Proceso de Anonimización de datos DANE.

De datos protegidos al análisis de información:

Anonimización de datos para actores públicos.

Subdirección de Prospectiva, Información y Evaluación Estratégica

Unidad de producción de Información Estadística
Equipo de clasificación Socioeconómica
Plan Estadístico Distrital - Política de Gestión de Información Estadística

Mayo de 2026

Alcaldía de Medellín
Distrito de
Ciencia, Tecnología e Innovación

Contenido

Departamento Administrativo de Planeación

- **Proceso de Anonimización**
  - Las 6 etapas del proceso de anonimización de datos
- **Caso de Anonimización ECV**

De datos protegidos al análisis de información:

Anonimización de datos para actores públicos.

Alcaldía de Medellín
Distrito de
Ciencia, Tecnología e Innovación

LA REFINERÍA DE DATOS: LAS 6 ETAPAS DEL PROCESO DE ANONIMIZACIÓN

METODOLOGÍA DEL SISTEMA ESTADÍSTICO NACIONAL (SEN) 2024

ETAPA 1: EVALUACIÓN DE RIESGO (RISK ASSESSMENT)

MÉTRICA: CÁLCULO DE RE-IDENTIFICACIÓN.
ENTRADA: DATOS CRUDOS.
PROCESO: ANÁLISIS DE VARIANZAS.
SALIDA: PERFIL DE RIESGO.

ETAPA 2: DISEÑO DEL MODELO (MODEL DESIGN)

MÉTRICA: SENSIBILIDAD UTILIDAD/PRIVACIDAD.
ENTRADA: TEOREMA DE BAYES.
PROCESO: SELECCIÓN DE ALGORITMO (K-ANONYMITY, L-DIVERSITY).
SALIDA: ESTRATEGIA DE ANONIMIZACIÓN.

ETAPA 3: TRANSFORMACIÓN DE DATOS (DATA TRANSFORMATION)

MÉTRICA: FRECUENCIA DE PROCESAMIENTO.
ENTRADA: DATOS CRUDOS Y ESTRATEGIA.
PROCESO: APLICACIÓN DE TÉCNICAS.
SALIDA: DATOS PROCESADOS.

ETAPA 4: VALIDACIÓN DE PRIVACIDAD (PRIVACY VALIDATION)

MÉTRICA: CUMPLIMIENTO NORMATIVO (SDPS, SEN).
ENTRADA: DATOS TOKENIZADOS.
PROCESO: PRUEBAS DE RE-IDENTIFICACIÓN.
SALIDA: VERIFICACIÓN DE PRIVACIDAD.

ETAPA 5: EVALUACIÓN DE UTILIDAD (UTILITY EVALUATION)

MÉTRICA: PÉRDIDA DE INFORMACIÓN.
ENTRADA: DATOS ANONIMIZADOS Y ORIGINALES.
PROCESO: COMPARACIÓN ESTADÍSTICA.
SALIDA: INFORME DE CALIDAD DE DATOS.

ETAPA 6: GESTIÓN Y DISTRIBUCIÓN (MANAGEMENT & DISTRIBUTION)

MÉTRICA: CONTROL DE ACCESO Y AUDITORÍA.
ENTRADA: DATOS PROTEGIDOS.
PROCESO: FORMACIÓN SOBRE ACCESO CONTROLADO.
SALIDA: ACCESO A DATOS ANONIMIZADOS.

¿Qué es la Anonimización y por qué es esencial?

DEFINICIÓN (Decreto 1743/2016)

*"Proceso técnico para transformar datos individuales de modo que **no sea posible identificar** a las personas o sus características individuales, preservando las propiedades estadísticas."*

Tensión central

Acceso máximo a la información vs. protección de la privacidad individual. La anonimización resuelve esta disyuntiva.

Crecimiento de microdatos

Mayor demanda de datos desagregados, big data y registros administrativos exige marcos metodológicos actualizados.

Reserva estadística

Ley 79/1993: garantiza confidencialidad de fuentes. La anonimización permite publicar sin violar la reserva.

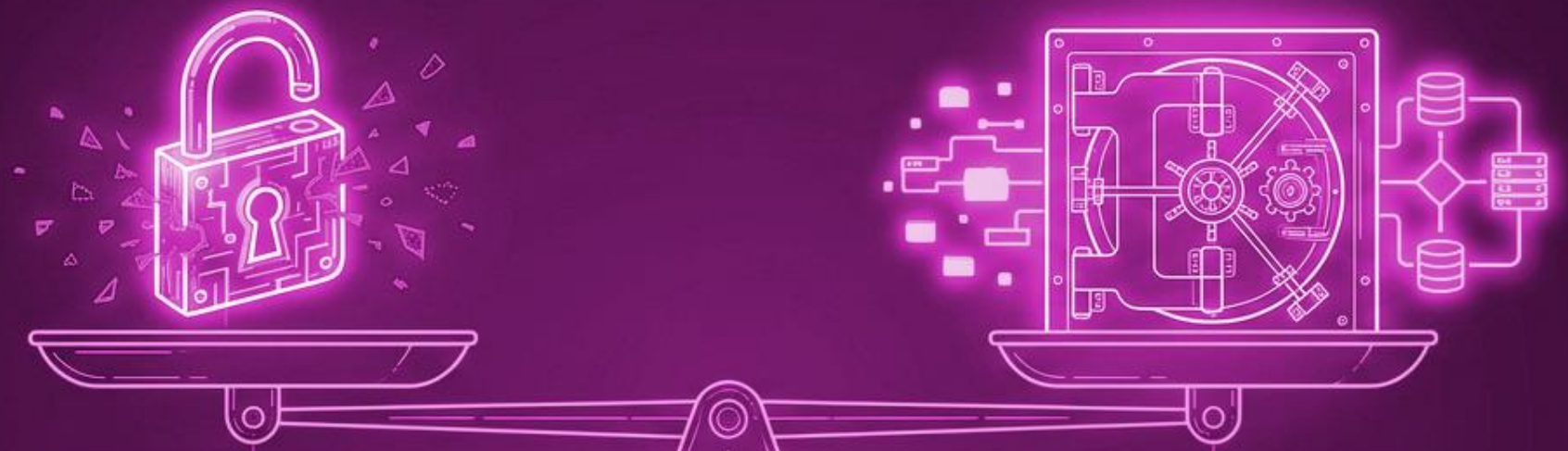
PRINCIPIOS SEN (Código BPE)

