LINEAMIENTOS PARA LA ANONIMIZACIÓN DE DATOS DEL SISTEMA NACIONAL DE ESTUDIOS Y ENCUESTAS POBLACIONALES PARA LA SALUD

MINISTERIO DE SALUD Y PROTECCIÓN SOCIAL

DIRECCIÓN DE EPIDEMIOLOGÍA Y DEMOGRAFÍA

Grupo gestión del Conocimiento y Fuentes de Información
Sistema Nacional de Estudios Encuestas poblacionales para la salud

Contenido

Índice de Figuras

En concordancia con el Código Nacional de Buenas Prácticas para las Estadísticas Oficiales (CNBP) de El Departamento Administrativo Nacional de Estadística (DANE), como ente rector, coordinador y regulador del Sistema Estadístico Nacional (SEN), quien promueve el mejoramiento de la calidad y la credibilidad de las estadísticas en nuestro país, el Ministerio de Salud y Protección Social incorpora las directrices dadas en el documento de lineamientos para la anonimización, con el ánimo de fortalecer la calidad de los procesos y de las estadísticas oficiales generadas[11].

Introducción

Paralelo al crecimiento de los volúmenes de información que se generan desde las diferentes áreas del conocimiento crece también el interés por la utilización de los mismos. Académicos, investigadores, estudiantes, organizaciones interesadas en ejercer control social y la comunidad en general se podrían beneficiar si dispusieran de datos de libre acceso; sin embargo garantizar que sean respetados los derechos de todas las personas cuya información reposa en las bases de datos, que se les garantice la protección de sus datos personales y su vida privada es una prioridad en este proceso.

La aplicación adecuada de técnicas de anonimización puede aportar en este propósito; por un lado se garantizaría la privacidad y por otro permitiría conservar con un buen grado de fidelidad la información subyacente según la necesidad de los potenciales usuarios. Es importante tener en cuenta que esta no es una tarea sencilla por cuanto ninguna técnica es infalible y por otro lado un conjunto de datos anonimizado puede conllevar problemas residuales inherentes al mismo proceso de anonimización, lo que podría provocar su total inutilización o simplemente no cumplir a cabalidad con el propósito del usuario.

Glosario

Anonimización: es el proceso que impide la identificación de las unidades de estudio que son fuente para los registros individuales del conjunto de microdatos.[1]

Microdato: son los datos sobre las características de las unidades de una población, (individuos, hogares, establecimientos, entre otros), que constituyen una unidad de información en una base de datos y que son recogidos por medio de una operación estadística.[1]

Datos personales: Es toda información numérica, alfabética, gráfica, fotográfica, acústica o de cualquier otro tipo, susceptible de recogida, registro, tratamiento y transmisión, concerniente a personas físicas identificadas o identificables (tales como nombre, apellidos, estado civil, sexo, edad, domicilio, número de la seguridad social, número de matrícula del empleado, identificación personal, número de teléfono, etc.)[1]

Información sensible: es la información considerada como estrictamente confidencial. Información y características referentes a la edad, procedencia, salud, raza, religión, ideología, afiliación, finanzas, etc., se consideran de carácter sensible y requieren de una protección especial.[1]

Tratamiento de los datos personales: Hace referencia a cualquier operación o procedimiento técnico, sea o no automatizado, que permita la recogida, grabación, conservación, elaboración, modificación, consulta, utilización, cancelación, bloqueo o supresión, así como las cesiones de datos que resulten de comunicaciones, consultas, interconexiones y transferencias.[2]

Legislación

Legislación Nacional

En Colombia la legislación vigente referente a la protección de la confidencialidad se encuentra consagrada en la Constitución Política, en su Artículo 15, donde se establece la primera directriz materia de derecho a la intimidad, protección de datos, entre otros:

> *Todas las personas tienen derecho a su intimidad personal y familiar y a su buen nombre, y el Estado debe respetarlos y hacerlos respetar. De igual modo, tienen derecho a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en los bancos de datos y en archivos de entidades públicas y privadas. En la recolección, tratamiento y circulación de datos se respetarán la libertad y demás garantías consagradas en la Constitución. La correspondencia y demás formas de comunicación privada son inviolables.*[7]

Este derecho constitucional fue desarrollado por la Ley Estatutaria 1266 de 2008 por la cual se dictan las disposiciones generales el habeas data y se regula el manejo de la información contenida en bases de datos personales, en especial la financiera,

crediticia, comercial, de servicios y la proveniente de terceros países y se dictan otras disposiciones con la recolección, tratamiento y circulación de datos personales.[8]

En 2012, la Ley 1581, por la cual se dictan disposiciones generales para la protección de datos personales, parcialmente reglamentada por la Ley 1377 de 2013, dispone en sus principios referentes al acceso y circulación restringida de seguridad y de confiden-cialidad, que el acceso a los datos se debe restringir y la información debe estar sujeta a tratamiento por parte del responsable, como lo indica en su Artículo 4:

> *f)Principio de acceso y circulación restringida: el Tratamiento se sujeta a los límites que se derivan de la naturaleza de los datos personales, de las disposiciones de la presente ley y la Constitución. En este sentido, el Tratamiento sólo podrá hacerse por personas autorizadas por el Titular y/o por las personas previstas en la presente ley.*
>
> *Los datos personales, salvo la información pública, no podrán estar disponibles en Internet u otros medios de divulgación o comunicación masiva, salvo que el acceso sea técnicamente controlable para brindar un conocimiento restringido sólo a los Titulares o terceros autorizados conforme a la presente ley.*[9]
>
> *g) Principio de seguridad: la información sujeta a Tratamiento por el Responsable del Tratamiento o Encargado del Tratamiento a que se refiere la presente ley, se deberá manejar con las medidas técnicas, humanas y administrativas que sean necesarias para otorgar seguridad a los registros evitando su adulteración, pérdida, consulta, uso o acceso no autorizado o fraudulento.*
>
> *h) Principio de confidencialidad: todas las personas que intervengan en el Tratamiento de datos personales que no tengan la naturaleza de públicos están obligadas a garantizar la reserva de la información, inclusive después de finalizada su relación con alguna de las labores que comprende el Tratamiento, pudiendo sólo realizar suministro o comunicación de datos personales cuando ello corresponda al desarrollo de las actividades autorizadas en la presente ley y en los términos de la misma.* [9]

Posteriormente la Ley 1712 de 2013, por medio de la cual se crea la ley de transparencia y del derecho de acceso a la información pública nacional, cuyo objetivo es regular el derecho de acceso a Ia información pública, los procedimientos para el ejercicio y garantía del derecho y las excepciones a Ia publicidad de información, establece la divulgación de la información para promover y generar una cultura de

la transparencia.

La Ley en su Artículo 21, da las pautas para la divulgación total o parcial, indicando que la información puede tener una versión pública siempre y cuando mantenga la reserva únicamente de la parte indispensable.

> *i) Principio de Ia divulgación proactiva de Ia información. El derecho de acceso a Ia información no radica únicamente en Ia obligación de dar respuesta a las peticiones de Ia sociedad, sino también en el deber de los sujetos obligados de promover y generar una cultura de transparencia, lo que conlleva Ia obligación de publicar y divulgar documentos y archivos que plasman Ia actividad estatal y de interés público, de forma rutinaria y proactiva, actualizada, accesible y comprensible, atendiendo a límites razonables del talento humano y recursos físicos y financieros.*[10]

Finalmente, en el Código Nacional de Buenas Prácticas para las Estadísticas Oficiales se hace referencia a la anonimización de los microdatos en el principio 5:

> *Confidencialidad. Las entidades pertenecientes al Sistema Estadístico Nacional -SEN- deben garantizar la protección y la confidencialidad de la información con la que se producen las estadísticas oficiales, así como evitar la identificación de las fuentes.*[11]

