INTRODUCCIÓN

En un entorno donde los datos son cruciales para la toma de decisiones, la correcta documentación, construcción y estandarización de las variables en las bases de datos es esencial para garantizar la calidad, coherencia y comparabilidad de la información. La falta de uniformidad en la definición y uso de variables puede llevar a errores, malinterpretaciones y dificultades en el análisis. Esta guía tiene como objetivo establecer lineamientos claros para la documentación de metadatos, la estandarización de las variables y los atributos esenciales que deben cumplir las bases de datos y las variables que incorporan, promoviendo así la consistencia, interoperabilidad y fiabilidad de la información en todos los procesos de gestión de datos de la entidad.

En este contexto, se busca no solo optimizar la calidad de los datos, sino también facilitar su utilización en análisis y reportes, promoviendo de esta manera una toma de decisiones más informada y basada en datos de alta fidelidad.

1. OBJETIVO

Ofrecer un marco integral de mejores prácticas para la correcta documentación y estandarización de las variables en las bases de datos de la entidad, asegurando así su uniformidad y consistencia. Este propósito abarca, entre otros aspectos, el establecimiento de criterios claros y detallados sobre los atributos que deben cumplir las variables, la forma en que deben ser estructuradas y documentadas, y las políticas operativas que deben observarse en todos los procesos relacionados.

2. ALCANCE

Esta guía está dirigida a todos los equipos de la entidad que gestionan datos, desde su recolección hasta su análisis y reporte. El documento se enfoca en la estandarización de las variables dentro de las bases de datos, abarcando su definición, clasificación, formato y documentación. Además, cubre aspectos relacionados con la calidad de los metadatos y los atributos necesarios para asegurar la integridad y consistencia de las bases de datos a lo largo de su ciclo de vida. También incluye políticas operativas esenciales para mantener la homogeneidad de las variables y cumplir con las normativas establecidas, tanto internas como externas.

3. DEFINICIONES

Base de Datos: es un conjunto organizado de datos personales que son objeto de tratamiento. Esta definición se extiende a los archivos (depósito ordenado de datos incluidos datos personales – Sentencia C-748 de la Corte Constitucional).

Datos categóricos: “se refieren a una forma de información que puede almacenarse e identificarse basándose en sus nombres o etiqueta”. (QestionPro, 2024)

Diccionario de datos: “conjunto de metadatos que contiene las características lógicas y puntuales de los datos que se van a utilizar en el sistema, incluyendo descripción, alias, contenido y organización” (McCalla, 2012)

Distribución temporal: “patrón o tendencia de ocurrencia de un fenómeno específico durante un período de tiempo.” (Zhuo , 2020)

Metadatos: “conjunto de atributos o elementos necesarios para describir un recurso determinado, que funciona como identificador de los materiales digitales diseñados. (Agudelo Benjumea, 2020)

Variables categóricas: “las variables categóricas contienen un número finito de categorías o grupos distintos”. (Soporte de Minitab, 2024)

Variables finales: “una variable final es una constante; una vez inicializada, su valor no puede cambiarse”. (Datacamp, 2024)

Variables intermedias: “son variables situadas dentro de un orden causal entre la variable dependiente y el factor de estudio, pero a diferencia de los factores de confusión forman parte de la cadena causal”. (Solís Sánchez, 1999)

4. NORMATIVIDAD ASOCIADA

- Ley 1581 de 2012: por la cual se dictan disposiciones generales para la protección de datos personales.
- Ley 1712 de 2014: por medio del cual se crea la ley de transparencia y del derecho de acceso a la información pública nacional.
- Resolución 1419 de 2017: por la cual se expiden los Lineamientos para el proceso estadístico en el Sistema Estadístico Nacional (SEN).
- Resolución 140 de 2022: por medio de la cual se adopta la política de seguridad y privacidad de la información en la Agencia Distrital para la Educación Superior, la Ciencia y la Tecnología – Atenea.
- Resolución 141 de 2022: por medio de la cual se adopta la política de protección de datos personales en la Agencia Distrital para la Educación Superior, la Ciencia y la Tecnología – ATENEA.

5. DESARROLLO

5.1. PRINCIPIOS

“La calidad trasciende de ser considerada una prioridad competitiva a convertirse en un requisito para que las organizaciones y la sociedad puedan aprovechar los datos que se producen a nivel interno y externo en la generación de valor”, (Rangel Carrillo, 2020), por tanto, la gestión de calidad de datos en la Agencia serán principios generales los siguientes:

i. *Documentación continua y exhaustiva:* Toda actividad relacionada con la recolección, transformación, y uso de los datos debe ir acompañada de una documentación adecuada.