PRINCIPIO 10
Accesibilidad: máximo detalle posible para todos los usuarios

PRINCIPIO 11
Confidencialidad: técnicas de anonimización para proteger fuentes

LEY 2335/2023
Fortalece aprovechamiento de registros administrativos en el SEN

LA TENSIÓN CENTRAL DE LA INFORMACIÓN PÚBLICA



El Dilema

La demanda creciente de microdatos y big data exige un mayor detalle estadístico, pero la Ley 79/1993 y el Código BPE obligan a garantizar la confidencialidad absoluta de las fuentes.

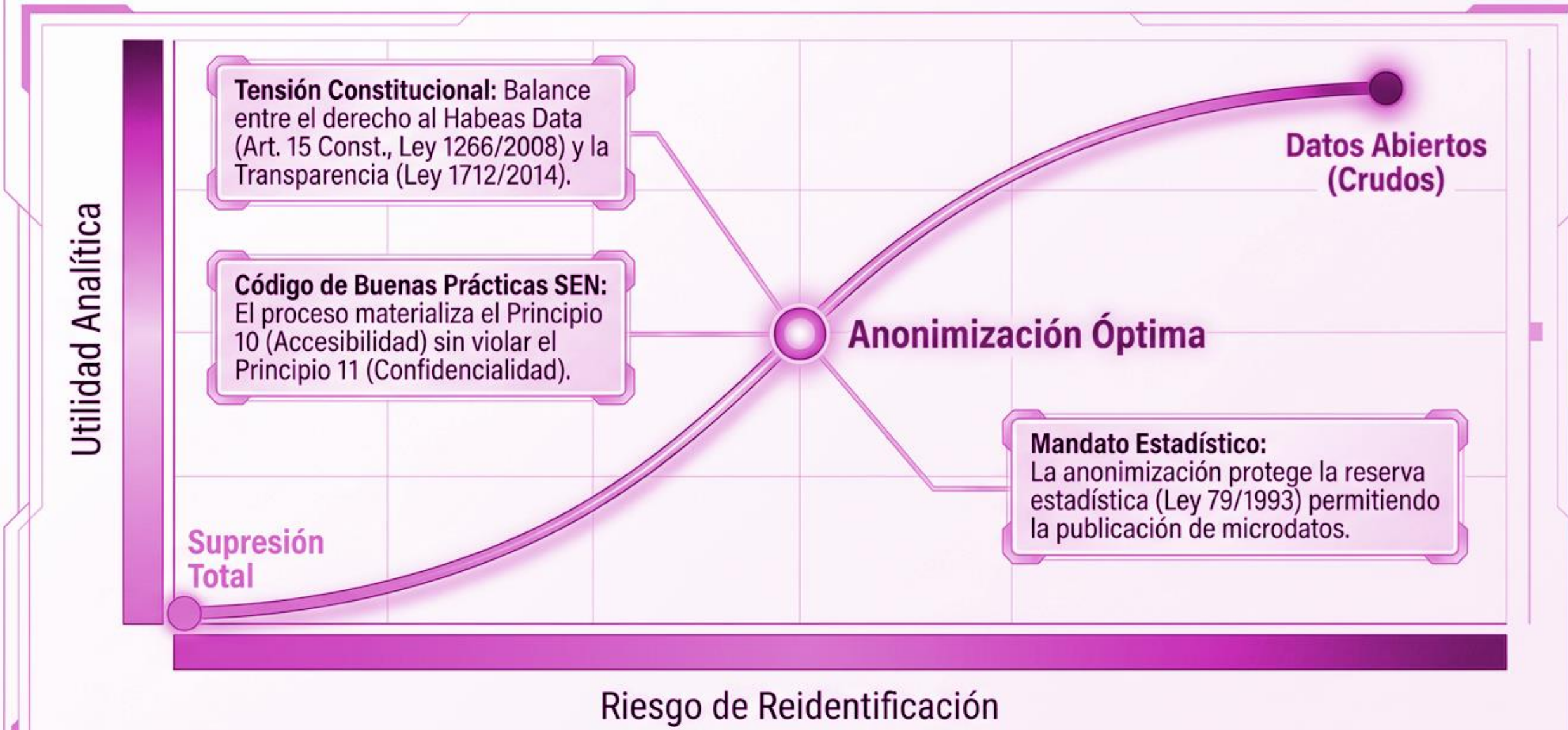
La Solución Técnica

La anonimización no es ocultar datos; es la resolución de esta disyuntiva. (Decreto 1743/2016).

