COLOMBIA POTENCIA DE LA VIDA

DANE INFORMACIÓN PARA TODOS 70 AÑOS

“Los datos abiertos en un mundo de 8.000 millones de personas”

Conmemoración Día de la Población en Colombia

Julio 27 de 2023

COLOMBIA POTENCIA DE LA VIDA

DANE INFORMACIÓN PARA TODOS 70 AÑOS

Avances y Desafíos Metodológicos

Producción, Administración, Interoperabilidad de Datos Abiertos

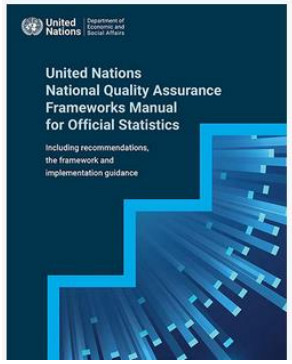
Leonardo Trujillo
Subdirector DANE

Outline

1. DANE -Política de Datos Abiertos
2. ANDA – Archivo Nacional de Datos
3. Técnicas de Anonimización

1994, reafirmados en 2014

2014



2015

2019

2020

UNSD - Data Stewardship Working Group

2021

El DANE un centro de investigación aplicada: gestión del conocimiento en los ecosistemas estadísticos global y nacional

SEN 2.0
Sistema Estadístico Nacional-Colombia

Lineamientos, normas y estándares
Evaluación de la calidad estadística
Marco de Interoperabilidad
Datos abiertos
Marco ético - SETE

CASEN - Consejo Asesor del SEN

CAD – Comité Administración de Datos

CES – Comités Estadísticos Sectoriales

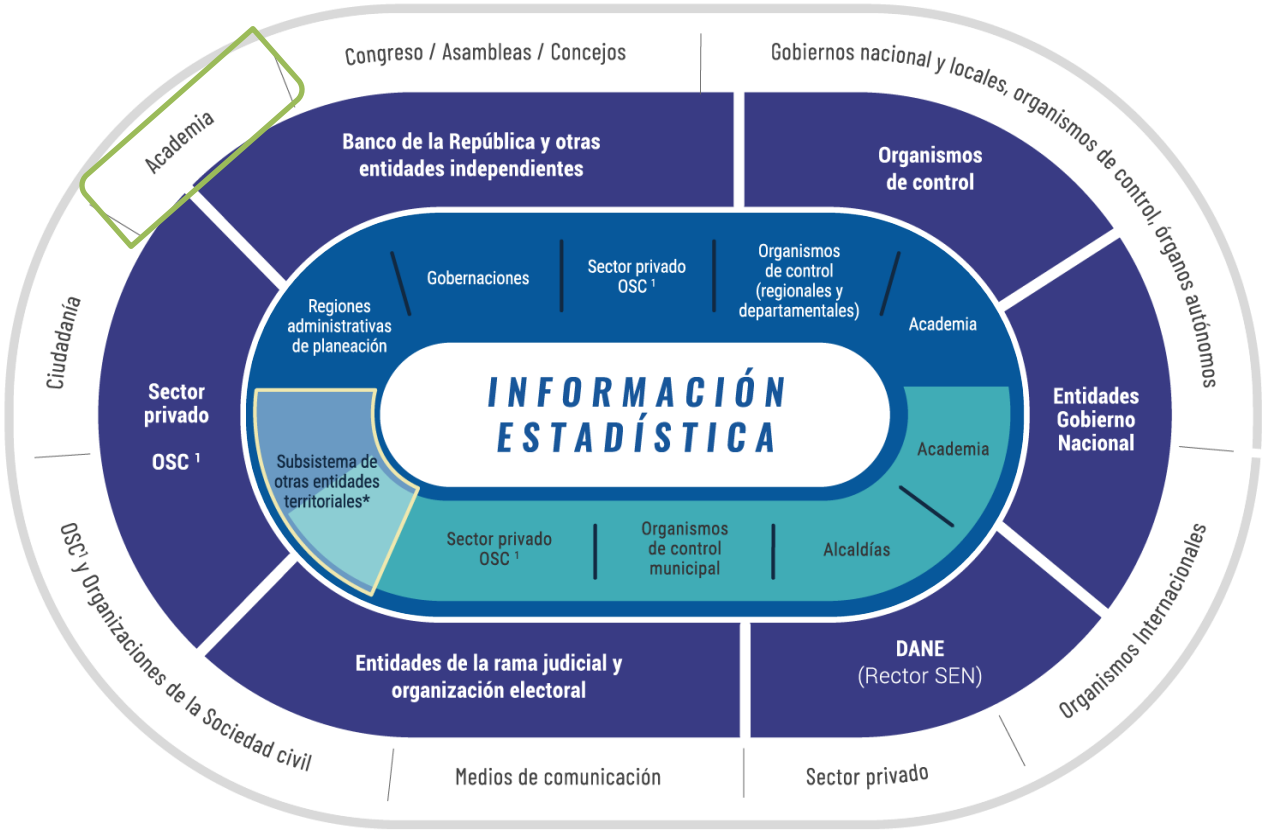
CSES – Comité de Seguimiento a Estadísticas Estratégicas

Ecosistema de datos de las estadísticas oficiales

- **Sistema Estadístico Nacional – SEN**
- **Subsistema Departamental**
- **Subsistema Municipal**
- **Usuarios de información estadística**

*** Subsistema de otras entidades territoriales**

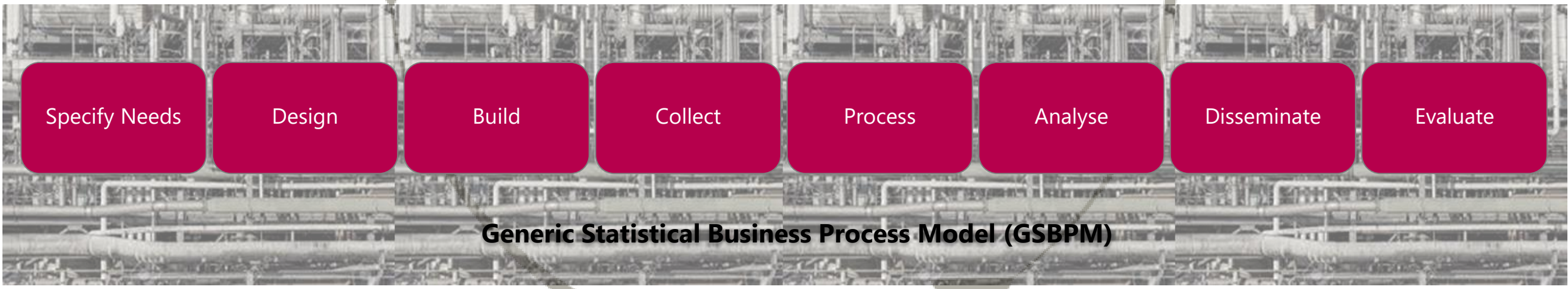
- Provincias
- Áreas metropolitanas
- Territorios indígenas
- Territorio colectivo de comunidades negras
- Asociaciones de municipios
- Territorios PDET



[1] **Artículo 155 de la Ley 1955 de 2019.** Desde la oferta, además de las entidades del orden nacional y subnacional que producen información estadística; están las personas jurídicas, públicas o privadas, que prestan servicios públicos y las personas jurídicas que posean, produzcan o administren registros administrativos en el desarrollo de su objeto social, que sean insumos necesarios para la producción de estadísticas oficiales. Dentro de las organizaciones de la sociedad civil (OSC) se incluyen: las organizaciones no gubernamentales (ONG), entidades sin ánimo de lucro (ESAL) las cuales a su vez, incluyen las cámaras de comercio; gremios; organizaciones sociales, comunitarias, sindicales, de profesionales, étnicas, académicas; entre otras.