Esto incluye diccionarios de datos, metadatos, procesos de validación y cambios en los datos.

ii. *Estandarización conforme a lineamientos de la entidad:* Todos los datos deben seguir los estándares definidos por la entidad para asegurar consistencia y coherencia en los análisis y reportes.

iii. *Validación y control de calidad en todas las etapas de producción de información*: Implementar validaciones de calidad de los datos en cada etapa de la cadena de producción de información, desde la recolección hasta la difusión.

iv. *Transparencia y trazabilidad*: todo dato debe ser trazable, lo que significa que se debe poder identificar su origen, las modificaciones realizadas, y los responsables de dichas modificaciones.

v. *Accesibilidad y seguridad*: Los datos deben ser fácilmente accesibles para quienes los necesiten, pero solo dentro de los niveles de acceso permitidos, garantizando la seguridad y confidencialidad de la información.

vi. *Mejora Continua*: La gestión de la calidad de los datos debe verse como un proceso de mejora continua. Las actividades de control de calidad deben ser evaluados y ajustados regularmente.

5.2. ANÁLISIS EXPLORATORIO DE DATOS

El primer paso en la gestión de la calidad de los datos es el entendimiento de los conjuntos de datos que procesa la entidad. Esto ayudará a identificar patrones, anomalías y relaciones que guíen análisis más profundos y revelen posibles problemas de calidad de los datos. Las actividades recomendadas en este punto son:

a. **Revisión de la estructura y tipos de datos**: inspeccionar la estructura de los datos, verificando los tipos de variables (numéricas, categóricas, fechas, etc.) y asegurarse de que cada campo tiene el tipo correcto.

b. **Evaluación de las distribuciones**: visualizar las distribuciones de las variables numéricas y categóricas, esto permitirá detectar posibles valores atípicos, sesgos en la distribución o transformaciones necesarias.

c. **Identificación de valores atípicos**: detectar valores fuera de los rangos esperados o extremos que no correspondan a patrones esperados y verificar la veracidad de estos valores.

d. **Análisis de datos faltantes**: revisar la proporción de datos faltantes por variable y registro, así como su patrón de ausencia.

e. **Detección de inconsistencias**: identificar inconsistencias lógicas entre variables relacionadas.

f. **Exploración de distribuciones temporales**: detectar patrones estacionales, tendencias o irregularidades en la recolección de datos.

g. **Identificación de duplicados**: identificar registros duplicados en los archivos de datos.

5.3. CRITERIOS DE EVALUACIÓN DE CALIDAD DE LOS DATOS

Una vez que se han entendido los datos, el siguiente paso es evaluar los diferentes campos del conjunto de datos validando como mínimo los siguientes criterios:

- *Completitud*: el grado en que los datos están completos y no tienen valores faltantes. Esto implica evaluar los valores faltantes por variable y registro.
- *Coherencia:* los datos deben ser consistentes entre sí y no deben presentar contradicciones lógicas. Para evaluarlo es necesario la definición y validación de reglas lógicas entre los campos.
- *Exactitud:* los datos deben reflejar con precisión la realidad que intentan representar. Para lograrlo es fundamental el método de recolección o la comparación con otras fuentes.
- *Validez:* los datos deben ajustarse al tipo de datos esperado y a los dominios de valores aceptados. Esto conlleva a validar si los datos se encuentran en los dominios y rangos esperados o si se encuentran en los formatos adecuados.
- *Uniformidad o Estandarización:* los datos deben seguir los formatos y convenciones definidos, de manera consistente en toda la base o conjunto de datos.
- *Precisión (Granularidad):* los datos deben tener el nivel de detalle adecuado según los requerimientos del análisis o el proceso.
- *Trazabilidad:* los datos deben tener un registro claro de su origen y cualquier transformación que hayan sufrido a lo largo del tiempo.
- *Actualidad (Tiempos de actualización):* Los datos deben estar actualizados y reflejar la realidad en el tiempo correcto.

5.4. EVALUACIÓN DE LA CALIDAD DE DATOS

Se hace necesario monitorear la calidad de los datos almacenados, de modo que puedan ser identificadas las anomalías y establecer un proceso de corrección sobre las mismas, evitando así la toma de decisiones sobre la base de información incorrecta (Yanes Pavón, 2019). Para la Agencia Atenea, los criterios de evaluación se caracterizan por:

Criterio de Calidad	Descripción	Métrica
Completitud	Grado en el que los datos no contienen valores faltantes.	Porcentaje de valores faltantes por columna.
Validez	Grado en que los datos cumplen con las reglas predefinidas (rango, formato).	Porcentaje de datos válidos (cumplen con el formato esperado). Número de violaciones de reglas.
Exactitud	Qué tan precisos y correctos son los datos comparados con una fuente confiable.	Error promedio comparado con una fuente externa (si aplica).
Consistencia	Nivel de coherencia entre diferentes conjuntos de datos relacionados.	Número de inconsistencias detectadas entre bases de datos relacionadas. Porcentaje de duplicados.