El Objetivo

Transformar datos individuales para que sea imposible identificar a las personas, preservando intactas sus propiedades estadísticas.

La Curva de Intercambio Privacidad-Utilidad



Fuente:: Guía Para La Anonimización de Datos Estructurados – DANE. Sistema Estadístico Nacional. 2024 De Medellín

Evolución del Marco Normativo en Colombia

Derechos Fundamentales (1991 - 2008)

- **1991:** Art. 15 Constitución (Habeas data fundamental).
- **1993: Ley 79** (Reserva estadística DANE).

Protección de Datos Personales (2008 - 2016)

- **2008:** Ley 1266 (Habeas data en bases financieras/crediticias).
- **2012:** Ley 1581 (Protección de datos personales).
- **2016: Decreto 1743** (Marco del Sistema Estadístico Nacional - SEN).

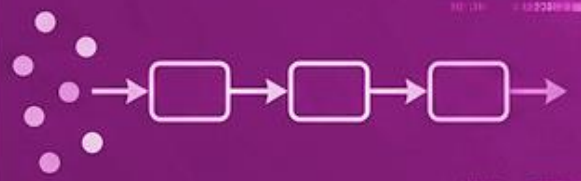
Mandatos Estadísticos (2023 - Presente)

- **2023: Ley 2335** (Nueva Ley Estadística, fortalecimiento de registros administrativos).

Evolución Metodológica: De la Lista de Chequeo al Bucle Iterativo

DANE 2018

Estructura: 3 Etapas Lineales.



Estructura: 3 Etapas Lineales.
Enfoque: Frecuencias básicas.
Software: Código SAS.
Alcance: Encuestas + Registros Administrativos.

Medellín 2022

Estructura: 4 Etapas (Adaptación distrital).

Estructura: 4 Etapas / Filtral.
Enfoque: Unicidad / rareza.
Software: Sin código implementado.
Alcance: Primer piloto subnacional.

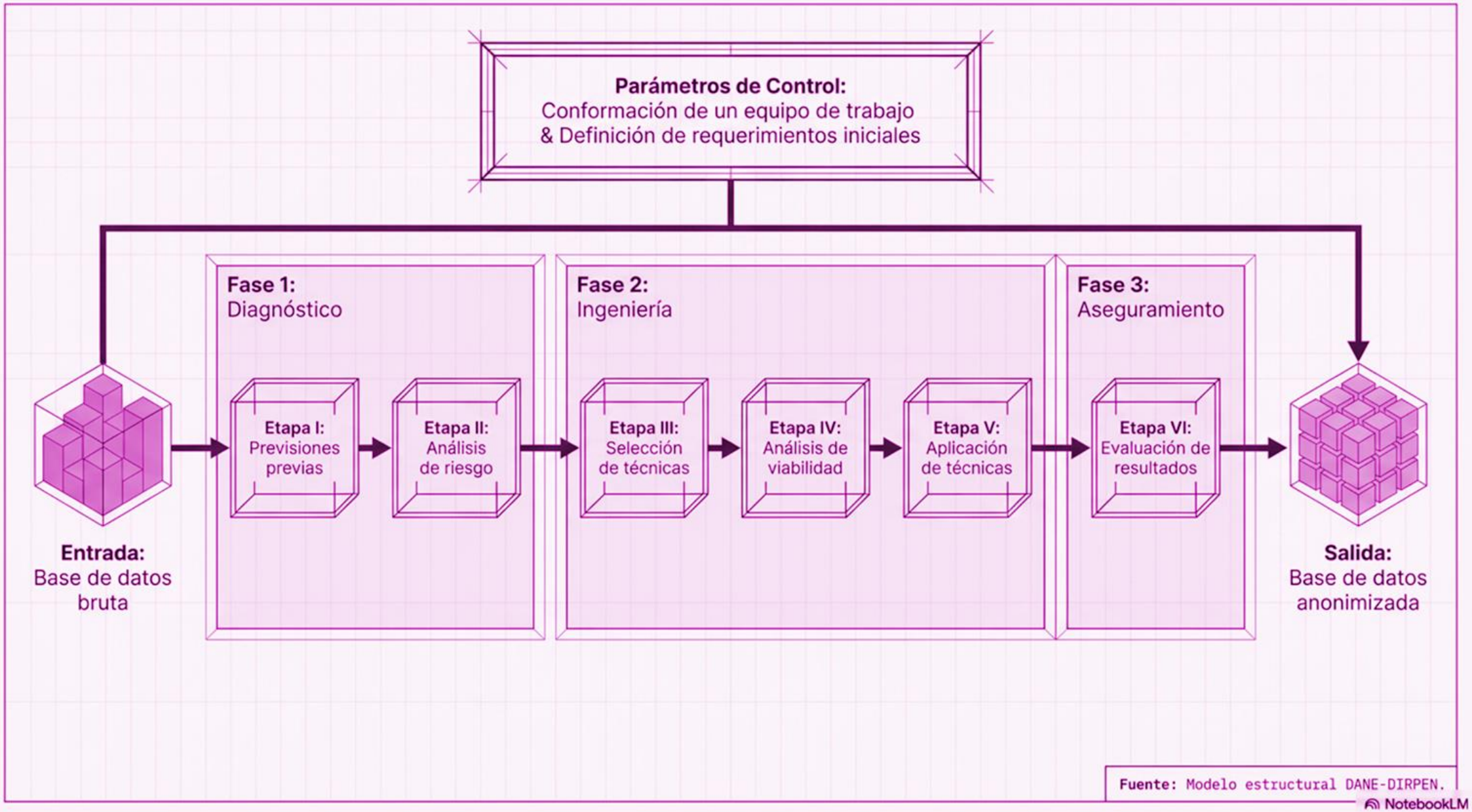
DANE 2024 (Actual)