Precisa en el numeral 5.3 que se debe:

> *5.3. Asegurar que la publicación de las estadísticas oficiales no permita la identificación individual de las fuentes.* [11]

Adicionalmente en los numerales 5.4 y 5.6 el Código, enfatiza que se debe contar con un protocolo para anonimizar los datos:

> *5.4 Aplicar protocolos para la protección y seguridad de la información.*
> *5.6. El acceso a microdatos anonimizados por parte de los usuarios debe estar sujeto a protocolos que garanticen la confidencialidad.* [11]

Legislación Internacional

Según el principio establecido por la División de Estadísticas de Naciones Unidas, los datos deben ser manejados con confidencialidad y exclusividad.

> *Los datos que reúnan los organismos de estadística para la compilación estadística, ya sea que se refieran a personas naturales o jurídicas, deben ser estrictamente confidenciales y utilizarse exclusivamente para fines estadísticos.*[3]

En el contexto jurídico, la Unión Europea, desde su Directiva 95/46/CE, en el considerando 26, excluye los datos anonimizados del alcance de la legislación sobre protección de los mismos considerando que

> *. . . los principios de la protección deberán aplicarse a cualquier información relativa a una persona identificada o identificable; que, para determinar si una persona es identificable, hay que considerar el conjunto de los medios que puedan ser razonablemente utilizados por el responsable del tratamiento o por cualquier otra persona, para identificar a dicha persona; que los principios de la protección no se aplicarán a aquellos datos hechos anónimos de manera tal que ya no sea posible identificar al interesado; que los códigos de conducta con arreglo al artículo 27 pueden constituir un elemento útil para proporcionar indicaciones sobre los medios gracias a los cuales los datos pueden hacerse anónimos y conservarse de forma tal que impida identificar al interesado.*[4]

El 10 de abril de 2014, el grupo de trabajo sobre protección de datos del articulo 29 (WP29) de la comisión Europea publicó ?Anonymisation Techniques onto the web? que fue adoptado como ?Opinion 05/2014 con Anonymisation Techniques? La justificación de este documento es muy acertada, ya que parte del concepto de ?datos abiertos? y de los beneficios que la liberación y reutilización de datos pueden aportar a los ciudadanos grupos de investigación y sociedad en general, para reconocer la importancia de aplicar las tecnicas de anonimización adecuadas que garanticen un tratamiento correcto de los datos de carácter personal.[5]

Las normas internacionales ISO 29100 tribuyen al proceso de anonimización la irreversibilidad como una característica esencial; la definen como:

> *El proceso por el cual la información de identificación personal se modifica de forma irreversible de tal manera que el responsable del tratamiento no puede identificar, directa o indirectamente, ya sea por sus propios medios o en colaboración con algún tercero, a la persona asociada a dicha información de identificación personal (ISO 29100:2011).*[6]

Proceso de Anonimización de los datos.

A la luz de la legislación existente, se entiende la anonimización como el resultado de un proceso de tratamiento de los datos personales con el fin de evitar la identificación de una persona de manera directa o indirecta y de forma irreversible.[4] Este proceso debe procurar controlar el riesgo de identificación de las personas que acceden a

entregar su información con fines estadísticos según los medios disponibles, teniendo en cuenta que se debe preservar la utilidad y aprovechamiento de los datos que necesitan los directos responsables o terceros.

Cabe señalar que los datos personales pueden estar contenidos en tablas de resumen de información, gráficos y demás documentos y publicaciones derivadas del quehacer estadístico; en estos casos el análisis o desagregación de la información de manera muy detallada, puede revelar información sensible o confidencial que conlleve a la identificación de una persona en particular.

