# Guía de Calidad e Interoperabilidad de Datos Abiertos

Octubre 2024

TIC

**Ministerio de Tecnologías de la Información y las Comunicaciones**

Oscar Mauricio Lizcano Arango – Ministro de Tecnologías de la Información y las Comunicaciones

Belfor Fabio García Henao – Viceministro de Transformación Digital

William Fernando Oviedo Camargo – Director de Gobierno Digital

Natalia Albañil – Subdirectora(e) de Fortalecimiento de Capacidades Públicas Digitales

Luisa Fernanda Medina – Líder de la iniciativa de datos abiertos

**Equipo Técnico BDGUIDANCE**

Andrés Briceño – Gerente de proyecto

Genaro Muñoz Obregón – Consultor senior analítica de datos

Alirio Otálora – Científico de datos

Yesid Patiño – Científico de datos

Milton García – Ingeniero de datos

| **Versión** | **Observaciones** |
|---|---|
| Versión 1 | **Guía de calidad e interoperabilidad** |
| Julio 2022 | Inclusión y actualización de contenido, capitulación, ajuste de formato. |
| Octubre 2024 | Corrección ortográfica y ajuste de línea gráfica |

Comentarios, sugerencias o correcciones pueden ser enviadas al correo electrónico: datosabiertos@mintic.gov.co

Guía de estándares de calidad e Interoperabilidad de los datos abiertos del Gobierno de Colombia.

# Contenido

# Lista de tablas

# Lista de ilustraciones

# Introducción

En Colombia se han llevado a cabo en los últimos años, importantes esfuerzos para impulsar la apertura de datos públicos, los cuales han tenido resultados significativos, como se puede evidenciar en el portal nacional de datos abiertos www.datos.gov.co, herramienta dispuesta para que todas las entidades públicas en Colombia publiquen sus datos abiertos y garanticen el derecho de acceso a la información pública a los ciudadanos.

Además de lograr que las instituciones publiquen los datos que producen como resultado de su trabajo, es necesario que esto se haga bajo estándares que garanticen la calidad, y permitan la interoperabilidad de los datos. Lo anterior constituye un importante desafío en el país dado que el propósito principal de la apertura de datos está orientado al uso y explotación de los datos abiertos para la generación de valor económico y social, dicha reutilización solo es posible si existe calidad en los datos publicados.

El objetivo de la presente guía es proporcionar orientaciones y buenas prácticas para el desarrollo de estrategias de apertura de datos, que cumplan con los principales estándares de calidad e interoperabilidad de datos abiertos de los países mejor calificados en índices internacionales como el Open Data Barometer[1], el Global Open Data Index[2] y el Our Data Index de la OCDE[3]. Además de lo establecido en la Hoja de Ruta de Datos Abiertos Estratégicos para el Estado Colombiano, las guías del dominio de información del Marco de Referencia del MinTIC[4], el Modelo de Calidad de Datos ISO 25012[5], la Carta Internacional de Datos Abiertos[6] y el Marco de Interoperabilidad[7]. Con base en este análisis se definieron 17 criterios fundamentales para la apertura de datos de calidad por parte del Gobierno de Colombia, los cuales son: confidencialidad, relevancia, actualidad, trazabilidad, conformidad, exactitud, completitud, consistencia, precisión, portabilidad, credibilidad, comprensibilidad, accesibilidad, eficiencia, recuperabilidad, disponibilidad y unicidad.

---

1 https://opendatabarometer.org/
2 https://index.okfn.org/
3 https://www.oecd-ilibrary.org/governance/open-useful-and-re-usable-data-ourdata-index-2019_45f6de2d-en
4 https://www.mintic.gov.co/arquitecturati/630/articles-9253_recurso_pdf.pdf
5 https://iso25000.com/index.php/normas-iso-25000/iso-25012
6 https://opendatacharter.net/principles-es/
7 http://lenguaje.mintic.gov.co/marco-de-interoperabilidad

Los criterios enunciados están orientados a la estandarización del proceso de apertura de datos, la cual permite utilizar y reutilizar la información para que pueda ser interoperable por diferentes entidades a nivel nacional e internacional.

