METODOLOGÍA DE ANONIMIZACIÓN DE DATOS

2020

Planta Docente, EPBM y ES

Rafael Ramos Ballesteros

Grupo de Información y Análisis Sectorial

25/09/2020

Contenido

INTRODUCCIÓN

El presente documento da a conocer la metodología, lineamientos y políticas que permita generar un proceso de anonimización de la información, con la ayuda de algoritmos de enmascaramiento de datos a las bases de datos del Ministerio de Educación, esto con el fin de poder entregar a terceros dicha información, sin tener que vulnerar los derechos de privacidad y salvaguarda de la integridad física y psicológica de los estudiantes y docentes que tengan su información consignadas en las bases de datos del Ministerio de Educación.

El Ministerio de Educación como ente encargado de la consolidación y divulgación de la información estadística sobre los registros de matrícula de EPBM, ES y Planta Docente, es consultado continuamente por diversos sectores y actores, sobre temas y reportes relacionados a los registros de matrículas, establecimientos y planta docente. Por lo tanto, se hace imperioso garantizar la protección de la información privada y/o sensible, de los estudiantes y docentes con respecto al tratamiento de sus datos personales consignados en las bases de datos del MEN. En los capítulos siguientes se desarrollará un algoritmo de permutación y encriptación de caracteres de las variables que contienen información reservada.

OBJETIVO

- Diseñar y desarrollar la metodología de intercambio de los datos del Ministerio de Educación a las entidades externas, teniendo en cuenta los niveles de seguridad requeridos.
- Garantizar la protección y privacidad de los datos sensibles del Ministerio para los Docentes y Estudiantes
- Clasificar los datos que van a ser privados, sensibles y públicos de cada una de las Bases de datos del Ministerio de Educación, donde permita la singularización, vinculación e inferencia de la persona.
- Establecer un método de anonimización de los datos que se consideran Sensibles y Privados por parte del Ministerio de Educación

CONCEPTOS BASICOS

En este capítulo se citarán los lineamientos jurídicos dados por el gobierno colombiano respecto al tema, además se abordarán algunos conceptos sobre datos personales, enumerados en la introducción.

Conceptos legislativos

El Gobierno Colombiano, en el año 2012 a través del decreto 1581[1] de Octubre, definió el tratamiento que debe ser dado a los datos personales por parte del responsable del mismo, ya sea una persona natural o jurídica, pública o privada. En el año 2013 con el Decreto 1377[2] de junio, se reglamenta parcialmente el 1581, en el artículo 12 ***Requisitos especiales para el tratamiento de datos personales de niños niñas y adolescentes; en el que*** se dictan los parámetros y requisitos necesarios para dar tratamiento de menores de edad.

Conceptos Técnicos[3]

El objetivo de la codificación de la información es evitar la Singularización, Vinculabilidad e Inferencia, a continuación, se definen cada una

Singularización:
Es la posibilidad de extraer de un conjunto de datos algunos registros que puedan identificar a un alumno

Vinculabilidad
Es la capacidad de vincular como mínimo dos registros de un único interesado o de un grupo de interesados, ya sea de en la misma base de datos o en dos o más bases distintas, si el si el consultor puede determinar que dos registros están asignados al mismo grupo de personas, y estas no pueden ser singularizadas.

Inferencia
Es la posibilidad de deducir con una probabilidad significativa el valor de un atributo a partir de los valores de un conjunto de otros atributos.

ESTRATEGIA DEL PROCESO DE ANONIMIZACIÓN

Se establece un proceso hibrido con el desarrollo de un algoritmo donde se define un tratamiento posterior al almacenamiento de los datos sensibles reportados por los establecimientos educativos, el cual se asigna un código único de identificación a los datos, sin permitir ningún tipo de singularización

[1] http://www.sic.gov.co/drupal/sites/default/files/normatividad/Ley_1581_2012.pdf
[2] http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=53646
[3] http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

o reconocimiento de un estudiante o docente. De igual manera este proceso de anonimización tampoco permitirá vinculación alguna a través de las demás variables contenidas en cada registro, inferencia en la ubicación del estudiante o docente, para enfocarse exclusivamente en la generación de estadísticas descriptivas que cumplen con el objeto de información y a la vez de garantía de los derechos fundamentales de los niños, niñas y adolescentes.