129 Statistical Operations
(different collection methods, regularity and coverage)

Agricultural	Economics	Commerce and Industry	Indexes	National Accounts	Social	Technology and Innovation	Government and Institutions
5	11	19	74	46	21	4	6

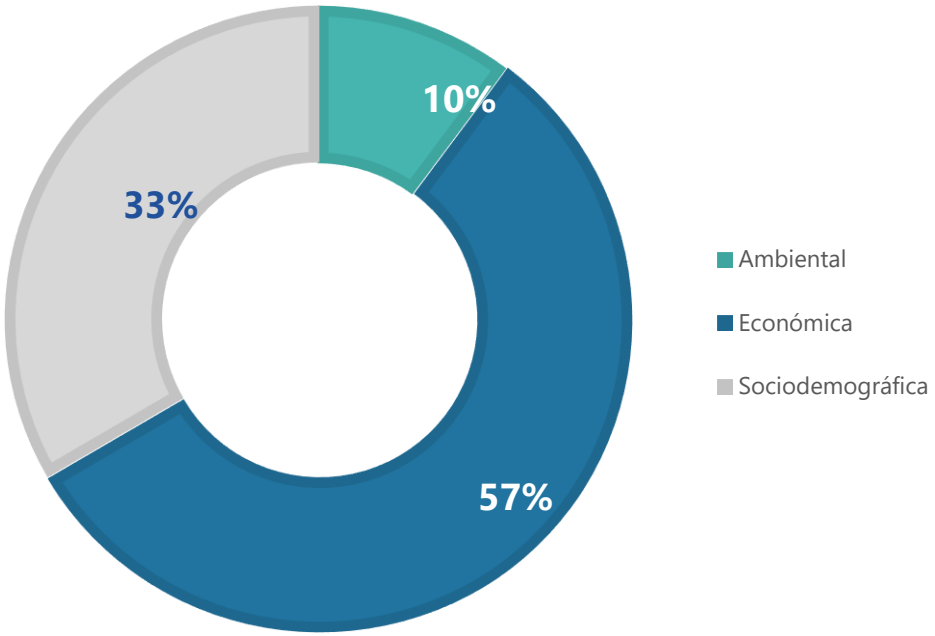


Generic Statistical Business Process Model (GSBPM)

DANE principal productor de estadísticas oficiales

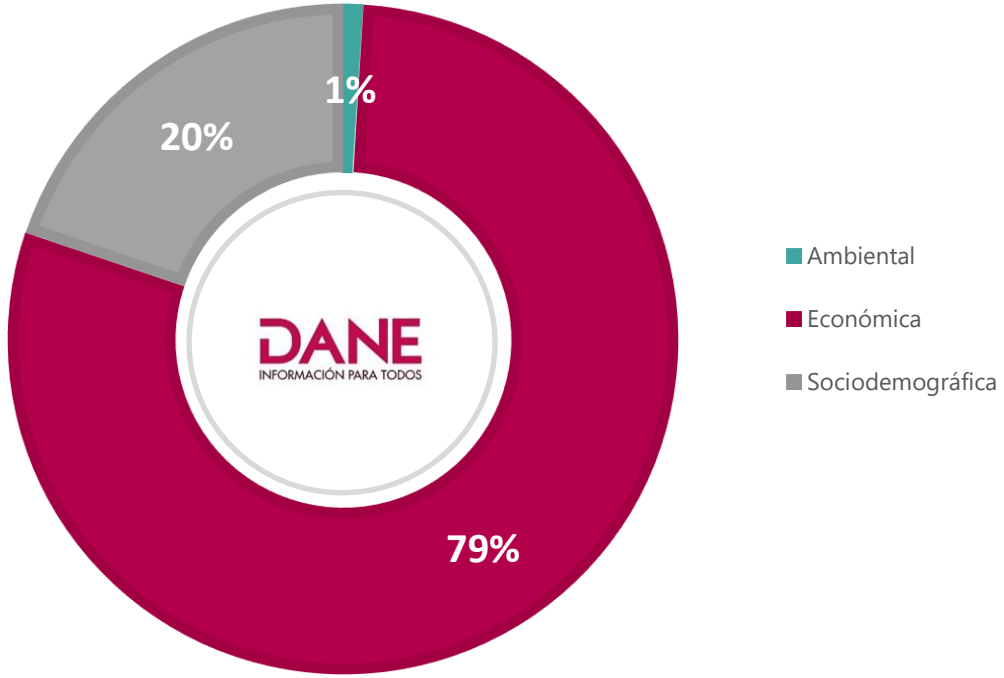
Los datos e información estadística del SEN son activos disponibles para la investigación aplicada y la toma de decisiones.

Operaciones Estadísticas*



El país tiene **401 operaciones estadísticas** producidas por **103 entidades.**

Operaciones Estadísticas DANE *



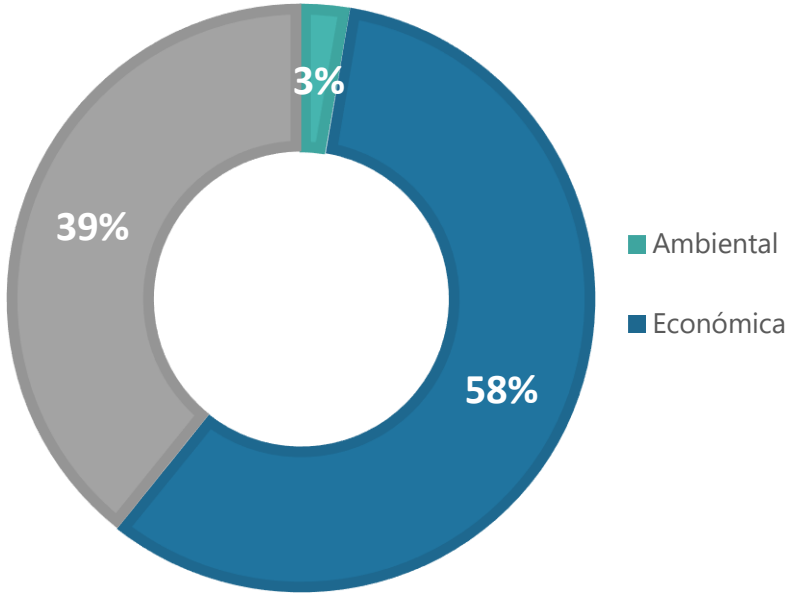
Del total de **operaciones estadísticas**, el DANE produce **106**

* OO.EE activas

Fuentes secundarias disponibles y su uso estadístico

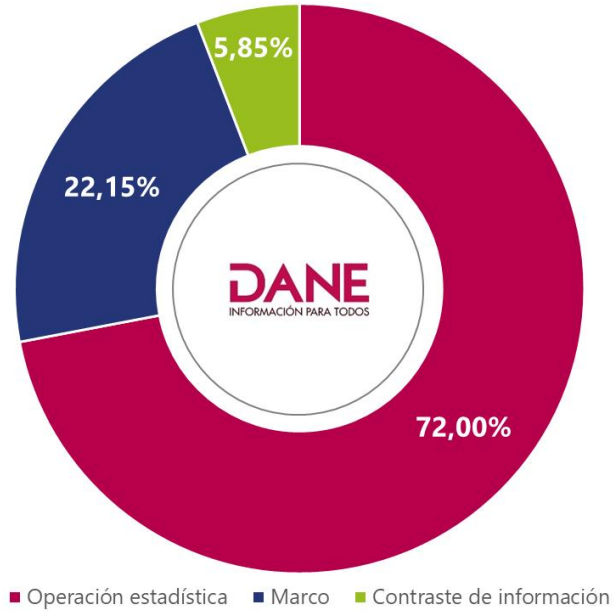
Los datos e información estadística del SEN son activos disponibles para la investigación aplicada y la toma de decisiones.

Participación de los RR.AA. por temática



El país tiene **527 Registros administrativos** producidos por **95 entidades.**

Participación de los principales usos de RR.AA. por el DANE



El DANE aprovecha estadísticamente **233 Registros administrativos.**

* OO.EE y RR.AA en estado activo

Fuente: Inventario de usos de registros administrativos 2022, revisado al 3 de octubre de 2022

De acuerdo con las bases del Plan Nacional de Desarrollo 2022-2026 **“Colombia potencia mundial de la vida”,** se cuenta con los siguientes actores diferenciales para el cambio:

Genero y OS
Mujer
LGBTIQ+

Ciclo de vida
Niños y niñas
Jóvenes
Adulto Mayor

Discapacidad

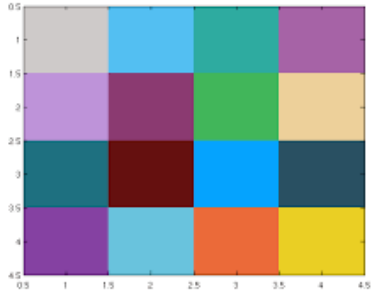
Víctimas Desplazados

Grupos étnicos

Territorio
Municipal y Sub
Ruralidad
PDET

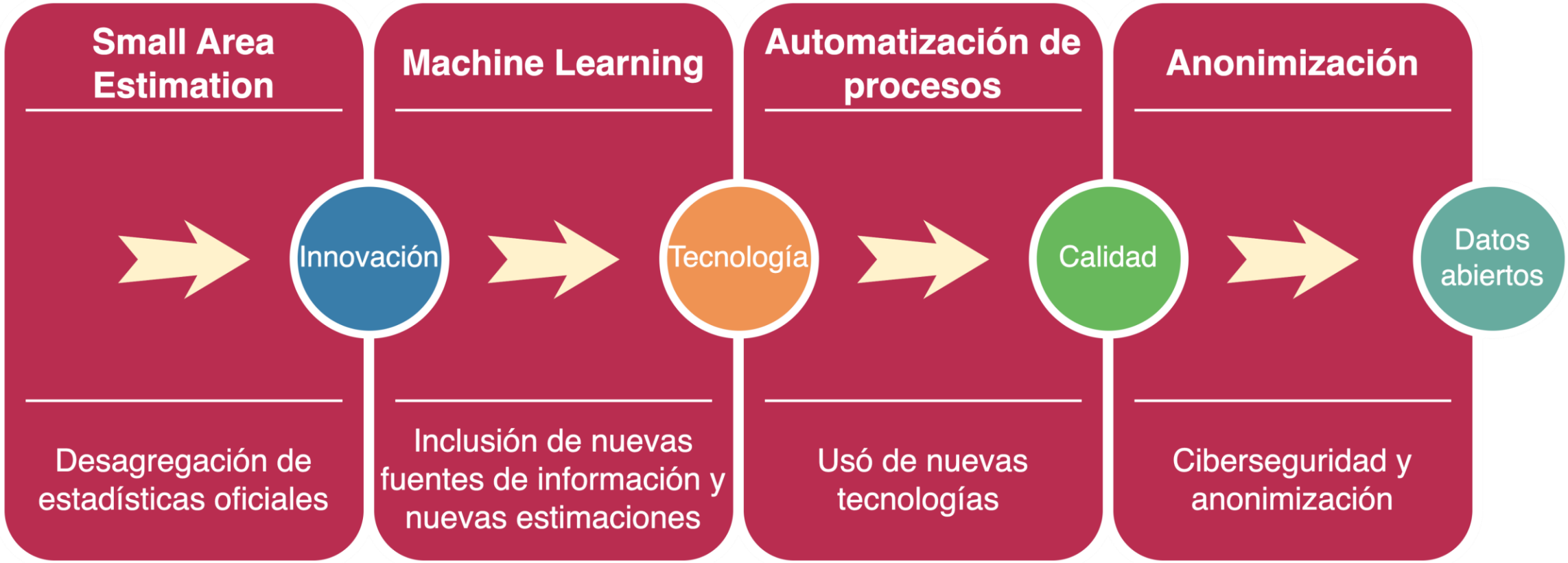
Campesinos

Economía Popular



Interseccionalidad

Methodologies required



Principios Fundamentales de las Estadísticas Oficiales

Adoptados en 1994 por la Comisión de Estadística de las Naciones Unidas y aprobados por la Asamblea General de las Naciones Unidas en 2014.

Pertinencia, Imparcialidad e Igualdad de Acceso

Información al servicio del gobierno, la economía y el público. Estadísticas que respeten el derecho de los ciudadanos a la información pública.

Normas Profesionales y Ética

Normas para mantener la confianza en las estadísticas oficiales en los procedimientos de recogida, tratamiento, almacenamiento y presentación de datos estadísticos.

Responsabilidad y Transparencia

Presentación de la información según normas científicas que faciliten una correcta interpretación de los datos.

Prevención de Mal Uso

Evitar la interpretación errónea y el mal uso de las estadísticas.

Fuentes de Estadísticas Oficiales

Censos, encuestas y registros administrativos.

Principios Fundamentales de las Estadísticas Oficiales

Adoptados en 1994 por la Comisión de Estadística de las Naciones Unidas y aprobados por la Asamblea General de las Naciones Unidas en 2014.

Confidencialidad

Los datos individuales recogidos por el DANE tanto si se refieren a personas naturales como jurídicas deben ser estrictamente confidenciales y utilizarse exclusivamente con fines estadísticos.

Legislación

Las leyes, los reglamentos y las medidas con las que funcionan los sistemas estadísticos deben hacerse públicos.

Coordinación Nacional

La coordinación entre los organismos estadísticos de los países es esencial para lograr coherencia y eficacia.

Uso de Normas Internacionales

Utilización de conceptos, clasificaciones y métodos estandarizados que favorezcan la coherencia.

Cooperación Internacional

La cooperación bilateral y multilateral en materia de estadísticas contribuye a la mejora de los sistemas de estadísticas oficiales en todos los países.

Datos Abiertos

En Colombia, el acceso a la información pública es un derecho consagrado en la Constitución de 1991 (artículos 20 y 74) y los datos abiertos son una de las maneras más efectivas de hacer realidad este derecho (MINTIC, 2021, Guía para el Uso y Aprovechamiento de Datos Abiertos en Colombia).

Este acceso a los datos se enmarca en la Política de Gobierno Digital y contribuye al propósito de tomar decisiones basadas en datos. Los Datos Abiertos son información pública dispuesta en formatos que permiten su uso bajo licencia abierta y sin restricciones legales para su aprovechamiento. En Colombia, la Ley 1712 de 2014 sobre Transparencia y Acceso a la Información Pública Nacional los define como:

“todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos”.

Ley 1266 de 2008 (Habeas Data), Artículo 15 Constitución.

Datos Abiertos

De este modo, la Ley establece la obligatoriedad de las entidades públicas de divulgar datos abiertos teniendo en cuenta las excepciones de acceso a la información asociadas a información clasificada y reservada establecidas en su título tercero, artículos 18 y 19 y en el marco de la seguridad y privacidad de la información.

Ejemplos de datos abiertos en entidades públicas en Colombia se encuentran disponibles en el Portal de Datos del Estado Colombiano www.datos.gov.co (MINTIC):

- Información contenida en el Archivo Nacional de Datos – ANDA del DANE http://www.dane.gov.co/files/anda/index.html
- Casos positivos de Covid 19 dispuestos por el INS.
- Contrataciones públicas dispuestas por la Agencia de Contratación - Colombia Compra Eficiente (a quién le compra el estado, cuánto gastó y en qué gastó).
- Directorio de entidades públicas del orden nacional y territorial dispuestas por el Departamento Administrativo de la Función Pública.
- Los establecimientos educativos activos de preescolar, básica y media a nivel nacional del sector oficial y privado dispuestos por el Ministerio de Educación.
- Los establecimientos a nivel nacional que actualmente son sujetos de actividades de inspección, vigilancia y control por parte del Invima.

Beneficios de los Datos Abiertos

- Transparencia y control social.

Permiten que la sociedad civil, las entidades públicas, la academia y los medios realicen procesos de investigación y control social a sus gobernantes.

- Mejoramiento o creación de productos, servicios y negocios innovadores.

Ayudan a las entidades públicas a mejorar la prestación de servicios a los ciudadanos así como optimizar procesos internos: cuál es el tiempo de espera de un ciudadano para acceder a un servicio y proponer mejoras para optimizarlo; saber qué trámites son los más demandados; entender en qué zonas del país un suceso es más recurrente y adelantar acciones para su monitoreo y control.

- Pronosticar y prevenir fenómenos.

Mapear el estado de un sector y mejorar la cobertura en salud, el estado de vías en una ciudad, asignación de recursos de una entidad, prevención del crimen, pronóstico del tiempo, análisis de enfermedades.

- Generar nuevo conocimiento.

Insumo fundamental donde la academia y los medios pueden generar artículos, estudios de mercado y tesis.

DANE
INFORMACIÓN PARA TODOS

ANDA – Archivo Nacional de Datos

El futuro es de todos
Gobierno de Colombia

Datos Abiertos DANE

Metadatos

Los metadatos estadísticos son una herramienta que brinda a los usuarios de los datos, el contexto necesario de la operación estadística para su correcta interpretación, detallando información relativa a conceptos, al diseño del proceso estadístico, procedimientos de recolección, procesamiento y análisis de los datos, documentos relacionados entre otros aspectos.

Microdatos

Los microdatos constituyen el menor nivel de desagregación de los datos en una operación estadística. Utilizando software estadístico los usuarios pueden procesar sus propios tabulados; sin embargo, se advierte que deben considerar el diseño muestral correspondiente y evaluar los errores muestrales para su inferencia estadística.

Página web: https://www.dane.gov.co/

Pestaña: Estadísticas por Tema o Servicios al Ciudadano.

ANDA – Archivo Nacional de Datos

Es un instrumento realizado por el DANE con el apoyo del Banco Mundial que centraliza los microdatos y metadatos de las operaciones estadísticas de las entidades productoras en Colombia a través de herramientas de búsqueda virtual.

Es pionero en la promoción del uso de datos abiertos en los procesos de captura, procesamiento y divulgación. Mediante un catálogo los usuarios pueden buscar, comparar, solicitar acceso y descargar información relacionada con censos, encuestas por muestreo y registros administrativos.

Son beneficiarios del ANDA:

a) Los productores de información estadística que difunden lo que producen y asimilan cómo debe ser la correcta captura, producción y difusión de cifras

b) Los ciudadanos (académicos, gobiernos, estudiantes, empresas) que usan los datos abiertos como instrumento para la investigación y la toma de decisiones.