Estructura: 6 Etapas Iterativas

Estructura: 6 Etapas Iterativas
Enfoque: Simulación de ataques externos.
Software: R (SDCmicro) + SAS.
Alcance: SEN + Big Data + Privacidad Diferencial en el horizonte.

Anatomía del Riesgo en los Microdatos

NotebookLM


Parámetros de Control:
Conformación de un equipo de trabajo
& Definición de requerimientos iniciales
Fase 1:
Diagnóstico
Fase 2:
Ingeniería
Fase 3:
Aseguramiento
Etapa I:
Previsiones previas
Etapa II:
Análisis de riesgo
Etapa III:
Selección de técnicas
Etapa IV:
Análisis de viabilidad
Etapa V:
Aplicación de técnicas
Etapa VI:
Evaluación de resultados
Entrada:
Base de datos bruta
Salida:
Base de datos anonimizada
Fuente: Modelo estructural DANE-DIRPEN.
NotebookLM

El Motor Iterativo de Anonimización (DANE 2024)

1. Revisiones Previas
Normativa, necesidades de usuarios, y análisis exploratorio.

2. Análisis de Riesgos
Clasificación de variables y modelamiento de escenarios de ataque.

3. Selección de Técnicas
Mapeo de riesgos vs. perturbación/no perturbación.

4. Viabilidad
Verificación de utilidad analítica teórica.

5. Aplicación
Procesamiento matemático y alteración de los microdatos.

6. Documentación y Reevaluación
Verificación final frente a ejercicios de rompimiento.

1. Revisiones Previas | 2. Análisis de Riesgos | 3. Selección de Técnicas | 4. Análisis de Viabilidad | 5. Aplicación de Técnicas | 6. Evaluación de Resultados

Etapa I: Revisiones Previas (Los Cimientos)

Análisis Exploratorio

- Definir el diccionario de datos.
- Clasificar tipos de variables (cuantitativas vs. categóricas).
- Evaluar el volumen y las dimensiones de la base bruta.

Revisión Normativa

- Identificar el marco legal (Habeas Data, Ley 1581).
- Revisar el histórico de demandas de usuarios (¿Qué datos piden los investigadores?).

Propiedades Estadísticas

- El requisito innegociable: Definir las medidas descriptivas y distribuciones que deben conservarse exactas o con variación mínima (ej. promedios departamentales).

2. Análisis de Riesgos

Etapa II: Análisis de Riesgos (La Matriz de Amenazas)

Clasificación de la Vulnerabilidad

⚠ Identificadores Directos:
Nombres, cédulas
(Requieren eliminación inmediata).

⚠ Pseudoidentificadores:
Edad, municipio, profesión
(Peligrosos cuando se combinan).

⚠ Variables Sensibles:
Ingresos, etnia, salud
(El objetivo del atacante).

Simulación de Escenarios

- Análisis de unicidad de registros.
- Ejercicios de ataque simulado cruzando bases externas.
- **Salida:** Identificación precisa de las Unidades de Observación Riesgosas.

1. Revisiones Previas | 2. Análisis de Riesgos | **3. Selección de Técnicas** | 4. Análisis de Viabilidad | 5. Aplicación de Técnicas | 6. Evaluación de Resultados

Etapa III: Selección de Técnicas (El Arsenal Técnico)

No Perturbación (Reducción de detalle)

- **Eliminación de variables:** Supresión total de identificadores directos.
- **Recodificación global:** Agrupación de edades o salarios en rangos/intervalos.
- **Supresión local:** Reemplazar valores específicos por vacíos para lograr k-anonimato.

Perturbación (Modificación controlada)

- **Microagregación:** Reemplazar valores extremos por el promedio del grupo.
- **Intercambio de datos:** Cruce aleatorio de registros entre unidades seguras y riesgosas.
- **Adición de ruido:** Inyección de varianza aleatoria controlada.

Tecnologías de Horizonte (SEN 2024): Datos sintéticos y Privacidad Diferencial (ε-DP) para censos masivos.

Las Técnicas en Acción: Transformación a Nivel Micro

Microagregación (Perturbación)

Dato Original: Ingresos de 3 personas: $1.2M, $1.3M, $1.4M (Fácilmente identificables).

Dato Transformado: Las 3 personas ahora reportan $1.3M (El promedio del grupo). El valor global se mantiene, el individuo se esconde.