Unicidad	Grado en que los registros son únicos, sin duplicados.	Porcentaje de duplicados. Número de claves primarias duplicadas.
Uniformidad	Medida en que los datos siguen el mismo formato o convención.	Porcentaje de valores con el formato correcto (ej., fecha en formato ISO).
Integridad	Relación correcta entre diferentes conjuntos de datos o tablas.	Porcentaje de claves foráneas válidas. Porcentaje de relaciones completas entre tablas.
Oportunidad	Actualización de los datos dentro del tiempo esperado o requerido.	Latencia en la actualización de los datos (días/horas). Porcentaje de cumplimiento del cronograma de actualización.
Accesibilidad	Facilidad de acceso y disponibilidad de los datos.	Tasa de éxito en el acceso a los datos. Tiempo promedio de respuesta en la recuperación de datos.

5.5. ESTANDARIZACIÓN DE DATOS

Después de validar los criterios de evaluación de calidad de un conjunto de datos, un siguiente paso tiene que ver con la estandarización. A continuación, se presentan los estándares generales a manejar en la Agencia.

a. Estándares de formatos de datos:

- Números
  - Separador decimal: coma (,).
  - Separador de miles: punto (.).
- Fechas
  - Se usará el formato ISO 8601 (AAAA-MM-DD), que es el formato internacional de fechas. (ISO)
- Porcentajes:
  - Expresar porcentajes siempre como decimales (por ejemplo, 0.25 en lugar de 25%) en bases de datos, dejando la conversión para la presentación.

b. Estándares de Codificación de datos categóricos:

- Listas y dominios oficiales
  - Para las variables categóricas se usarán listas o clasificaciones oficiales, en este sentido se tomará de referencia el documento “CONCEPTOS Y CLASIFICACIONES PARA LA CARACTERIZACION DE BENEFICIARIOS V1 que genera y actualiza periódicamente la Subgerencia de Análisis de Información y Gestión del Conocimiento - SAIGC basadas en referentes internos o externos.

c. Estándares de nombres de variables:

- Convención de nombres

- Seguir una convención clara y coherente para los nombres de las variables, preferiblemente en snake_case (ej. nivel_formación)

- Uso de prefijos y sufijos
  - Usar prefijos o sufijos para indicar categorías comunes o jerarquías dentro de las variables. Ejemplo: fecha_nacimiento, codigo_ies
- Longitud de las variables
  - Limitar el nombre de las variables a un máximo de 20 caracteres

d. Estándares de nombres de archivos:

Los archivos deberán ser nombrados con la fecha de corte en formato (AAAA_MM_DD) seguido del nombre del grupo que lo genera, el nombre del archivo de datos y la versión, el siguiente es un ejemplo del nombre del archivo generado por SAIGC con corte al 31 d de octubre de 2024 de la versión 7 de la base maestra de beneficiarios del programa Jóvenes a la E, ej:**20241031_SAIGC_MAESTRA_JE_BENEFICIARIOS_V7**.

e. Estándares de codificación y texto:

- Codificación UTF-8
  - Se usará UTF-8 para la codificación de caracteres, que es el estándar más común y ampliamente soportado para textos multilingües.
- Uso de mayúsculas y minúsculas
  - Sobre un conjunto de datos se debe validar el uso de mayúsculas o minúsculas en todas las categorías textuales

f. Control de versionamiento de archivos

- Control de versionamiento: Los cambios en los archivos de datos deberán contar con versionamiento semántico para identificar los cambios, por ejemplo:
  - v1.0.0: Primera versión.
  - v1.1.0: Actualizaciones menores.
  - v2.0.0: Cambios significativos.

g. Estándares para la integración de datos

- Llaves primarias y foráneas
  - Estandarizar la definición de llaves primarias para identificar de manera única cada registro y llaves foráneas para relacionar datos entre tablas.
- Codificación unificada para identificadores

- Se usarán identificadores únicos y unificados en todos los sistemas.

h. Intercambio entre áreas de archivos grandes

- Formatos *csv o *txt, quoting
  - (doble comilla)
- Campos de cadena o texto y separador pipeline (|) en archivos planos
  - Sin espacios al principio y al final del campo (trim) o caracteres especiales no visibles.

5.6. DOCUMENTACIÓN DE LOS DATOS

Todo conjunto de datos debe estar acompañado de un documento de metadatos que describa claramente el origen, propósito, estructura y uso de los datos, este documento se debe elaborar en el formato de diccionario de datos que se anexa al presente documento y se denomina “Formato Diccionario de Datos”.