No obstante las diferentes referencias en materia de legislación concernientes al proceso de anonimización, la ley no establece mecanismos o instrumentos para la protección de la información. Por ello, las herramientas o metodologías utilizadas para llevar a cabo la anonimización de un conjunto de datos deben ser constantemente evaluadas y actualizas según la tecnología disponible y los elementos contextuales, teniendo en cuenta que este proceso lleva implícito un factor de riesgo.

Por otro lado, existen problemas en la definición de mecanismos que eviten la identificación de quienes proveen la información, es decir los operadores, quienes de manera directa hacen manipulación de los datos, (incluyendo aquella información considerada como estrictamente confidencial) referentes a la edad, procedencia, salud, raza, religión, ideología, afiliación, etc., y que requieren de una protección especial.

Esta información sensible puede ser utilizada con fines inapropiados generando, lo que produciría una reducción en la confianza de quienes entregan información al Ministerio y afectando considerablemente la credibilidad de la entidad. Por tanto es primordial entregar directrices a los operadores con el fin de salvaguardar la integridad de la información hasta que sea trasmitida al Ministerio para su custodia.

Etapas del proceso de anonimización

Para lograr un proceso efectivo de anonimización de los datos es necesario eliminar de ellos los elementos suficientes para que no se pueda identificar a una persona mediante el conjunto de los medios que puedan ser razonablemente utilizados por el responsable del tratamiento o por terceros. Lo óptimo es decidirlo caso por caso y puede requerir la aplicación de una o varias técnicas, sopesando siempre el menor daño residual en la calidad de los datos. Es conveniente tener en cuenta que el proceso en todas sus etapas debe quedar documentado.

Etapa 1. Preanonimización

En qué consiste la preanonimización

En esta etapa se debe determinar claramente cuáles son las variables y los identificadores que de manera directa o indirecta pueden aportar en la identificación de una persona. Para tal fin es necesario clasificar cada una de las variables recolectadas para incorporarlas en el archivo o base de datos, teniendo en cuenta las siguientes categorías:

I. Variables de identificación geográfica.

II. Variables de identificación directa de personas o empresas.

III. Variables con magnitudes o numéricas.

IV. Variables de carácter sensible o confidencial.

V. Variables sin restricción para el acceso público.

Estas variables se clasificarán como directas o indirectas según puedan llevar a la identificación de una persona. Se debe procurar minimizar la cantidad de información personal a recolectar, teniendo especial cuidado con aquellos datos que contienen información sensible.

En qué momento se hace la preanonimización

Este proceso se debe realizar paralelo a la definición de las variables que se tendrán en cuenta en cada estudio o encuesta poblacional.

Quiénes participan en la preanonimización

Participan, como responsables de la primera etapa del proceso de anonimización un comité conformado por los líderes temáticos responsables del estudio, un equipo delegado de la Dirección de Epidemiología y Demografía y un equipo delegado de la Oficina de Tecnología de la Información (OTIC).

Etapa 2. Anonimización de microdatos de uso interno

En qué consiste la anonimización del microdato

Es la implementación de las actividades de protección de la privacidad de los datos sobre las características de las unidades de una población y que constituyen una unidad de información, una vez hagan parte de las bases de datos del Ministerio. Su finalidad es pasar a los temáticos de cada estudio la información estrictamente necesaria para la realización del respectivo análisis, evitando que otros funcionarios del Ministerio tengan acceso ilimitado a las variables de identificación.

En qué momento se hace la anonimización del microdato

La anonimización del microdato se debe realizar justo antes del inicio de la etapa de producción estadística.

Quiénes participan en esta anonimización del microdato

En primera instancia un equipo delegado de la Dirección de Epidemiología y Demografía, el cual estaría a cargo de la custodia de la información de ubicación geográfica (Sector, Sección, Manzana, USM), dado que estos datos solo tienen relevancia en el marco de la Muestra Maestra y no representan información relevante para los temáticos; el equipo de sistemas de la Oficina de Tecnologías de la Información y Comunicación (OTIC), se encargará de seudoanonimizar los datos, es decir, ocultar la identidad de las personas mediante seudónimos, ocultamiento y/o supresión de los datos personales según requerimiento.