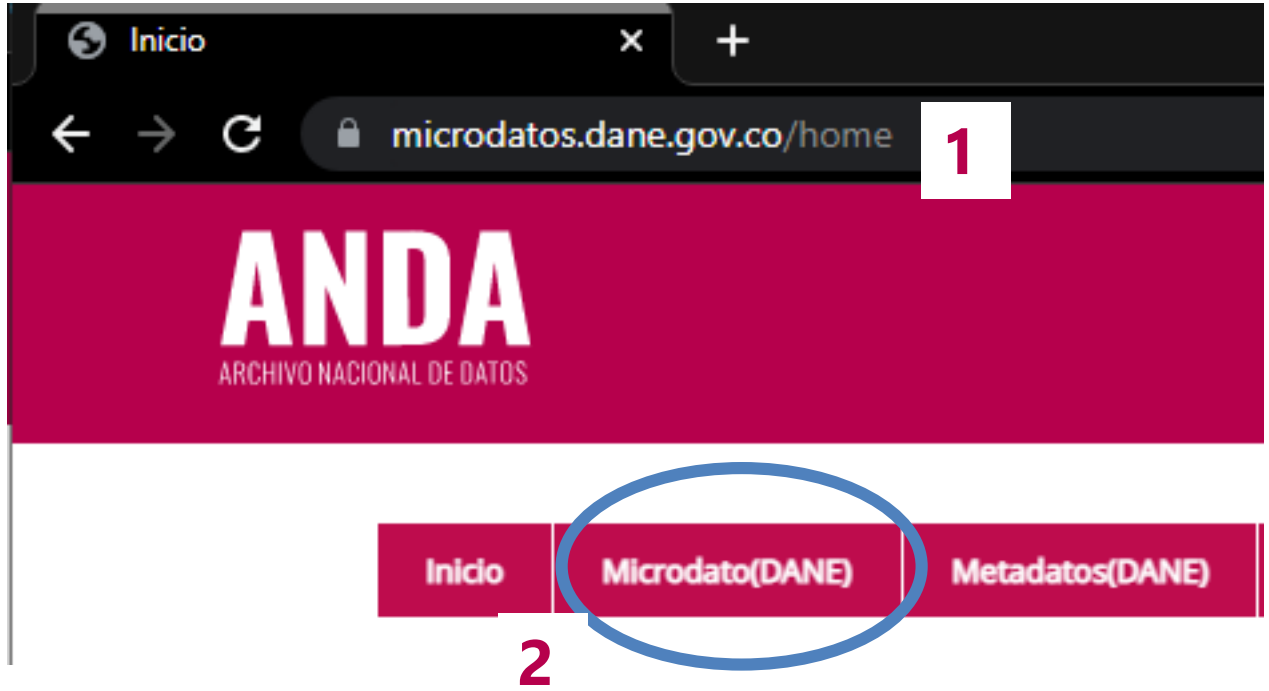
Constituye un repositorio de datos de cifras confiables y de calidad.

1. **Acceda al catalogo de datos del DANE - ANDA**

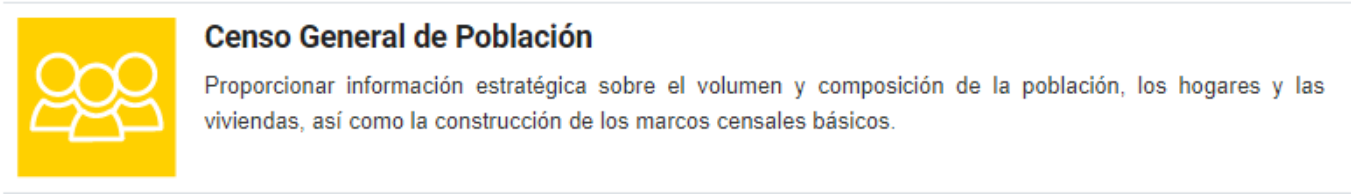
https://microdatos.dane.gov.co/home

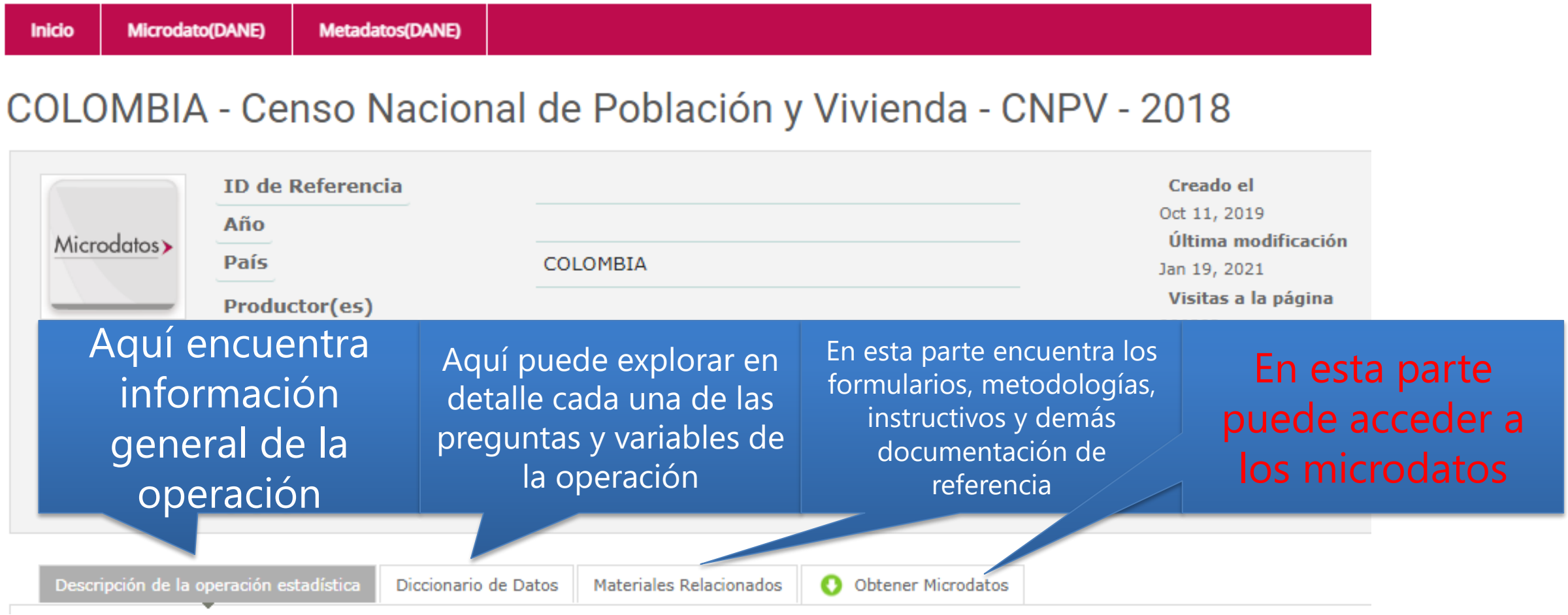
2. **Ingrese a Microdatos(DANE)**

3. **Ubique la operación estadísticas**



3




Inicio
Microdato(DANE)
Metadatos(DANE)
COLOMBIA - Censo Nacional de Población y Vivienda - CNPV - 2018
Microdatos
ID de Referencia
Año
País
COLOMBIA
Productor(es)
Creado el
Oct 11, 2019
Última modificación
Jan 19, 2021
Visitas a la página
Aquí encuentra información general de la operación
Aquí puede explorar en detalle cada una de las preguntas y variables de la operación
En esta parte encuentra los formularios, metodologías, instructivos y demás documentación de referencia
En esta parte puede acceder a los microdatos
Descripción de la operación estadística
Diccionario de Datos
Materiales Relacionados
Obtener Microdatos

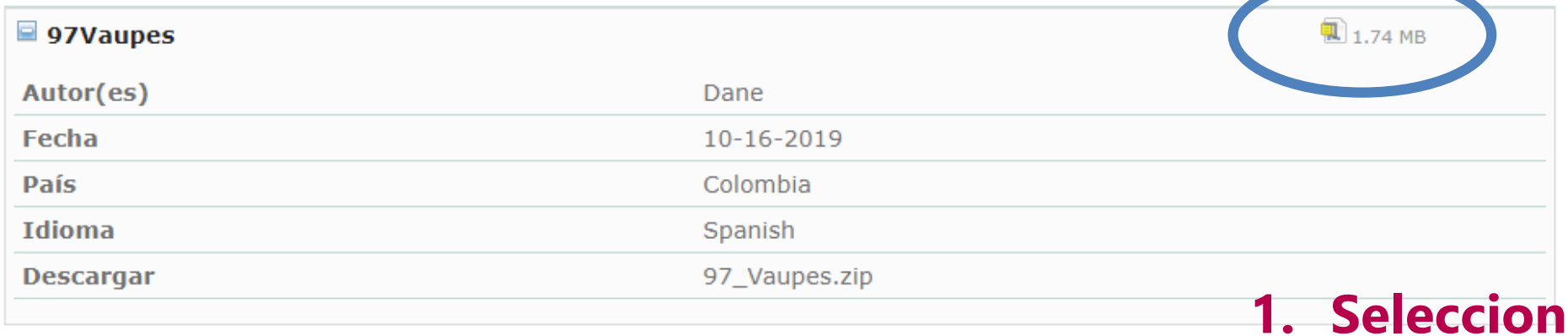
Documentación Técnica

Manual de diligenciamiento	1.53 MB
Manual de conceptos	339 KB
Diccionario de Datos CNPV2018	81.21 KB
Instructivo uso del Marco Geoestadístico Nacional (MGN) integrado a la información del CNPV 2018	577.17 KB
Instructivo de integracion de información estadistica del CNPV2018 con el MGN	2.13 MB
Ficha_metodologica_CNPV_2018.	694.38 KB
Metodologia_CNPV_2018	7.59 MB