Uno de los componentes fundamentales de los metadatos serán los diccionarios de datos que al momento de documentar tendrán las siguientes características mínimas:

a. Diccionario de datos.
b. Nombre de variables y descripciones: No utilizar espacios, caracteres especiales o acentos en los nombres de las variables.
c. Documentación de dominios.
d. Definición de mínimos.
e. Validaciones de consistencia.
f. Definición de conceptos.
g. Variables finales versus Variables intermedias.

5.7. POLÍTICAS DE OPERACIÓN

Se hace necesario definir de manera general las decisiones que corresponden al manejo de la información durante su procesamiento.

a. Si base no corresponde con diccionario, actualizar diccionario, socializar y documentar
b. Gobierno de metadatos transversales es SAIGC
c. Manejo de cortes y generación de versiones
d. Validación no solo sobre la base de ingreso sino también de salida
e. La articulación de dominios con fuentes oficiales

6. ANEXOS:

- A1_G5_DE Anexo - Documento de conceptos y clasificaciones en la caracterización de beneficiarios.

7. DOCUMENTOS DE REFERENCIA:

(Secretaría de Educación Pública. Perkins International Latín America & Sense International. (2011).

ACNUR. (s.f.). Estatuto Temporal de Protección de Migrantes Venezolanos.

Agudelo Benjumea, M. (2020). *Los Metadatos*. Obtenido de http://biblioteca.udgvirtual.udg.mx/jspui/handle/123456789/3631

American Psychiatric Association. 2014. (2014).

Código Civil Colombiano . (s.f.). Artículo 41.

Constitución Política de Colombia. (1991). Artículos 286 y 298. Colombia.

DANE. (2020). Guía para la inclusión del enfoque diferencial e interseccional.

Datacamp. (05 de 12 de 2024). *datacamp.com*. Obtenido de https://www.datacamp.com/es/doc/java/final

Departamento Nacional de Planeación. (s.f.).

IDECA. (s.f.). Infraestructura de Datos Espaciales de Bogotá.

ISO. (s.f.). 8601.

Ley 115 de 1994, Ley 30 de 1992. (s.f.).

Ley 136. (1994). Ley 136.

Ley 142. (1994). (Art. 101.8).

Ley 1448. (2011).

McCalla, F. (2012). Diccionario de Datos: Un enfoque semántico, de seguridad y usabilidad. *Revistas Académicas UTP*, 32-48.

Min. Rel. Exteriores. (2017). (Resolución 5797 de 2017, resolución 1272 de 2017).

Ministerio de la Protección Social & ACNUR, 2011. (s.f.).

Ministerio del Interior. (2018). Decreto 762 de 2018. Artículo 2.4.4.2.1.10.

OIM. (2006). Glosario sobre Migración. .

Organización Internacional del Trabajo (OIT). (1989). Manual del Convenio 169.

QestionPro. (05 de 12 de 2024). *https://www.questionpro.com/blog/es*. Obtenido de https://www.questionpro.com/blog/es/datos-categoricos/#:~:text=Los%20datos%20categ%C3%B3ricos%20se%20refieren,en%20lugar%20de%20medirse%20num%C3%A9ricamente.

Rangel Carrillo, A. M. (diciembre de 2020). Principles, guidelines, dimensions, and attributes for the quality evaluation of Open Government Data. *Aibi revista de investigación, administración e ingeniería*, 54-65.

SIVIGE. (2016).

Solís Sánchez, G. (1999). Epidemiología y metodología científica aplicada a la pediatría (VI): Confusión e interacción. *EDUCACION CONTINUADA*, 91.

Soporte de Minitab. (05 de 12 de 2024). *https://support.minitab.com/es*. Obtenido de https://support.minitab.com/es-mx/minitab/help-and-how-to/statistical-modeling/regression/supporting-topics/basics/what-are-categorical-discrete-and-continuous-variables/

Yanes Pavón, J. (2019). La evaluación de la calidad de datos: una aproximación criptográfica. *Computación y Sistemas*, Epub 10-Mar-2021.

Zhuo , L. (2020). Agent-based modelling and flood risk management: A compendious literature review. *Journal of Hydrology*, 125600.

8. RELACIÓN DE FORMATOS:

CODIGO	NOMBRE DEL FORMATO
F1_G5_DE	Formato Diccionario de Datos

9. CONTROL DE CAMBIOS:

Fecha	Versión	Descripción del Cambio

VALIDACIÓN	NOMBRE	CARGO	FECHA
Elaboró	Raúl Andrés Gómez Aldana	Contratista - Profesional	05/12/2024
Elaboró	Yehiman Alberto Bernal Hernández	Contratista - Profesional	05/12/2024
Revisó	John Alejandro Torres Sichacá	Profesional especializado	12/12/2024
Aprobó	Javier Andrés Rubio Sáenz	Subgerente SAIGC	13/12/2024