Etapa 3. Anonimización

En qué consiste la anonimización

Consiste en la aplicación de una o varias técnicas tendientes a eliminar al máximo el riesgo de identificación de las personas o entidades, causando el menor dao posible a los datos.

En qué momento se hace la anonimización

Una vez finalizada la anonimización de microdatos de uso interno se puede proceder con la anonimización.

Quiénes participan en la anonimización

Un comité integrado por los líderes temáticos, un representante de OTIC y un representante la Dirección de Epidemiología y Demografía, debe establecer la técnica de anonimización más apropiada para cada una de las variables, según los recursos más actuales y disponibles; la implementación de los algoritmos propuestos será efectuada por (OTIC), quien se encargará también de realizar las pruebas necesarias y retornar la información a los líderes temáticos del estudio.

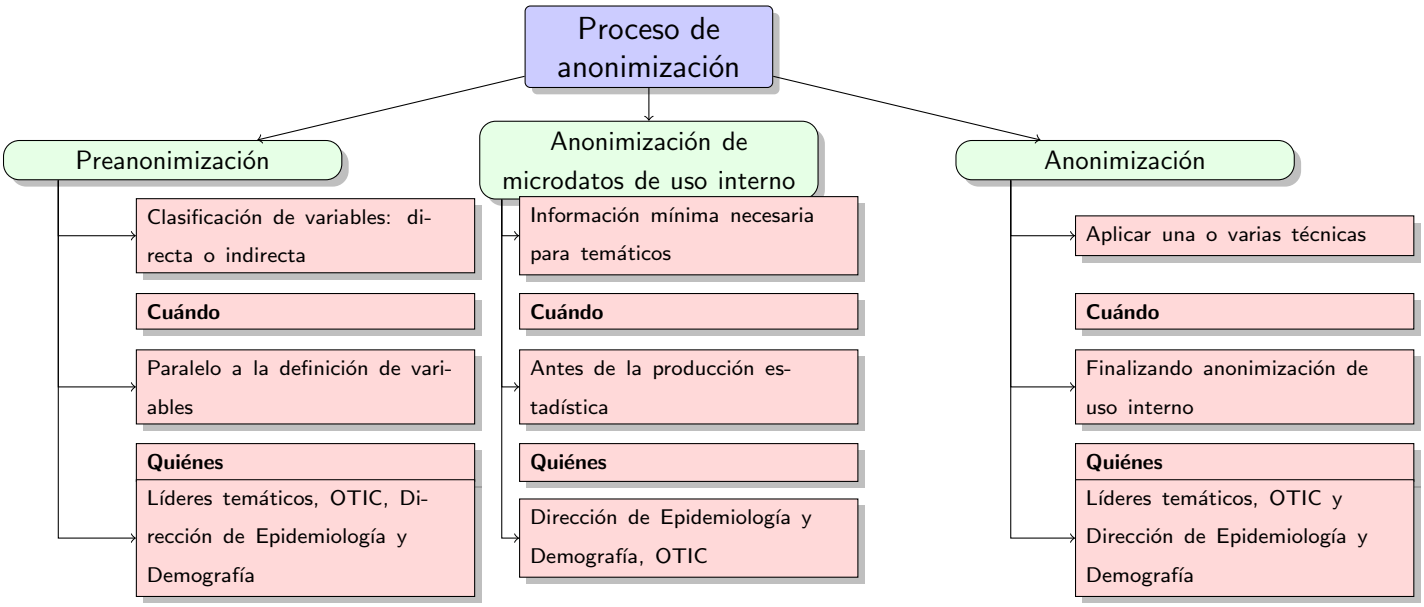

Figura 1: Proceso de anonimización

Técnicas de anonimización

Antes de hacer la selección de una o varias técnicas de anonimización es necesario hacer una valoración teniendo en cuenta que se debe minimizar el riesgo de:

1. Identificación o singularización de una persona.
2. Vinculabilidad entre los datos de la misma base o con datos provenientes de otra(s) bases de datos.
3. Inferencia o deducir a partir de los valores o a partir del análisis de la información contenida en la base de datos.