En este documento encuentra información detallada sobre el tratamiento realizado en el CNPV a las variables del Marco Geoestadístico

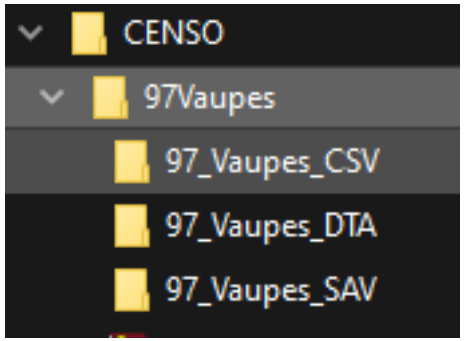
En este documento encuentra información detallada de cómo integrar el marco geoestadístico en las bases del CNPV

1

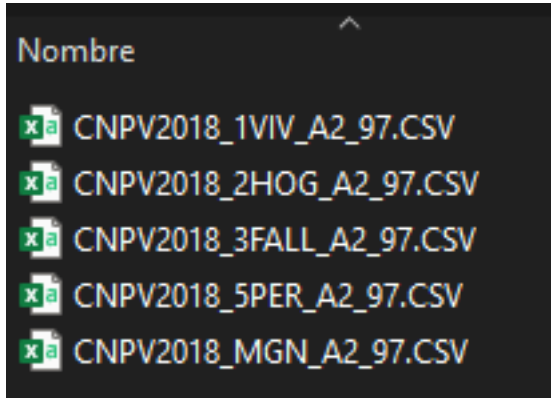


2

3



4



1. **Seleccione el departamento**
2. **Click en el archivo**
3. **Validación del sistema**
4. **Descomprima los archivos en su PC.**

Canales de difusión del CNPV2018

Los datos del CNPV 2018 se encuentran disponibles en los siguientes canales:

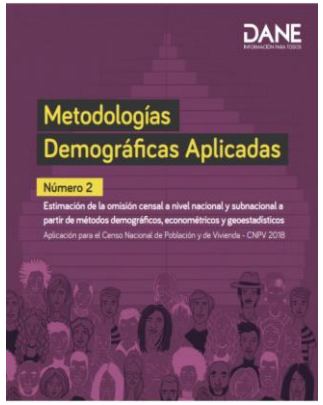
CANAL	ESTADO ACTUAL
ANDA	Publicación de datos abiertos a partir de bases anonimizadas a nivel de manzanas, con supresión y recategorización de variables.
REDATAM	Implementado, a partir de la base y lineamientos del ANDA
GEO PORTAL	Geovisores implementados
IPUMS	Por implementar
SALA DE PROCESAMIENTO	Base de datos anonimizada viviendas étnicas con coordenadas x,y. Acceso bajo licencia de uso y acuerdos de confidencialidad.

Google Data Search, Kaggle

Productos estadísticos

Informes de Estadística Sociodemográfica Aplicada

Consolidar información estadística de seis temáticas principales: Estudios urbanos, Étnicos, Migración, Mortalidad, Fecundidad y Población para diferentes periodos de referencia a partir de diversas fuentes de información como parte del fortalecimiento de la producción y difusión de información estadística para todos.

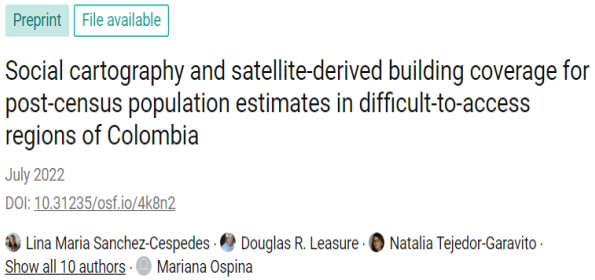


Metodologías Demográficas Aplicadas

Consolidar metodologías experimentales a partir de diversas fuentes de información como parte del fortalecimiento de la producción y difusión de información estadística para todos.

Notas Estadísticas

Incentivar los análisis y la toma de decisiones con base en los datos disponibles en temas que se consideran de interés.
Presentar en un solo documento, información proveniente de diferentes operaciones estadísticas del DANE para informar sobre un fenómeno de interés.
Análisis cuantitativos especializados en temas de la Agenda 2030.


Preprint File available

Social cartography and satellite-derived building coverage for post-census population estimates in difficult-to-access regions of Colombia

July 2022

DOI: 10.31235/osf.io/4k8n2

Lina Maria Sanchez-Cespedes · Douglas R. Leasure · Natalia Tejedor-Garavito · Show all 10 authors · Mariana Ospina

Social cartography and satellite-derived building coverage for post-census population estimates in difficult-to-access regions of Colombia

Desarrollamos modelos bayesianos jerárquicos para estimar el número de edificios y el tamaño de la población, los entrenamos usando enumeraciones de censos de cobertura total cercanos y evaluamos usando una validación cruzada de 10 veces.
Este artículo la fue aceptado para publicación en la Revista *Population Studies*

Otras Herramientas – Servicio al Ciudadano

- Microdatos - ANDA
- Atlas Estadístico (Geovisores) - Geoportal
- Biblioteca Virtual – Memoria histórica de la entidad / Biblioteca Documentación
- SIPSA
- Estratificación Socioeconómica para Servicios Públicos
- DIVIPOLA – División Político Administrativa
- Notas Estadísticas – Estadísticas Experimentales
- Historia del Dato
- Objetivos de Desarrollo Sostenible – ODS.

Generación de estadísticas experimentales

Una **estadística experimental** es aquella que se deriva de proyectos en desarrollo que cuentan con aspectos innovadores, ya sea por **aprovechamiento de nuevas fuentes de información, la metodología estadística utilizada o una temática nueva no medida anteriormente.**

las estadísticas experimentales son estadísticas oficiales por el decreto 2404 del 2019, y ofrecen nuevas formas de caracterizar cuantitativamente fenómenos en las tres dimensiones del **Desarrollo Sostenible: económica, social-demográfica y ambiental** del país y además contribuyen a:

1. **Mejorar la disponibilidad de estadísticas relevantes y oportunas con los niveles requeridos de desagregación** (por grupos de población o por dominios geográficos). Esto es clave para la producción y difusión de información estadística con enfoque territorial, diferencial e interseccional; requerida para la caracterización de grupos vulnerables.
2. **Incluir sectores donde se hayan identificado vacíos de información estadística,** tomando en cuenta la experiencia de los proyectos de estadísticas experimentales.
3. **Integrar fuentes tradicionales de información,** como censos y encuestas, con fuentes alternativas como imágenes de satélite o registros administrativos para la generación de nuevos indicadores. En este caso, la atención se centra en los fenómenos que se están investigando más que en las fuentes estadísticas utilizadas para describirlos.
4. **Generar capacidades en el personal de la entidad** para evitar la continua dependencia de agentes externos para **adelantar proyectos sobre nuevas fuentes.**

4. ESTABLECER CENTRO DE DATOS PARA CONSULTA DE MICRODATOS en SPEEE (Sala de Procesamiento Especializado Externo)

Sala de Procesamiento Especializado Externo ¿Qué es?

Para consultar y procesar microdatos anonimizados de distintas operaciones estadísticas, el **Departamento Administrativo Nacional de Estadística (DANE)** cuenta con la Sala de Procesamiento Especializado Externo (SPEE).

Fuente: DANE.

La Sala de Procesamiento Especializado Externo (SPEE) es un servicio que dispone el DANE para que investigadores especializados, nacionales o extranjeros, accedan y procesen los microdatos anonimizados de las operaciones estadísticas del DANE, no disponibles en el sitio web.

INFORMACIÓN PARA TODOS

4. ESTABLECER CENTRO DE DATOS PARA CONSULTA DE MICRODATOS en SPEEE (Sala de Procesamiento Especializado Externo)

Reglas de uso de la SPEE

El horario de atención es de lunes a viernes de 8:00 a.m. a 5:00 p.m.

Firmar el formato de asistencia al iniciar y terminar la visita.

Diligenciar las actas de compromiso de confidencialidad y de entrega de resultados cada vez que se requiera.

Hacer uso de los casilleros para guardar objetos personales. (maletas, libros, carpetas, etc)

Mantener el puesto de trabajo organizado

No es permitido más de dos investigadores en la misma terminal de manera simultánea

El uso de dispositivos móviles (celulares, tablets, portátiles, etc.) no es permitido al interior de la sala

No consumir alimentos o bebidas dentro de la sala

Fuente: DANE.

INFORMACIÓN PARA TODOS

4. ESTABLECER CENTRO DE DATOS PARA CONSULTA DE MICRODATOS en SPEEE (Sala de Procesamiento Especializado Externo)

¿Qué son los Centros de Datos?