Adicional a eso, se encriptará y cifrará la información entregada a la entidad externa con el fin de garantizar que solo pueda ser leída o consumida por el responsable de dicha entidad, por los medios de Distribución previamente pactados FTP, CD, WS entre otros.

METODOLOGÍA PARA LA ANONIMIZACIÓN

Antes de abordar la metodología que se propone se hará un abreve descripción de algunas técnicas de anonimización

TECNICAS DE ANONIMIZACIÓN

Aleatorización.

Este tipo de técnicas consisten en modificar el orden de los datos, el fin de estos procesos, es de minimizar las relaciones existentes entre los datos y las personas. Mediante la aplicación de estos métodos es posible establecer la singularidad de los registros, es necesario complementar con otras técnicas para reducir la probabilidad de identificación de personas a través de inferencias.

Adición de ruido.

Estas técnicas se caracterizan por modificar los atributos de un conjunto de datos adicionando información extra a los registros, la adición debe garantizar la integridad de la data. La adición de ruido maximiza su eficiencia en datos numéricos ya que permite manejar un rango (+ -), en los registros; este procedimiento debe combinarse con otras técnicas de anonimización.

Permutación

El objetivo de esta técnica es combinar los valores de dos o más variables de un conjunto de datos, este procedimiento es útil cuando se quiere conservar la composición exacta de cada registro de la data. La permutación necesita de la completitud y exactitud de las variables escogidas para tal fin.

MECANISMO DE ANONIMIZACIÓN DE DATOS

El mecanismo de anonimización para las bases de datos de matrícula se basa en un algoritmo de permutación, acompañado con un enmascaramiento de algunos caracteres permutados con

caracteres especiales o adición de ruido. A continuación, se dan los pasos a seguir para realizar el proceso para salvaguardar la privacidad de los estudiantes y docentes.

Paso 1. Selección de variables.

Un paso primordial para un proceso de protección de datos sensibles es la elección de los atributos considerados vulnerables a la singularización, vinculabilidad e inferencia.

De acuerdo a las variables consignadas en **SIMAT**, se tendrán en cuenta las siguientes variables.

Se hace la observación que debido a la sensibilidad y tratamiento especial que deben ser sometidos los datos de menores de edad, se consideran a aquellos atributos que puedan ayudar a determinar la ubicación de los estudiantes, serán consideradas como privadas.

- Variables de Singularización.

PER_ID	DIRECCION_RESIDENCIA
APELLIDO1	TELEFONO
APELLIDO2	FECHA_NACIMIENTO
NOMBRE1	NRO_DOCUMENTO
NOMBRE2	

- Variables de Vinculabilidad.

NOMBRE_ESTABLECIMIENTO	GENERO
NOMBRE_SEDE	GRUPO
EDAD	GRADO

- Variables de Inferencia

DPTO_CARGA	Divipola_MUNICIPIO
NAC_DEPTO	ETNIA
NAC_MUN	

Estas últimas variables en conjunto con cualquier de los dos otros grupos podrían determinar la identidad o ubicación de los educandos.

De acuerdo con las variables Consignada en el **SNIES**, se tendrán en cuenta las siguientes variables

Se hace la observación que debido a la sensibilidad y tratamiento especial que deben ser sometidos los datos de menores de edad, se consideran a aquellos atributos que puedan ayudar a determinar la ubicación de los estudiantes, serán consideradas como privadas.

- Variables de Singularización.

APELLIDO1	DIRECCION_RESIDENCIA
APELLIDO2	TELEFONO
NOMBRE1	FECHA_NACIMIENTO
NOMBRE2	NRO_DOCUMENTO

- Variables de Vinculabilidad.

NOMBRE_UNIVERSIDAD	GENERO
EDAD	CARRERA
EMAIL_PERSONAL	SEMESTRE

- Variables de Inferencia

	Divipola_MUNICIPIO_Universidad
DEPARTAMENTO_NAC	ETNIA
MUNICIPIO_NAC	

Paso 2. Mecanismo de Permutación.

En esta parte del proceso, se realizan dos permutaciones, ya que se debe anonimizar los atributos que permitan la singularización de los estudiantes y docentes, así como también los que permitan la ubicación geográfica. Como la información de georeferenciación de las matrículas, es muy requerida por usuarios internos y externos del MEN, se enmascararán los datos personales y la identificación de los establecimientos educativos o universidades.