Es importante señalar que aunque el propósito es minimizar el riesgo como se dijo anteriormente, éste nunca desaparece completamente.

Básicamente las técnicas de anonimización corresponden a experiencias que han sido implementadas y se consideran exitosas y se pueden clasificar en dos categorías, por un lado los métodos basados en la modificación sistemática de los datos, que puede ser vista como una modificación de la veracidad de los mismos, *métodos de aleatorización o perturbación* y por otro lado los métodos que producen supresiones ya sean parciales o totales o reducción del nivel de detalle del conjunto original, *métodos de reducción o generalización.*

Métodos de aleatorización o perturbación.

Microagregación

La microagregación es útil cuando la variable es numérica y consiste en reemplazar un conjunto integrado por k datos con la media calculada sobre ese mismo conjunto. El k mínimo aceptado es 3, sin embargo en la elección del k se debe sopesar la pérdida de información y de variabilidad.

Adición de ruido

Esta técnica es especialmente útil cuando los atributos pueden causar un importante efecto adverso en las personas y consiste en la modificación de atributos mediante la generación de valores aleatorios lo que proporciona un conjunto de datos menos exactos, sin embargo se debe mantener la distribución general de los datos originales.

La adición de ruido no es una medida suficiente para anonimizar un conjunto de datos; esta técnica se debe combinar con otra (s) con el fin de minimizar los riesgos. Por otro lado es importante tener en cuenta que el ruido introducido en un conjunto de datos debe respetar la lógica de los atributos; por ejemplo no se pueden superar los límites naturales de una variable.

Permutación o intercambio de registros

Se puede tomar como un caso particular de adición de ruido, la aplicación de esta técnica, la cual consiste en intercambiar los valores de un registro a otro. Esta permutación garantiza que el rango y la distribución de valores sean idénticos además de que mantiene el nivel de detalle; no obstante las correlaciones entre los valores y los individuos pueden quedar destruidas. Es útil para variables de tipo categórico.

Redondeo

Consiste en la sustitución del valor de las variables originales por valores redondeados. Se aplica en casos donde la variable a tratar es numérica.

Reajuste de los pesos o factores de expansión

En los casos donde se conoce el tipo de muestreo utilizado es posible revertir el proceso, lo cual aumenta la posibilidad de identificar de manera puntual a una persona; por tanto, es conveniente hacer una modificación de los pesos con el fin de disminuir este riesgo.

Métodos de reducción o generalización

Eliminación de variables

Es aconsejable que esta sea la primera técnica a implementar en el proceso de anonimización de los datos dado que algunas de las variables pueden contener información sensible o de identificación directa y no siempre se puede aplicar sobre estas otros métodos de anonimización; también es muy posible que las bases de datos contengan información irrelevante para los responsables temáticos. En el caso de los estudios del Ministerio, se tienen variables dentro de las bases que hacen referencia a una ubicación geográfica y otros requerimientos muestrales que no representan información de interés para los análisis. Otros ejemplos de variables que deben ser suprimidas son: nombres, documentos de identidad, números telefónicos, correos, fotografías, etc.

Eliminación de registros

Cuando al aplicar alguna(s) técnica(s) de anonimización de datos, no surta el efecto deseado porque siguen siendo identificables algunas de las personas, se puede recurrir a la eliminación de registros; cabe señalar que este debe ser un procedimiento de último recurso. Esta metodología se aplica a datos de tipo categórico por cuanto los datos numéricos tienen otras alternativas.