El centro de datos funciona como un punto DANE desde la institución universitaria que firma el acuerdo de entendimiento para la implementación de este sitio.

Los Centros de datos en universidades deben cumplir los requerimientos técnicos, físicos y administrativos para preservar la Ley de reserva estadística consagrada en la Ley 79 de 1993.

Fuente: DANE.

Aprovechamiento de las estadísticas oficiales para la investigación

Información pública y disponible para uso en investigación

DANE
INFORMACIÓN PARA TODOS

ANDA
ARCHIVO NACIONAL DE DATOS

- **458** estudios estadísticos, **330** con microdatos,
- **128** con metadatos; esto representado en **120**
- operaciones estadísticas del DANE.

Geoportal

- **38** Geovisores
- **54** Visores de mapas interactivos
- Servicios de descarga de datos geoestadísticos

Información disponible para uso ambientes controlados

Centros de Datos
8 Centros en universidades nacionales e internacionales

- U DEL ROSARIO
- U LOS ANDES
- U EAFIT MEDELLÍN
- U CARLOS III MADRID
- U DE SANTANDER
- U INSPER - BRASIL
- U BERKELEY-CALIFORNIA E.U.
- U SAN BUENAVENTURA

21 operaciones estadísticas con microdatos anonimizados disponibles

Flujo del tramite y aprobación de una solicitud
Artículos 4 al 10 de la resolución 2251 de 2019

- ## Condiciones de acceso - continuación

➢ En los casos en que el procesamiento requerido supere las capacidades actuales del área técnica, podrá plantearse la suscripción de convenios de cooperación técnica con recursos suficientes para efectuar los procesamientos solicitados o requeridos.

➢ Las salidas de información autorizadas a través de este mecanismo podrán realizarse en resúmenes numéricos que no permitan la identificación de las fuentes o proveedores de datos.

Acceso a microdatos anonimizados bajo licencia:

El acceso a los microdatos anonimizados que no son cubiertos por los mecanismos previstos en los casos mencionados anteriormente deberán ser autorizado por el Subcomité de Aseguramiento de la Reserva Estadística, cumpliéndose como mínimo las siguientes condiciones:

Subcomité de Aseguramiento de Reserva Estadística

El Subcomité de Aseguramiento de la Reserva Estadística, tiene el fin de garantizar el cumplimiento de la reserva estadística (artículo 5º de la Ley 79 de 1993), frente a la disposición de los microdatos anonimizados y los mecanismos de acceso a los mismos, promoviendo el mayor acceso y uso de la información estadística por parte de todos los usuarios.

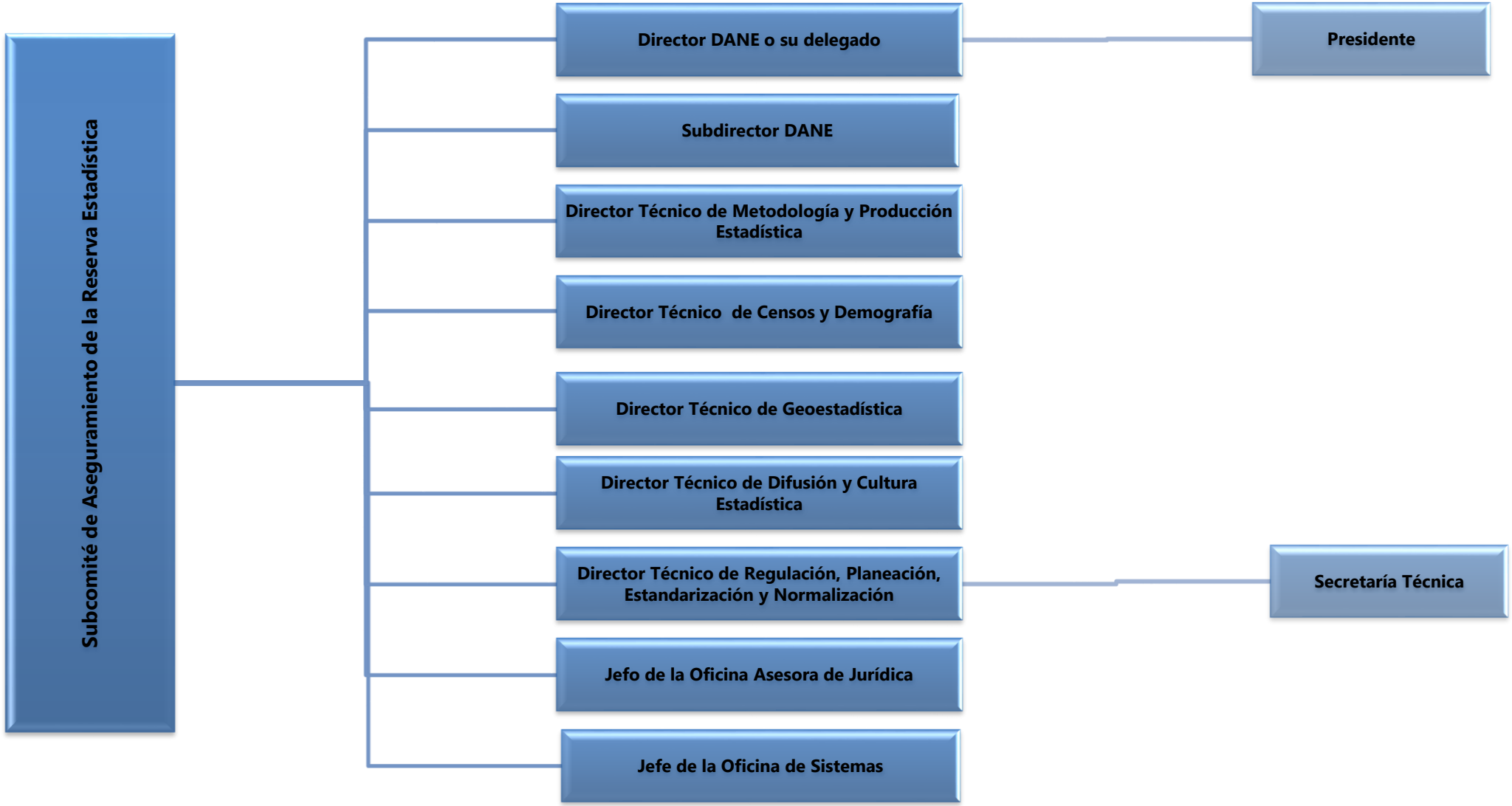
Antecedentes

En 2018, se encontraba vigente la Resolución 1503 del 16 de noviembre de 2011, la cual reguló la conformación del Comité de Aseguramiento de la Reserva Estadística del DANE. Así como amplió los conceptos y precisó definiciones, con el fin adaptarlos a las normas y patrones internacionales en materia de seguridad y manejo de la información estadística.

La anterior resolución fue modificada, por la **Resolución 2251 de 2019**, por la cual se dictaron y regularon los procesos necesarios para la gestión de las solicitudes de acceso a la información estadística, así como creo y reglamento el funcionamiento el Subcomité de Aseguramiento de la Reserva Estadística.

El artículo tercero del decreto 458 de 2020, reguló el funcionamiento de la reserva estadística en el marco del Estado de Emergencia Económica, Social y Ecológica.

2. Conformación del Subcomité de Aseguramiento de la Reserva Estadística



3. Flujo del tramite y aprobación de una solicitud

Artículos 4 al 10 de la resolución 2251 de 2019

1. Las solicitudes para acceder a datos e información estadística producida por la entidad, pueden ser recibidas a través de :
   - ✓ Correo institucional contacto@dane.gov.co
   - ✓ Página web o
   - ✓ A través de cualquier funcionario del DANE, y

2. Las solicitudes deben ser radicadas a través del sistema de correspondencia establecido por la entidad para preservar su trazabilidad.

3. Las solicitudes deben contener la siguiente información:
   - ✓ Solicitante: persona natural o jurídica que suscribe la solicitud.
   - ✓ Operación estadística involucrada
   - ✓ Descripción o detalle de la solicitud, como por ejemplo objetivo, uso, justificación, etc.
   - ✓ Mecanismo de acceso *a través del cual se pude gestionar* la solicitud.

DANE
INFORMACIÓN PARA TODOS

Anonimización

El futuro es de todos
Gobierno de Colombia

ANDA – Archivo Nacional de Datos

Ley 79 de 1993 Artículo 5:

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

Solución a la tensión entre la demanda de información (artículos 20 y 74 de la Constitución, ley 1712 de 2014) y la restricción a la oferta - habeas data (artículo 15 de la misma Constitución, ley 1266 de 2008)?

ANONIMIZACION

Anonimización de Datos

La anonimización busca proteger los datos estadísticos de manera que puedan ser difundidos sin afectar la confidencialidad de la información que puede ser relacionada a individuos o entidades en particular.