Recodificación Global (No Perturbación)

Dato Original:
Edad exacta: 34, 36, 35 años.

Dato Transformado: Edad agrupada: Rango 30-40 años. Se pierde precisión microscópica, pero se gana anonimato garantizado.

Etapa IV: Análisis de Viabilidad (La Frontera del Riesgo)

Dato Original

- **Dato Original:** Alta utilidad, pero supera la línea de riesgo (No publicable).

Riesgo de Divulgación

Máximo Riesgo Tolerable

Dato Publicado

- **Dato Publicado (El Objetivo):** Maximiza la utilidad manteniéndose justo por debajo del Máximo Riesgo Tolerable.

- **Sin Datos**: Cero riesgo, pero utilidad nula (Inútil para política pública).

Sin Datos

Utilidad de los Datos

Si el dato anonimizado pierde demasiada utilidad estadística, el proceso se declara No Viable.

Etapa V: Aplicación de Técnicas (La Sala de Máquinas)

Selección del Entorno

Implementación de rutinas algorítmicas en software especializado (R vía paquete SDCmicro, rutinas en SAS, o herramientas como τ-Argus).

Ejecución de Código

- Cargue de la base bruta en entornos seguros.
- Consolidación de riesgos.
- Ejecución de la transformación programada.

Seguridad Estricta

La documentación y el código de esta etapa son de acceso restringido absoluto. Si se filtran las rutinas, un tercero podría aplicar ingeniería inversa y revertir la anonimización.

6. Evaluación de Resultados

Etapa VI: Evaluación de Resultados (El Filtro de Calidad)

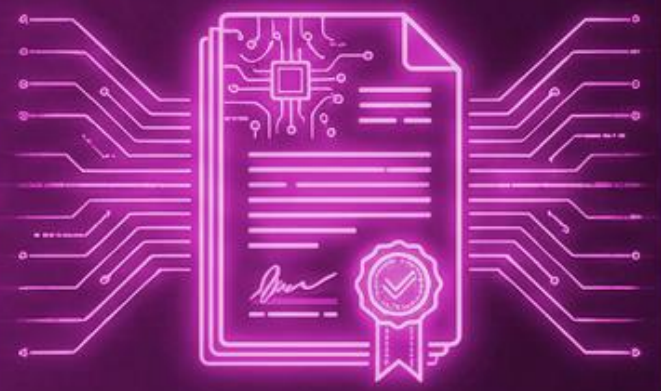
Validación 1: Propiedades Estadísticas

¿Se conservaron los promedios y distribuciones definidos en la Etapa I?

(Tolerancia típica < 5% de variación).

Validación 2: Riesgo Residual

¿Existen nuevas unidades riesgosas creadas accidentalmente por las técnicas de enmascaramiento?