Recodificación

Una ventaja de la recodificación es que se puede aplicar tanto a variables categóricas como numéricas y consiste en colapsar categorías con el fin de hacerlas menos específicas. La unión de las categorías se hace teniendo en cuenta las frecuencias y los propósitos del análisis.

Supresión de celdas

Sí al hacer cruce de información a partir de tablas de contingencia estas muestran celdas que pueden revelar información (frecuencias bajas) que conduzca a la identificación de alguna(s) persona(s) se puede suprimir una o varias celdas de la tabla. Esta técnica es útil también cuando se trata de información presentada en gráficos.

Bibliografía

[1] Lineamientos para la Anonimización de Microdatos [Internet]. 2014. Recuperado a partir de: http://www.dane.gov.co/files/sen/lineamientos/DSO-020-LIN-08.pdf

[2] Derecho.com. Tratamiento de los Datos [Internet]. 2014. Recuperado a partir de: http://www.derecho.com/c/Tratamiento_de_datos

[3] United Nations Statistical Commission. Fundamental principles of official statistics. Off Rec Econ Soc Counc. 1994;

[4] Europeas C, de la Unión Europea C. Directiva 95/46/CE del Parlamento Europeo y del Consejo de 24 de octubre de 1995 relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales ya la libre circulación de estos datos. Agencia de Protección de Datos; 1997.

[5] European Commission. Sixteenth Report of the Article 29 Working Party on Data Protection Covering The year 2012 [Internet]. European Commision; 2014 nov. Report No.: 16. Recuperado a partir de: http://ec.europa.eu/justice/data-protection/article-29/documentation/annual-report/files/2014/16th_annual_report_en.pdf

[6] ISO Copyright for the freely available standards. INTERNATIONAL STANDARD ISO/IEC29100 [Internet]. First edition 2011-12-15. Geneva: ISO copyright office; 2011. 28 p. Recuperado a partir de: http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html

[7] De Colombia CP. Constitución política de Colombia. Bogotá Colomb Leyer. 1991.

[8] Estatutaria L. 1266, 2008. Por Cual Se Dictan Las Disposiciones Gen Hábeas Data Se Regula El Manejo Inf Contenida En Bases Datos Pers En Espec Financ Credit Comer Serv Proveniente Terc Países Se Dictan Otras Disposiciones. 2008.

[9] Cuartas Rodríguez E, Jaller Escudero JD. El Habeas Data como Derecho fundamental y la Ley 1581 de 2012 y su decreto 1377 de 2013. 2014.

[10] Congreso de la República. Ley 1712 de 2014 [Internet]. 2014. Recuperado a partir de: http://www.mintic.gov.co/portal/604/articles-7147_documento.pdf.

[11] DANE. Código Nacional de Buenas Prácticas para las Estadísticas Oficiales.

[12] García PAD. ANONIMIZACIóN DE DATOS PERSONALES EN PROCESOS IMPLICADOS INFANTES Y ADOLESCENTES. LAS REGLAS DE HEREDIA. AR Rev Derecho Informático. 2008;(125):1-16.

[13] DAEDALUS JCG. Anonimización: Un Enfoque útil para la protección de la privacidad y de la confidencialidad. 21 Junio 2011 - Proces Semántico Tecnol Leng [Internet]. 21 de junio de 2011; Recuperado a partir de: http://www.daedalus.es/blog/es/anonimizacion-enfoque-util-proteccion-privacidad-confidencialidad/

[14] TicWeb.es - Tecnólogos de la información en la Web. Microdato [Internet]. 2011. Recuperado a partir de: http://www.ticweb.es/microdato-y-microformato-pilares-de-la-web-3-0-semantica/

[15] JEUSKAL ESTATISTIKA ERAKUNDEA - Instituto Vasco de Estadística. Técnicas de Protección y Seguridad de Datos Estadísticos. [Internet]. Donostia-San Sebastián, 1 01010 VITORIA-GASTEIZ; Recuperado a partir de: http://www.eustat.eus/documentos/datos/ct_02_c.pdf