El proceso de apertura de datos por parte de las diferentes instituciones a nivel nacional, regional y local; deberá contener los 17 criterios de calidad seleccionados a manera de estándar. Con la aplicación de estos criterios se busca lograr mejores calificaciones en las evaluaciones nacionales e internacionales de calidad, e interoperabilidad de datos; replicar las mejores prácticas de los países líderes en datos abiertos y contar con un modelo de calidad estándar como la ISO 25012, con el fin de generar gran valor y aprovechamiento por parte de los diferentes usuarios, así como una mejor gestión y gobernanza en los sistemas de información que los contienen, incluyendo la plataforma de datos abiertos del Gobierno de Colombia (www.datos.gov.co).

# 01. ¿Por qué es importante publicar datos abiertos de calidad?

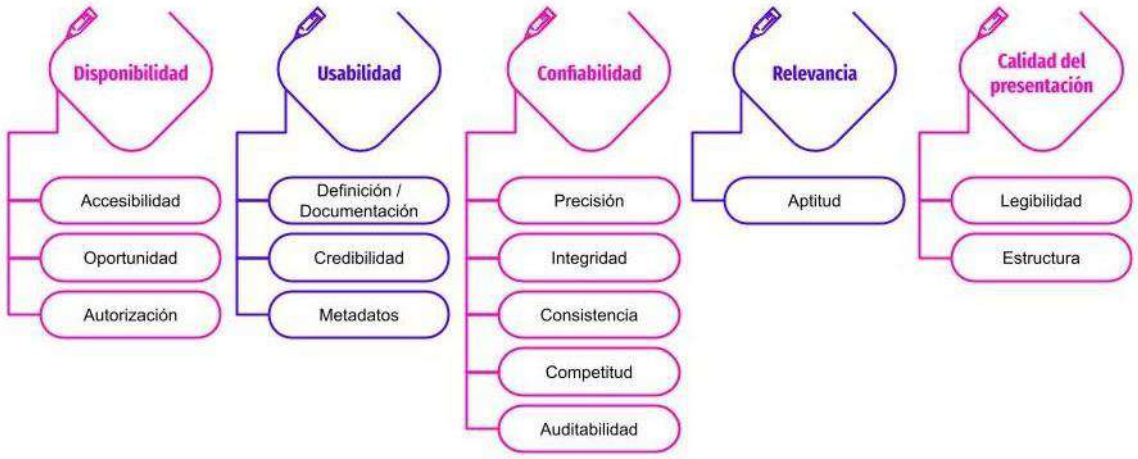
A partir década de 1950 los investigadores comenzaron a estudiar sobre los problemas de calidad, especialmente en los productos, con lo cual aparecieron una serie de definiciones que fueron publicadas. Algunas de ellas fueron: la calidad, entendida como "grado en el que un conjunto de características inherentes cumple con los requisitos" (Norma ISO 9000, 2000); "Aptitud para el uso" (Strong, 1996); "Conformidad con los requisitos" (Crosby, 1988). Más tarde, con el rápido desarrollo de las tecnologías de la información, la investigación se dirigió al estudio de la calidad de los datos.

La investigación sobre la calidad de los datos comenzó en el extranjero en la década de 1990, y muchos académicos propusieron diferentes definiciones de calidad de los datos y métodos de división de dimensiones de la calidad. El Grupo de Gestión de Calidad de Datos Total de la Universidad del MIT, dirigido por el profesor Richard Y. Wang, realizó investigaciones a fondo en el área de la Calidad de los Datos, definiéndola como la "aptitud para el uso" (Wang & Strong, 1996) y propuso que el juicio de la calidad de los datos dependía de los consumidores de estos. Así mismo, definieron la "dimensión de calidad de datos" como un conjunto de atributos que representan un único aspecto.

Con base en lo anterior, en los últimos años se han realizado grandes esfuerzos para la estandarización de los datos abiertos, teniendo como propósito, la reutilización de la información a través de estándares y la interoperabilidad de los datos por parte de cualquier usuario.

A continuación, se presentan los criterios que debe contener un conjunto de datos para que pueda ser de calidad e interoperable, definidos por el estándar universal de la calidad de los datos de 2 capas.