Datos tabulares: Publicar información estática agregada sin que información confidencial de individuos específicos pueda ser inferida.

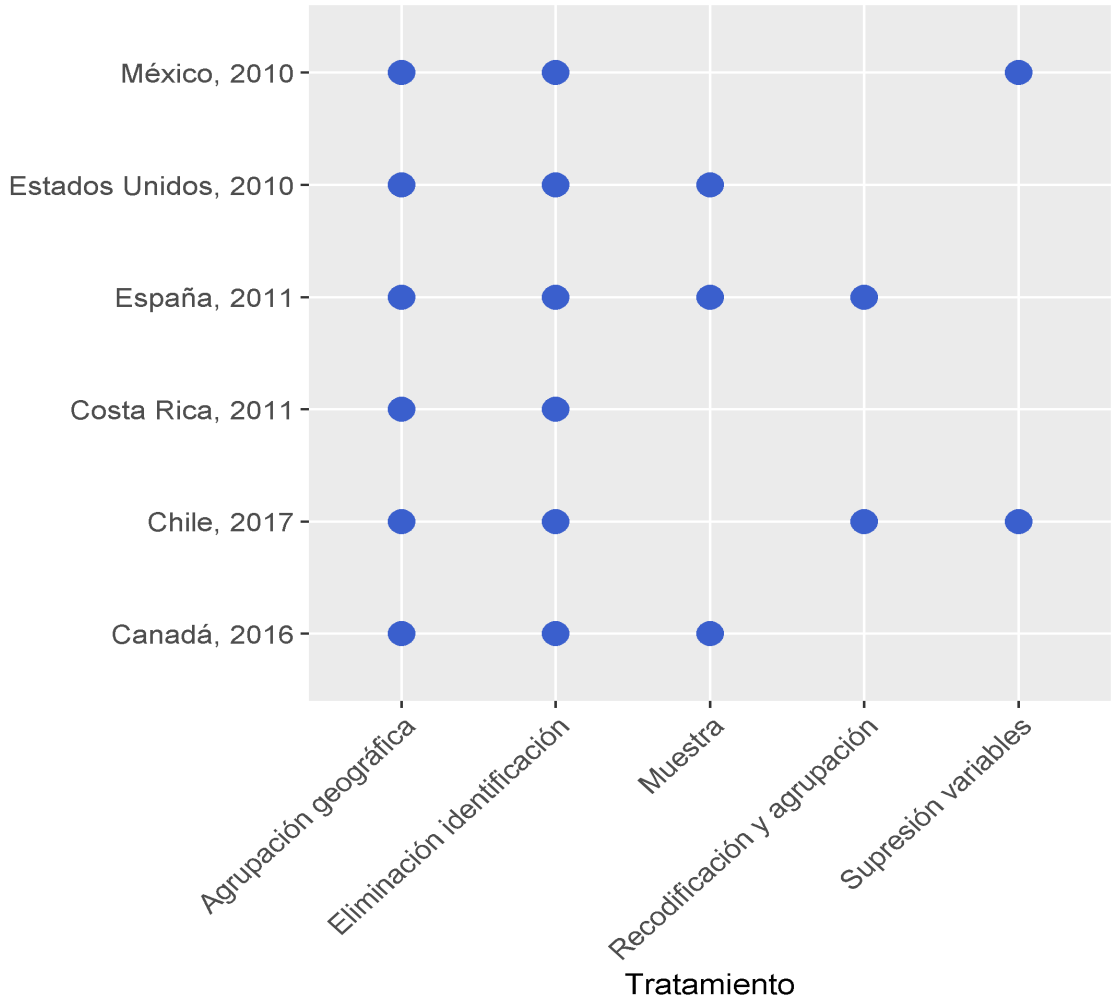
Microdatos: Conjuntos de datos que contienen las variables de interés para cada respondiente deben surtir algún proceso de anonimización para no interferir con la confidencialidad de una fila de la base de datos. Riesgos al cruzar información en bases anonimizadas.

Salidas de análisis estadísticos: Revisión de resultados de análisis que no infrinjan la confidencialidad de los individuos.

Dos técnicas: a) Perturbativas – Perturban los datos antes de ser publicados introduciendo a propósito un error aleatorio para preservar la confidencialidad. b) No perturbativas – reducen la cantidad de información por supresión o agregación.

Referentes internacionales en anonimización de censos.

La mayoría de Institutos de Estadística difunde datos agregados al público en general y microdatos solo a público especializado.



Anonimización de Datos

La Comisión Económica para Europa de las Naciones Unidas (UNECE) publicó el Manual sobre el control de divulgación estadística y el programa de anonimización ARGUS para implementar mecanismos de anonimización en oficinas de estadística nacional.

La Oficina de Censos de USA utiliza diferentes técnicas como eliminar variables de identificación directa, umbrales geográficos o categóricos de las variables, redondeo, infusión de ruido, recodificación, intercambio de registros basado en rangos o proximidad.

La CBS - Holanda ha desarrollado el software µ-ARGUS para anonimización de microdatos y el software t-ARGUS para anonimización de datos tabulares.

El Decreto 1743 de 2016 define la anonimización de microdatos en Colombia como "transformar los datos individuales de las unidades de observación, de tal modo que no sea posible identificar sujetos o características individuales de la fuente de información, preservando así las propiedades estadísticas en los resultados".

Proceso de anonimización

Entrada:
Base de datos

Etapa 1:
Revisiones previas

Etapa 2:
Análisis de riesgo

Etapa 3:
Selección de técnicas

Etapa 4:
Análisis de viabilidad

Etapa 5:
Aplicación de técnicas

Etapa 6:
Evaluación de resultados

Salida:
Base de datos anonimizada

Fuente: Guía para la anonimización de bases de datos en el Sistema Estadístico Nacional.

Anonimización de bases de datos

Las bases del CNPV 2018 fueron publicadas a nivel de microdato en octubre de 2019, en su anonimización fueron aplicadas dos técnicas:

- Supresión de variables. Fueron eliminadas de la base los identificadores directos y las variables consideradas sensibles a partir de criterios temáticos.
- Agregación de categorías dentro de variables. Fueron tratadas las variables: edad, nivel educativo y parentesco con el jefe de hogar.

Medición del riesgo de re-identificación

El riesgo de re-identificación fue evaluado a partir de la cuantificación de las personas únicas perfiladas a partir de sus características presentes en las bases de datos.

Fueron evaluados varios escenarios de re-identificación en donde se considera el efecto que tiene el número de variables de perfilamiento disponibles y el nivel geográfico al que se puede tener acceso en la base de datos.

También se consideró una medida de riesgo asociada con el riesgo conjunto presente por el hogar (personas ampliadas).

Disclosure Control – No Perturbativas

Eliminación de Variables
Se suprime toda la información de una variable cuando ésta contiene información de identificación directa de la unidad de observación (Cédula, Tipo de Identificación, Nombre, Apellidos, Dirección, Barrio, Municipio, Fecha de Nacimiento, RH).

Recodificación Global
Combinar diversas categorías de las variables categóricas en una más general que tenga mayor frecuencia y menor información. En variables continuas, agrupar por medio de intervalos (Grupo Étnico, Número de Habitaciones, Número de Viajes, Nivel de Escolaridad).

Codificación Superior e Inferior
Proteger la identificación de las unidades de observación que presentan los valores más altos o más bajos de cada variable.

Supresión Local
Reemplazar los valores de una o más variables de las unidades en riesgo por valores faltantes.

Disclosure Control – Perturbativas

Microagregación
Reemplazar los valores de algunas unidades en riesgo por el valor promedio (Edad, ingresos anuales, ingresos mensuales, número de hijos, número de personas en el hogar).

Redondeo
Sustituir los valores de las unidades en riesgo con decimales por enteros.