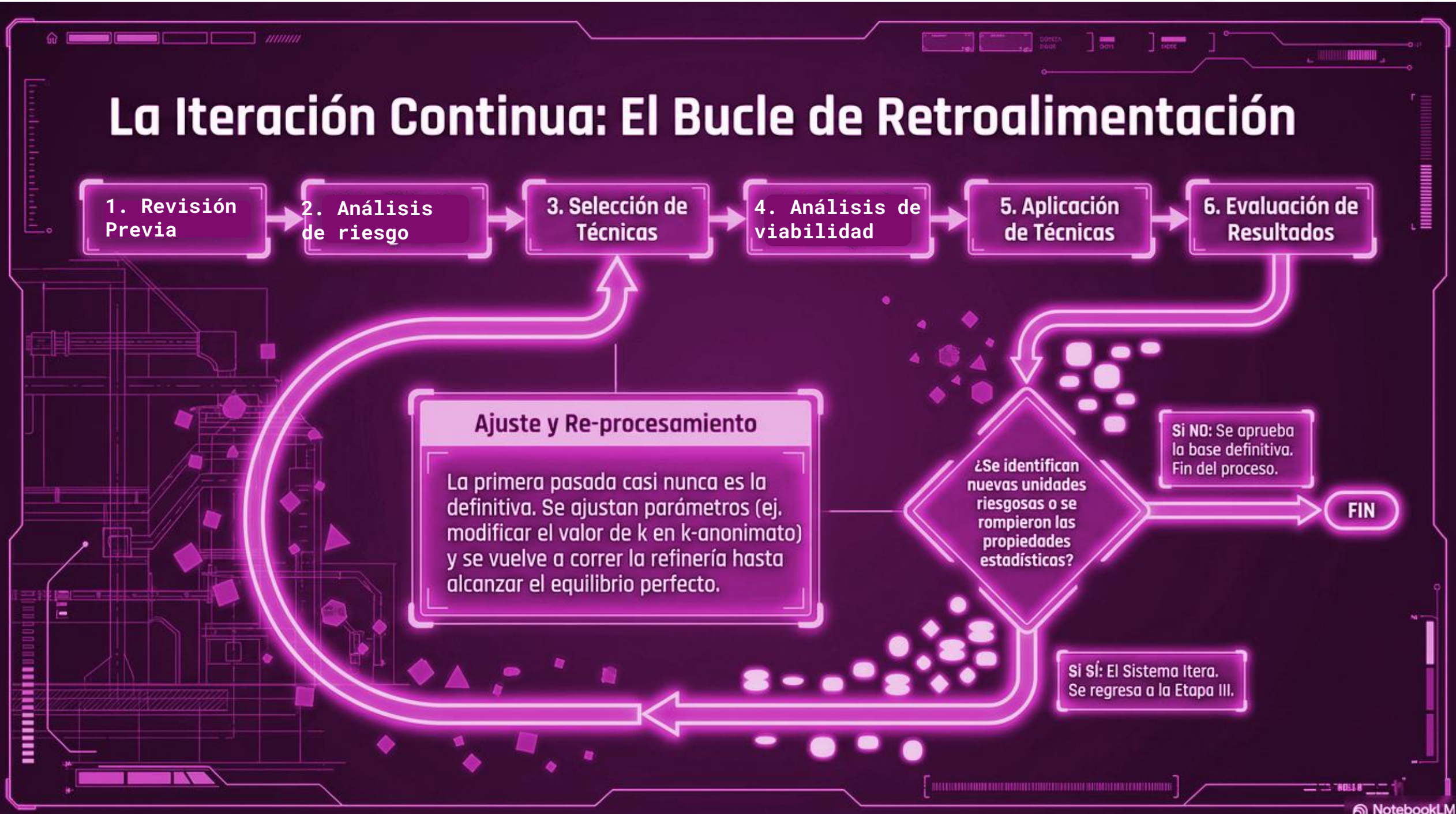


El Entregable: IFPA

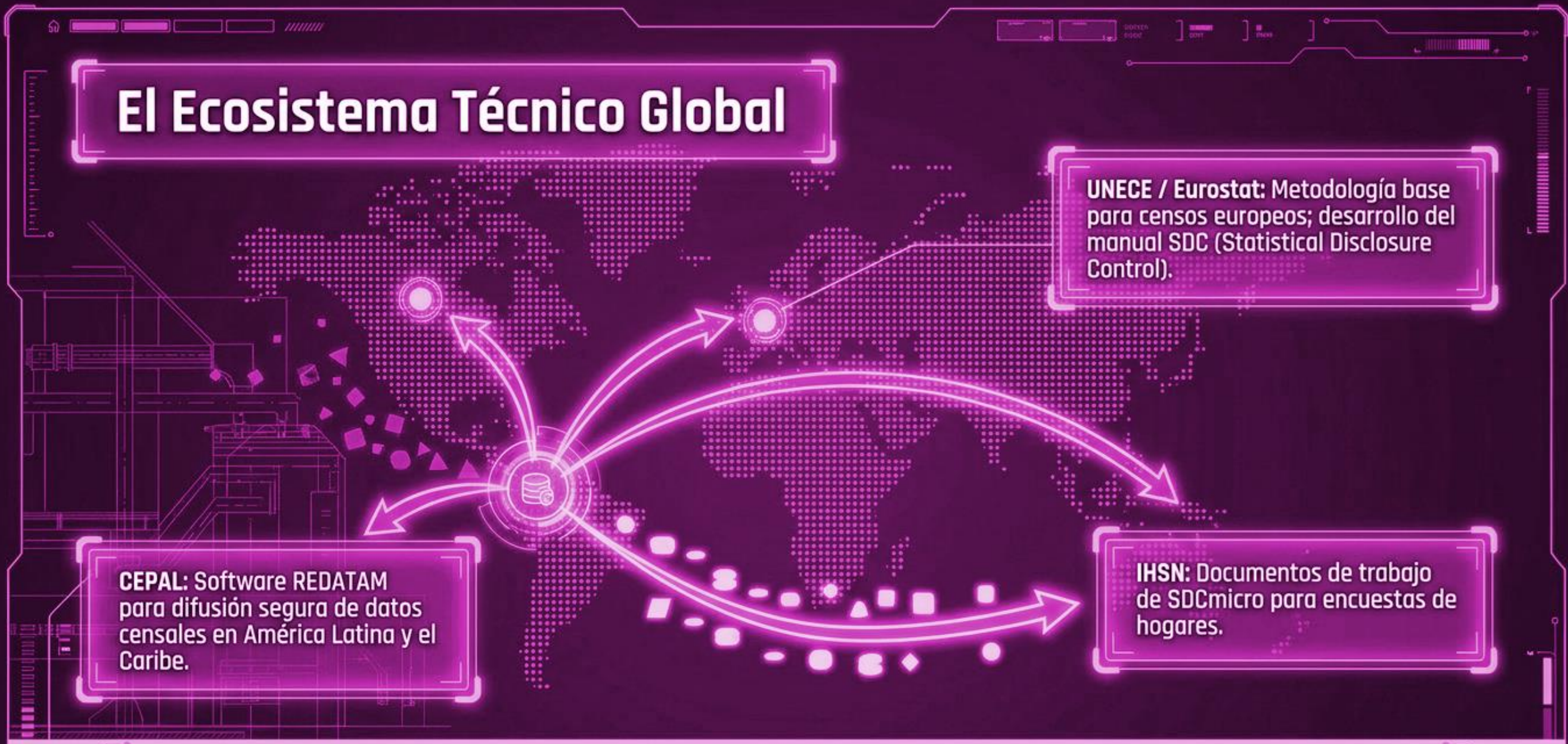
Creación del Informe Final del Proceso de Anonimización.