**Ilustración 1.** ***Estándar universal de calidad de dos capas***

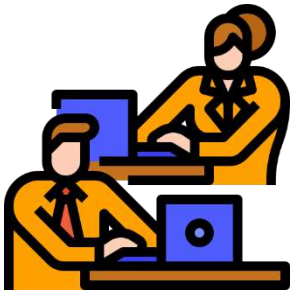



***Fuente****. Tomado de Los desafíos de la calidad de los datos y la evaluación de la calidad de los datos en la era del Big Data, por Cai. L. & Yangyon Z., 2015, Data Science Journal [8] (Cai, 2015)*

Los datos abiertos de alta calidad son una condición previa para analizarlos, reutilizarlos y garantizar el valor de los datos. Actualmente, no todos los conjuntos de datos proporcionados y publicados por las entidades del Estado en el portal de Datos Abiertos del Gobierno colombiano (www.datos.gov.co), tienen la misma calidad. La calidad de los conjuntos de datos determina qué tanto se puede confiar en ellos para reutilizarlos y transformarlos en productos, servicios o nuevas aplicaciones. Por lo anterior, es importante evaluar la calidad de los datos para verificar su exactitud, su confiabilidad y su aptitud para el uso; es decir la capacidad de los datos de prestar adecuadamente una función o servicio. En conclusión, los datos deben permitir transformarse, georreferenciar, analizar, reutilizar, visualizar y agregar sin necesidad de depurarse previamente.

[8] Cai. L. & Yangyon Z. (22 de mayo de 2015). *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.* Data Science Journal. *https://datascience.codata.org/articles/10.5334/dsj-2015-002/*

## 02. ¿Cómo mejorar la calidad de los datos?



La Guía para el uso y aprovechamiento de datos abiertos en Colombia[9], precisa que una estrategia de apertura de datos va más allá de la preparación y publicación de los datos por parte de las entidades; también involucra acciones concretas para promover el uso de los datos publicados, realizar seguimiento e identificar el valor agregado que se está generando con su utilización. En la segunda fase del proceso de apertura y uso de los datos, "monitorear la calidad y el uso", en la cual se centra esta guía, se propone analizar y presentar un paso a paso, para el cumplimiento de los criterios definidos para estandarizar la calidad y mejorar la interoperabilidad de los datos abiertos publicados, o a publicar en el portal de Datos Abiertos del Gobierno Colombiano; los cuales se detallan en el capítulo "4. Criterios de Calidad e Interoperabilidad" de esta guía.

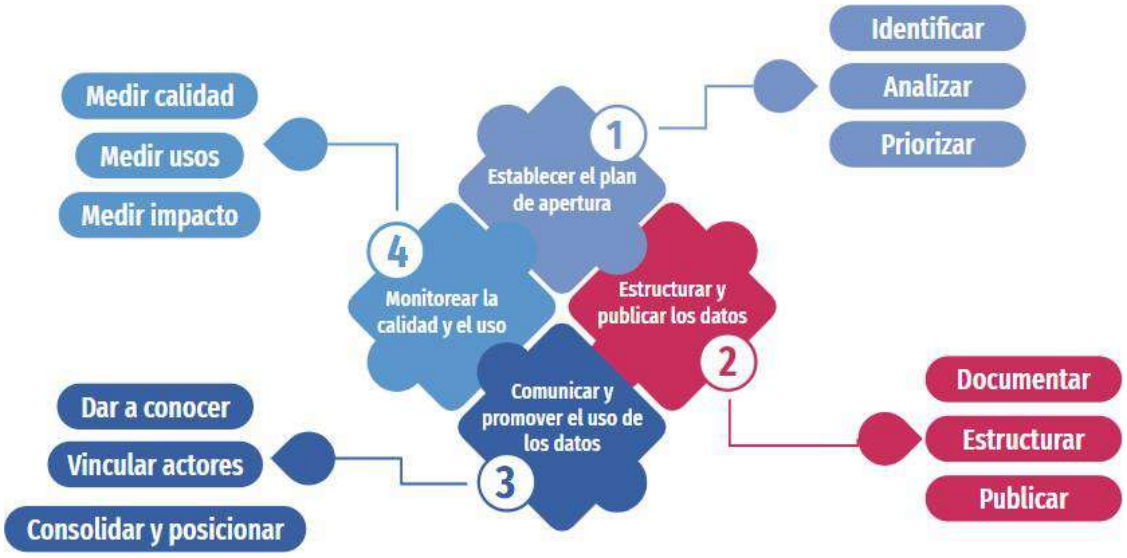
**RECUERDE**

Una estrategia de datos abiertos va más allá de la preparación y publicación de los datos por parte de las entidades.

También involucra acciones concretas para promover el uso de los datos publicados, realizar seguimiento e identificar el valor agregado que se está generando con su utilización.