- ***Anonimización de datos personales:***

  Las variables que se tendrán en cuenta son las siguientes.
  Municipio de Nacimiento
  Fecha de Nacimiento
  Primer Apellido
  Segundo Apellido
  Primer Nombre
  Segundo Nombre

  El código se construirá de la siguiente forma:
  - Tomar el Segundo carácter del Primer Apellido
  - Los tres primeros carácteres del municipio de Nacimiento

- La fecha de nacimiento sin separadores especiales con el siguiente formato ddmmaaaa
- El quinto carácter del Segundo Apellido
- El cuarto carácter del Primer Nombre
- Sexto carácter del Segundo Nombre

Una vez construido el código, este se enmascara con caracteres especiales de la siguiente forma:

- Cambiar el carácter de la posición 5 por el símbolo %
- Cambiar el carácter de la posición 10 por el símbolo #
- Cambiar el carácter de la posición 15 por el símbolo &

A continuación, se muestra un ejemplo con un registro aleatorio

Municipio Nacimiento	Fecha Nacimiento	Primer Apellido	Segundo Apellido	Primer Nombre	Segundo Nombre
76147	19/04/1998	PARRA	BOTERO	LISET	BIBIANA

Los caracteres resaltados en rojo son los seleccionados según el orden dado arriba, el código generado sería el siguiente

Código
R147190419198RTN

Este después de aplicar el enmascaramiento el código final queda de la siguiente forma

Código Final
R147%9041#198R&N

En aquellos casos donde no se cuente con los registros completos como en el caso de Segundo Apellido y Segundo Nombre el carácter correspondiente, será reemplazado por una “x”; de igual manera se procede para los faltantes en nombres o apellidos con pocos caracteres.

En los apellidos y nombres compuestos se deben eliminar los espacios en blanco, antes de ejecutar el procedimiento.

- ***Anonimización de datos instituciones educativas.***

Las variables que se tendrán en cuenta son las siguientes:

Departamento de Carga
Secretaria de educación
Municipio de Carga
Sede_ID

El código se construirá de la siguiente forma

- Tomar el tercer carácter del Departamento de carga
- Tomar el sexto carácter del Departamento de carga
- Tomar los datos de sede_ ID
- Tomar el séptimo carácter del municipio de carga
- Tomar el primer carácter del municipio de carga
- Tomar los tres primeros caracteres del código de la secretaria de educación

Una vez construido el código, este se enmascara con caracteres especiales de la siguiente forma.

- Cambiar el carácter de la posición 3 por el símbolo %
- Cambiar el carácter de la posición 6 por el símbolo #
- Cambiar el carácter de la posición 9 por el símbolo &

Departamento Carga	Secretaria educación	Municipio Carga	Sede_ID	Código	Código Final
ValledelCauca	4815	Jamundí	77494	LD77494IJ481	LD%74#4I&481

- ***Anonimización de datos Docentes asociados al anexo3A***

  Las variables se tendrán pendiente son las siguientes:

  Cedula
  Genero
  Cargo
  Municipio
  Departamento
  Fecha Vinculación
  Fecha Nacimiento

  El código se generará teniendo en cuenta las siguientes condiciones

  - Tres primeros números de la cedula
  - El Código del Cargo
  - Primer Carácter del Código del Municipio
  - Código del Departamento
  - El mes de nacimiento
  - Se imputa una variable con valor ‘XXX’

  Luego los campos siguientes se aplican diversas ecuaciones para la anonimización de tipo de datos Numéricos

Cedula → (Cedula+85)*8
CodigoDane→(CodigoDane+15)*2
FechaNacimiento→(FechaNacimiento+4)*7
CodigoMuncipio →(CodigoMunicipio+150)*3
CodigoDepartamento→(CodigoDepartamento+50)*8

CONSIDERACIONES.

Se debe considerar que la efectividad del algoritmo depende de la calidad de los registros, este nuevo código puede ser de apoyo a actividades relacionadas con la calidad de la data, se debe definir antes de la implementación el grupo encargado de llevar a la anonimización esto con el fin de blindar el desarrollo del proceso.

Recuento	8420854	
Recuento	8403617	17.237
	99,8%