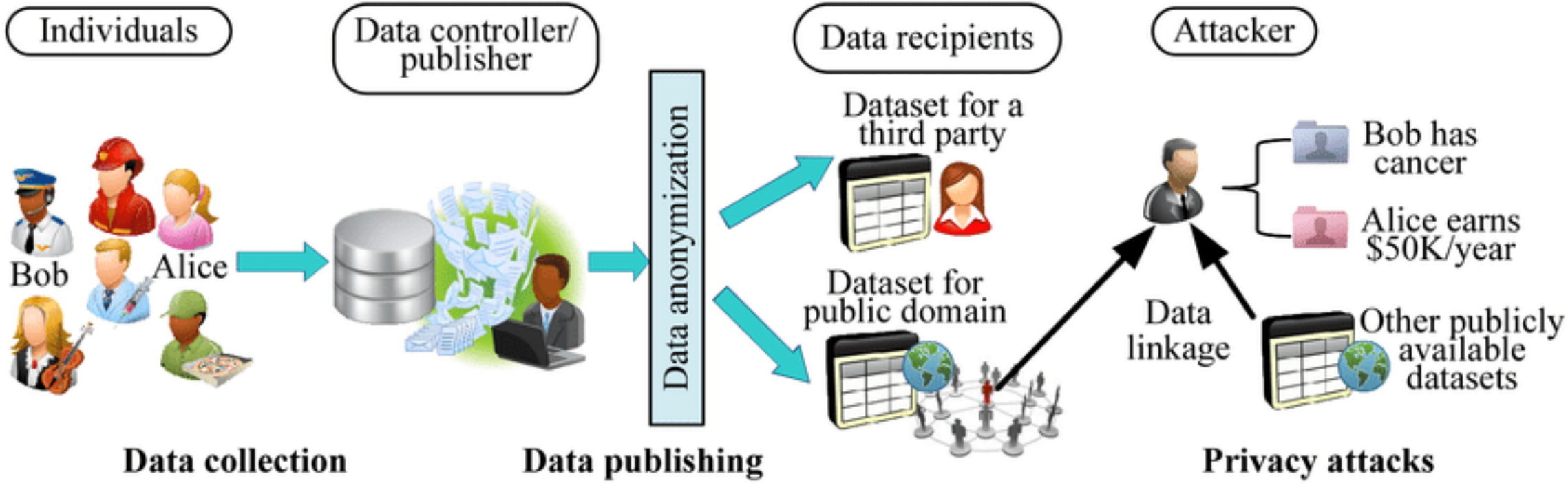
Intercambio de Datos
Intercambiar la información de unidades en riesgo con unidades que no tienen riesgo de manera aleatoria entre pares de observaciones.

Agregar Ruido
Añadir una cantidad aleatoria sobre los valores de las unidades en riesgo.

Datos sintéticos
Se produce una nueva base de datos mediante el uso de algoritmos de simulación que conserve las propiedades de la base original (regresión cuantílica, imputación adicional y datos combinados).

Privacy Enhancing Technologies - PET

Información obtenida a través de censos, encuestas y registros pero también a través de técnicas de ciencia de datos (webscrapping, imágenes satelitales, redes sociales) puede estar sujeta a riesgos de privacidad aunque se produzca con mayor granularidad y mayor oportunidad.



Privacy Enhancing Technologies - PET

Información obtenida a través de censos, encuestas y registros pero también a través de técnicas de ciencia de datos (webscrapping, imágenes satelitales, redes sociales) puede estar sujeta a riesgos de privacidad aunque se produzca con mayor granularidad y mayor oportunidad.

Combina datos sensible de múltiples fuentes de información.

Garantiza que los datos sólo sean usados para su propósito inicial.

Difunde información con la confianza que los datos originales (microdatos) no pueden ser reversados a través de algoritmos.

Privacidad de Entrada y de Salida

https://courses.openmined.org/courses

Input Privacy (la seguridad de lo que entra, hackers)

Secure Multiparty Computation (MPC)
Homomorphic Encryption (FHE)
Trusted Execution Environments (TEE)
Zero Knowledge Proofs (ZK Proofs)

Enclaves seguros

Output Privacy (la seguridad de lo que sale, ataques de reidentificación)

Disclosure Control. Aggregation. K-anonimity.
Privacidad Diferencial. Privacy loss parameter (épsilon). Noise mechanisms. No se puede reversar la información para deducir el valor de un individuo. Con perturbaciones a los datos se le da a cada individuo la misma privacidad como si hubiera sido borrado de la base.
Federated Learning. Técnica de ML que entrena un algoritmo a través de varios servidores o dispositivos que conservan muestras de la base de datos. Flower.

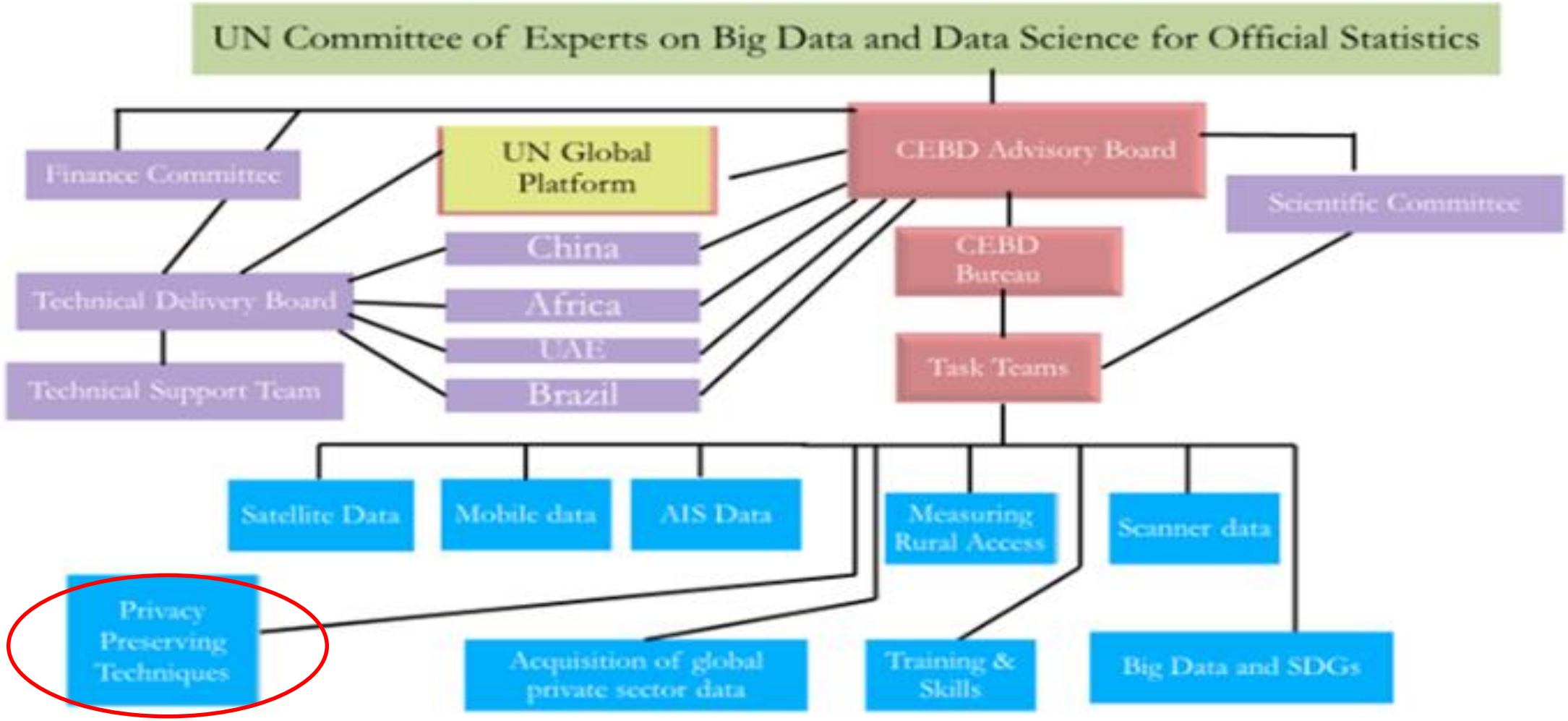
UN Big Data

COMITÉ DE EXPERTOS EN BIG DATA

Estados Miembros

Australia	Egypt	Morocco	Saudi Arabia
Bangladesh	Georgia	Netherlands	South Africa
Brazil	Germany	Oman	Switzerland
Cameroon	India	Pakistan	United Arab Emirates
Canada	Indonesia	Philippines	United Kingdom
China	Ireland	Poland	Tanzania
Colombia	Italy	Republic of Korea	United States of America
Denmark	Mexico	Rwanda	

COMITÉ DE EXPERTOS EN BIG DATA



Evolución Institucional y Logros

Interoperabilidad de datos con otras instituciones

Además del trabajo sobre las operaciones estadísticas bajo la responsabilidad del DANE, la entidad adelante un gran proceso de transformación en términos de acceso, evaluación, uso e interoperabilidad con los registros administrativos. Entre otros aspectos, este esfuerzo consiste en:

- **Investigar** los datos administrativos para identificar errores, incertidumbres y posibles sesgos en los datos
- Hacer esfuerzos para **comprender** por qué ocurren estos errores y gestionarlos
- **Comunicar** a los usuarios cómo estos podrían afectar las estadísticas y su uso

Consideración final

Los datos abiertos son esencialmente una construcción colectiva en la que participamos todos: cuando aceptamos participar de las encuestas y respondemos con veracidad, cuando nos enfrentamos a su uso y encontramos necesario plantear preguntas sobre su obtención, su coherencia y alcance, y cuando desde la posición que tengamos, como investigadores, planeadores, activistas, estudiantes, políticos, técnicos o profesional de cualquier área, hacemos un uso responsable que trascienda al beneficio de todos (DANE & UNFPA, Julio 26 de 2023).

COLOMBIA POTENCIA DE LA VIDA

DANE INFORMACIÓN PARA TODOS 70 AÑOS

Avances y Desafíos Metodológicos

Producción, Administración, Interoperabilidad de Datos Abiertos

Leonardo Trujillo
Subdirector DANE