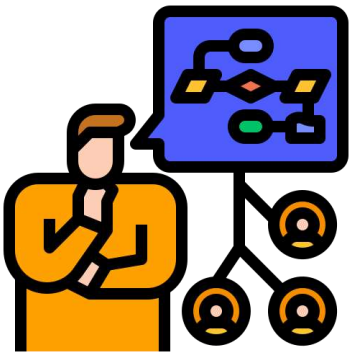
[9]https://herramientas.datos.gov.co/sites/default/files/Guia%20de%20Datos%20Abiertos%20de%20Colombia.pdf

**Ilustración 2. Ciclo de vida de los datos**




***Fuente.*** *Guía de uso y aprovechamiento de datos en Colombia.*

# 03. Conozca el Marco de Interoperabilidad



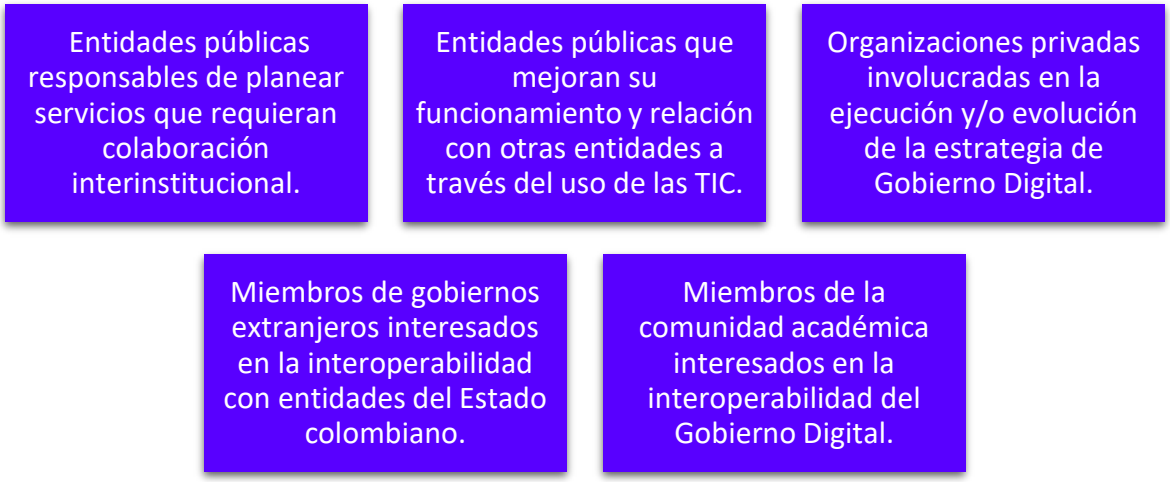
El Marco de interoperabilidad de Gobierno Digital surge con el propósito de contribuir en la entrega de servicios digitales, de manera completa, adecuada, minimizando los pasos y evitando el desplazamiento del ciudadano a diversas entidades para obtener la información necesaria de una entidad y acceder así a sus derechos y obligaciones con el Estado. La interoperabilidad permite fortalecer la visión de unidad del Estado, al tener una mayor capacidad de comunicación, entrega y uso de servicios digitales de valor para mejorar la calidad de vida de los ciudadanos. (Ministerio de Tecnologías de la Información y las Comunicaciones, 2019).

El Marco de Interoperabilidad es la estructura de trabajo común donde se alinean los conceptos y criterios que guían el intercambio de información. Define el conjunto de principios, recomendaciones y directrices que orientan los esfuerzos políticos, legales, organizacionales, semánticos y técnicos de las entidades, con el fin de facilitar el intercambio seguro y eficiente de información. Además, se desarrolla un modelo de madurez y un conjunto de principios, lineamientos, recomendaciones, protocolos, estándares y guías metodológicas, necesarias para que las entidades compartan información a través de servicios de intercambio de información de forma eficiente y segura, con el propósito de facilitar la prestación de sus servicios a ciudadanos, empresas y otras entidades públicas en Colombia. (Ministerio de Tecnologías de la Información y las Comunicaciones, 2019).

Los objetivos del marco de interoperabilidad son:

- Apoyar a las entidades públicas en sus esfuerzos por diseñar y ofrecer trámites y servicios en línea a otras entidades públicas, ciudadanos y empresas que, en la medida de lo posible, sean digitales por defecto, es decir, que proporcionen servicios y datos preferentemente a través de medios digitales, siendo accesibles para todas las entidades, los ciudadanos y que permitan la reutilización, participación, acceso y transparencia.
- Proporcionar orientación a las entidades públicas sobre el diseño y la actualización de los mecanismos de interoperabilidad, sus políticas, estrategias y directrices, así como la visión nacional que se promueve en interoperabilidad.
- Contribuir al fortalecimiento de mecanismos de interoperabilidad en las entidades públicas para la prestación de trámites y servicios en línea.