El documento maestro que certifica la viabilidad y asegura la trazabilidad.

La Iteración Continua: El Bucle de Retroalimentación

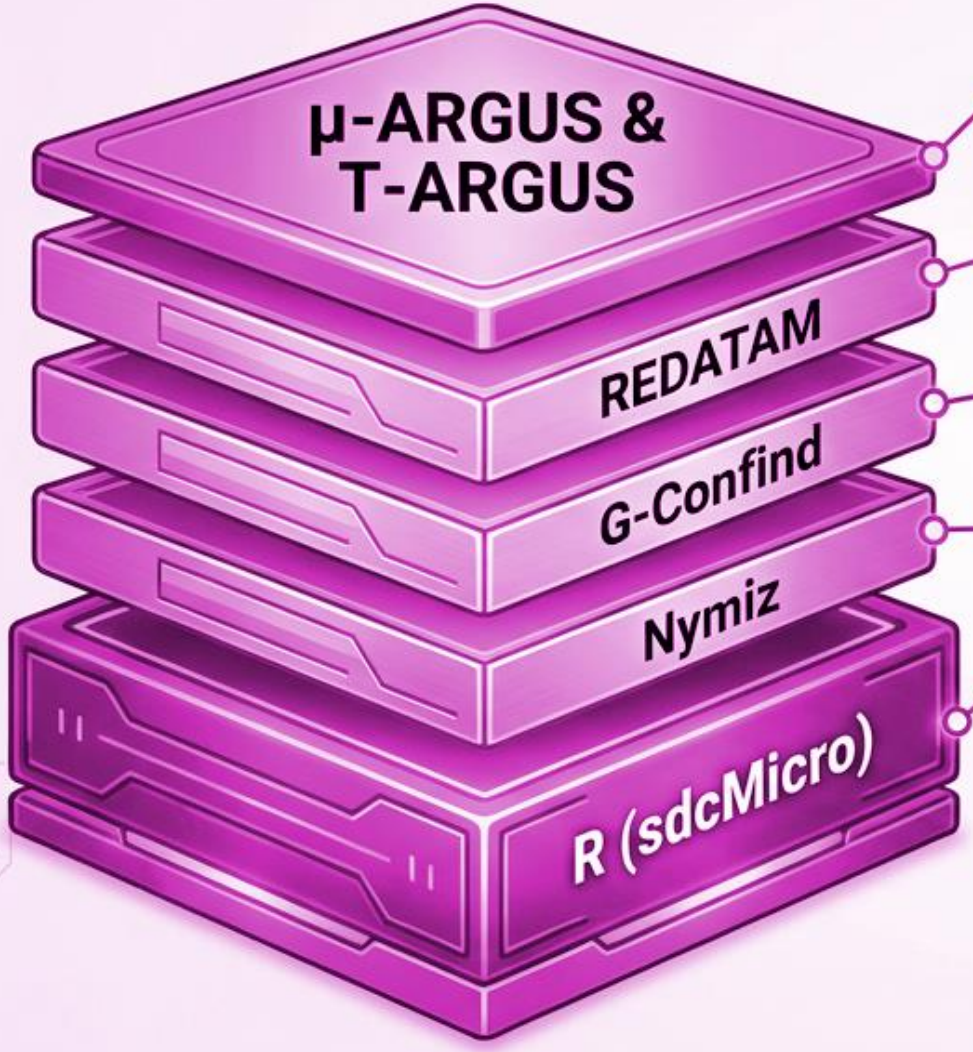


El Ecosistema Técnico Global



La guía 2024 no es una invención aislada, es la homologación del SEN de Colombia con los estándares internacionales

La Infraestructure de Software Global



- Pioneros (Eurostat / Países Bajos). Enfoque en gestión de riesgos y enmascaramiento.
- Plataforma interactiva censal (CEPAL). Perturbación post-tabular vía web.
- Supresión automatizada de celdas (Statistics Canada).
- Solución implementada a nivel subnacional (Ayuntamiento de Barcelona / Planeación Medellín).
- El estándar actual de anonimización del SEN (DANE 2024).

El Límite de Escalabilidad

sdcMicro presenta severos problemas de rendimiento al procesar bases superiores a ~30.000 registros, requiriendo nuevas soluciones de Big Data para censos completos.

Desbloqueando el Valor del SEN

Metodología Estructurada

Pasamos de la improvisación a una ingeniería iterativa de 6 pasos.

Transparencia Segura

Se maximiza el impacto del Big Data y los Registros Administrativos sin comprometer la Ley 79 de reserva estadística.

Valor Estratégico

La anonimización no limita la información; la libera para el análisis académico, la investigación y la creación de políticas públicas basadas en evidencia rigurosa.

Conclusiones y Retos del SEN

HALLAZGOS PRINCIPALES

Evolución progresiva y coherente
DANE 2024 construye sobre 2018 sin ruptura metodológica. Medellín 2022 aporta la dimensión subnacional.

Fortalecimiento del ecosistema institucional
La co-autoría DANE + AGN en 2024 integra perspectivas estadística y archivística, ampliando el alcance de aplicación.

Incorporación de estándares internacionales de vanguardia
DANE 2024 referencia 15+ países/organismos y anticipa tecnologías emergentes (privacidad diferencial, TEE).

