguientes funcionarios:

Dahann Valentina Pérez Zárate - dvperezz@dane.gov.co

Erik Stopwar Arciniegas Rincón - esarciniegasr@dane.gov.co

Grace Andrea Torres Pineda - gatorresp@dane.gov.co

Heidy Patricia Forero Muhete - hpforerom@dane.gov.co

Johana Catherine Avila Alvarado - jcavilaa@dane.gov.co

Juliana Catalina Pastás Pastás – jcpastasp@dane.gov.co

Milena Del Rosario Escobar Morillo - mrescobarm@dane.gov.co

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

Revisión de estilo por: Juan Camilo Giraldo Manrique – jcgiraldom@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 al correo: - jcavilaa@dane.gov.co