Transición de proceso lineal a iterativo
La introducción de ejercicios de ataque y reevaluación cíclica eleva el rigor frente a riesgos de reidentificación.

RETOS PENDIENTES SEN

- **Big data y RRSS** – Adaptar técnicas para datos no estructurados (audio, video, redes sociales)
- **Escalabilidad** – SDCmicro limitado a ~30K registros. Desarrollar herramientas para censos y grandes R.A.
- **Adopción subnacional** – Escalar la experiencia de Medellín 2022 a otras entidades territoriales del SEN
- **Privacidad diferencial** – Pilotear ε-DP para publicaciones agregadas en alianza con UN PET Lab

CASO PRÁCTICO ENCUESTA DE CALIDAD DE VIDA

Alcaldía de Medellín
Distrito de
Ciencia, Tecnología e Innovación

AGRUPACIÓN DE VARIABLES

I. **Variables geográficas:** dirección, barrio, comuna, departamento, distrito

II. Variables de identificación de personas o empresas: NIT, actividad económica, edad, sexo, sector, nivel educativo o etnia

III. Variables con datos numéricos: ingresos, edad, estatura, número de hijos, número de hermanos, área, diámetro

IV. Variables de carácter sensible o confidencial: número de identificación, NIT, ingresos, gastos, costos, impuestos declarados, teléfono, e-mail

V. Variables sin restricción de acceso al público: sexo, edad, años de escolaridad

Fuente:: Guía Para La Anonimización De Bases De Datos Que Se Produce Y Difunde El Distrito Especial De Ciencia, Tecnología E Innovación De Medellín

VARIABLES SUSCEPTIBLES DE ELIMINACIÓN

- Nombres.
- Fecha de nacimiento.
- Fecha de constitución en Cámara de Comercio.
- Números de teléfono y fax.
- Números de identificación: cédula de ciudadanía, pasaporte, tarjeta de identidad, números asociados a la seguridad social, licencias de conducción, NIT, RUT, RUP, RUE, etc.
- Direcciones de correo electrónico.
- Números de cuentas bancarias.
- Identificadores del vehículo, placa, etc.
- Identificadores de dispositivos móviles y números de serie.
- Direcciones de IP.
- Identificadores biométricos.
- Fotografías e imágenes similares.
- Cualquier otro número único de identificación.
- Dirección de domicilio.

Fuente:: Guía Para La Anonimización De Bases De Datos Que Se Produce Y Difunde El Distrito Especial De Ciencia, Tecnología E Innovación De Medellín

ETAPAS PREVIAS A LA ANONIMIZACIÓN

- Carga de información
- Consolidación de datos
- Revisión y Validación
- Verificación de la consistencia interna de los datos y ajustes
- Anonimización de microdatos

Fuente:: Guía Para La Anonimización De Bases De Datos Que Se Produce Y Difunde El Distrito Especial De Ciencia, Tecnología E Innovación De Medellín

TÉCNICAS DE ANONIMIZACIÓN EN ENCUESTA DE CALIDAD DE VIDA

ECV

Encuesta de Calidad de Vida

ELIMINACIÓN DE VARIABLES

Estas variables son eliminadas porque debido ala información contenida, es posible identificar directamente a las unidades de observación.

Nombre; Apellidos; Dirección; Barrio; Municipio; Teléfono

RECODIFICACIÓN GLOBAL

Variables combinarlas categorías de las variables y poder tener en cada nueva categoría una mayor frecuencia de unidades de observación.

Grupo Étnico; Número De Habitaciones De Lacasa; Numero De Viajes Realizados Fuera Del País; Nivel De Escolaridad

Fuente:: Guía Para La Anonimización De Bases De Datos Que Se Produce Y Difunde El Distrito Especial De Ciencia, Tecnología E Innovación De Medellín

Gracias

Alcaldía de Medellín
Distrito de
Ciencia, Tecnología e Innovación

Nota metodológica sobre herramientas de apoyo

Construcción de la estructura conceptual y metodológica:

Perplexity AI. (2026). Asistente de investigación con capacidades de búsqueda, síntesis y generación de reportes estructurados [Software de inteligencia artificial]. https://www.perplexity.ai

Utilizado para: organización de la arquitectura de la presentación, delimitación de ejes temáticos, comparación técnica entre guías de anonimización, identificación de diferencias metodológicas y estructuración de narrativa expositiva para formato de 30 minutos.

Desarrollo de recursos visuales y síntesis conceptual:

Google Labs. (2023). NotebookLM: Herramienta de investigación y aprendizaje asistida por IA [Software de inteligencia artificial]. https://notebooklm.google.com

Utilizado para: elaboración de imágenes de desarrollo conceptual, generación de resúmenes visuales basados en documentos fuente y producción de recursos explicativos derivados de las guías analizadas.

Aclaración metodológica:

El uso de estas herramientas complementan, pero no sustituyó, la revisión técnica de los documentos originales ni el criterio analítico aplicado en la comparación. Su función fue apoyar la estructuración inicial del contenido, facilitar la síntesis metodológica y enriquecer la presentación con recursos visuales, manteniendo como base el análisis técnico de las tres guías de anonimización y la validación conceptual del contenido.