El Marco de Interoperabilidad es genérico y aplicable a todas las entidades públicas y privadas en Colombia, el marco establece las condiciones básicas que se deben considerar para alcanzar la interoperabilidad tanto a nivel local, interinstitucional, sectorial, nacional o internacional y orienta a todos los involucrados en definir, diseñar, desarrollar y entregar servicios de intercambio de información, como son:




El Marco de Interoperabilidad conceptualmente se compone de cuatro (4) dominios, cada uno de ellos trata un tema específico de la interoperabilidad, amparados por un conjunto de principios que definen y orientan la concepción, diseño e implementación de los servicios de intercambio de información con un enfoque centrado en el ciudadano, seguro y de confianza. Transversalmente se presentan los lineamientos a cumplir, el marco se completa con un esquema de gobierno y un modelo de madurez que permite determinar el estado actual y proyectado de interoperabilidad para las entidades.

**Ilustración 3. Modelo conceptual del Marco de Interoperabilidad**


***Fuente.*** *Ministerio TIC.*

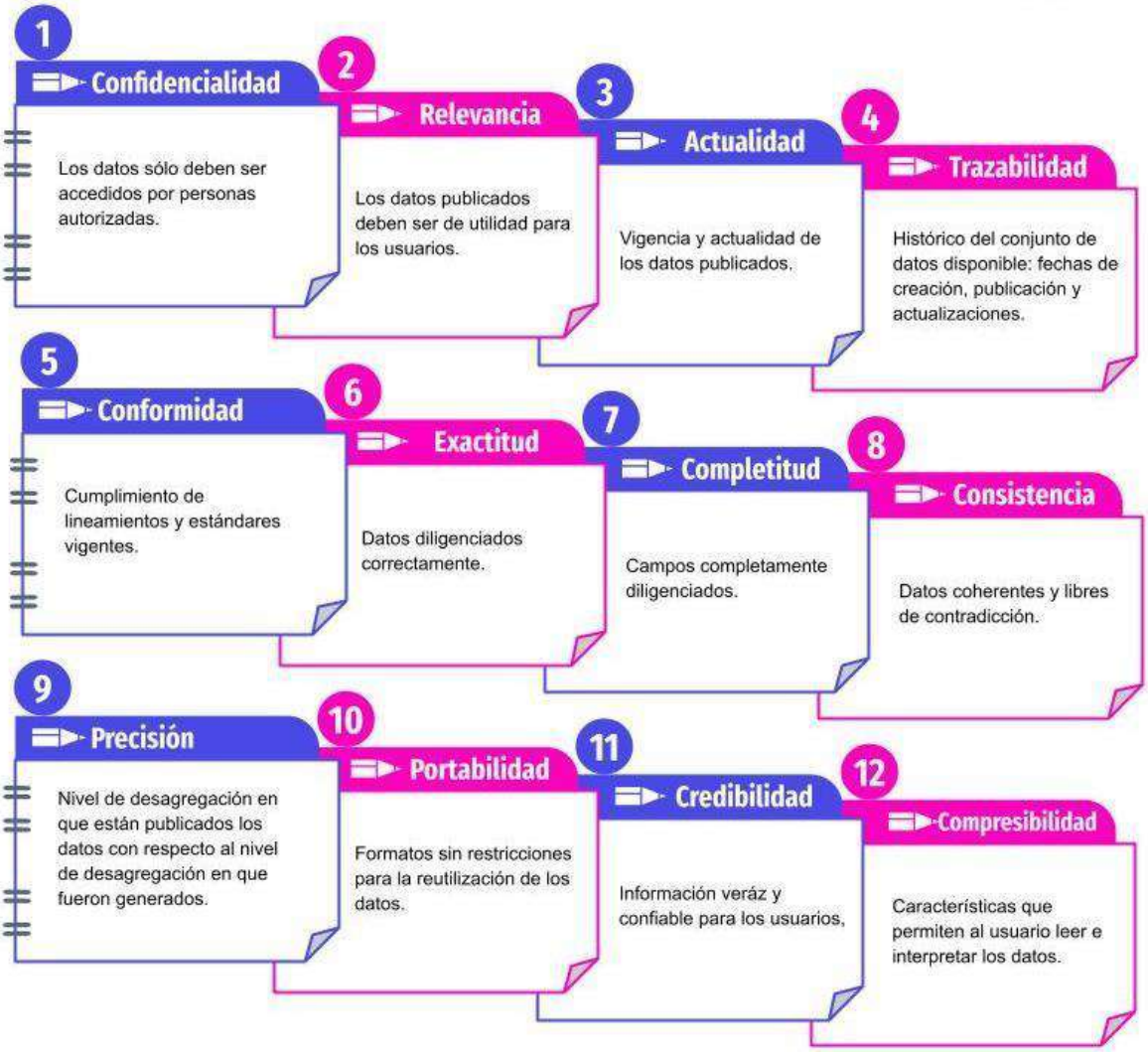
El Marco de Interoperabilidad contempla una iteración holística, es decir, desde múltiples iteraciones llamada dominios de interoperabilidad, los cuales son:

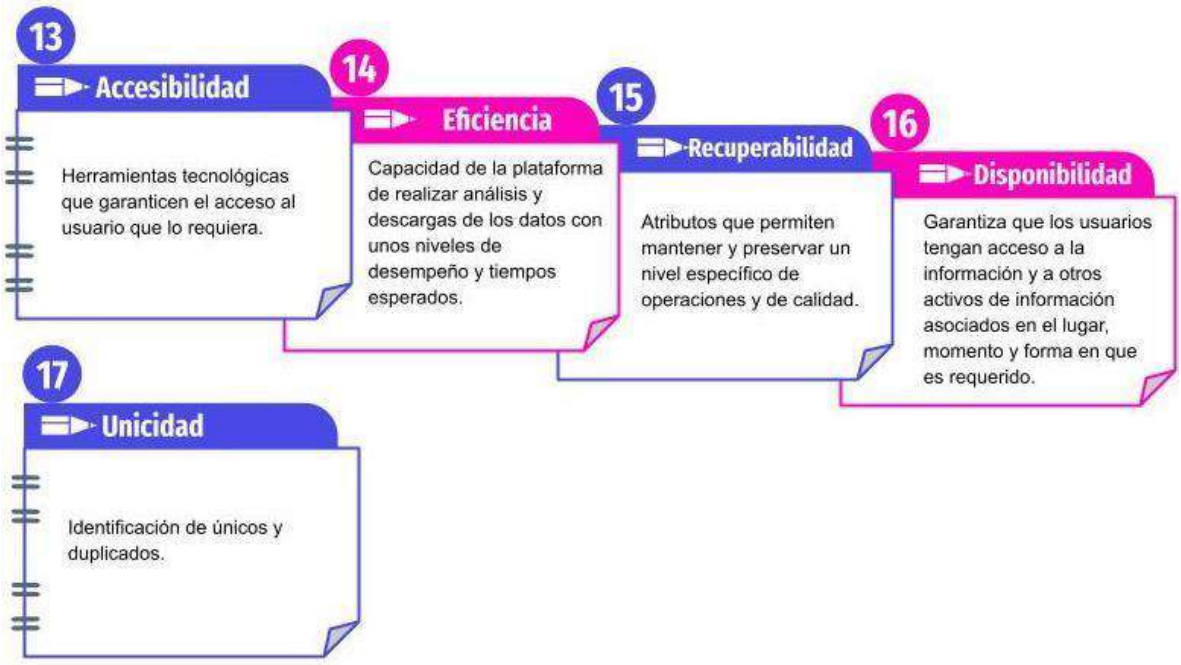
- **Político-legal:** Corresponde a la disposición de un conjunto de políticas y normas que permiten el intercambio de información.
- **Organizacional:** Modo en que las misiones, políticas, procesos de negocio y mecanismos de prestación de estos procesos interactúan con aquellos de otras entidades a través del intercambio de información.
- **Semántico:** Permite garantizar que, en el momento de intercambiar datos, el significado de la información sea el mismo para todos los actores involucrados.
- **Técnico:** Hace referencia a las condiciones que se deben cumplir en aplicaciones e infraestructura para conectar los sistemas de información, con el propósito de intercambiar información a través de servicios (Ministerio de Tecnologías de la Información y las Comunicaciones, 2010).

Con la adecuada implementación de los criterios de calidad desarrollados en esta guía, los dominios semánticos y técnicos se convierten en una consecuencia natural proveyendo las condiciones físicas necesarias y que servirán de base para la ejecución de los dominios político-legal y organizacional del marco de interoperabilidad.

# 04. Criterios de calidad e interoperabilidad

**Ilustración 4. Criterios de calidad e interoperabilidad**





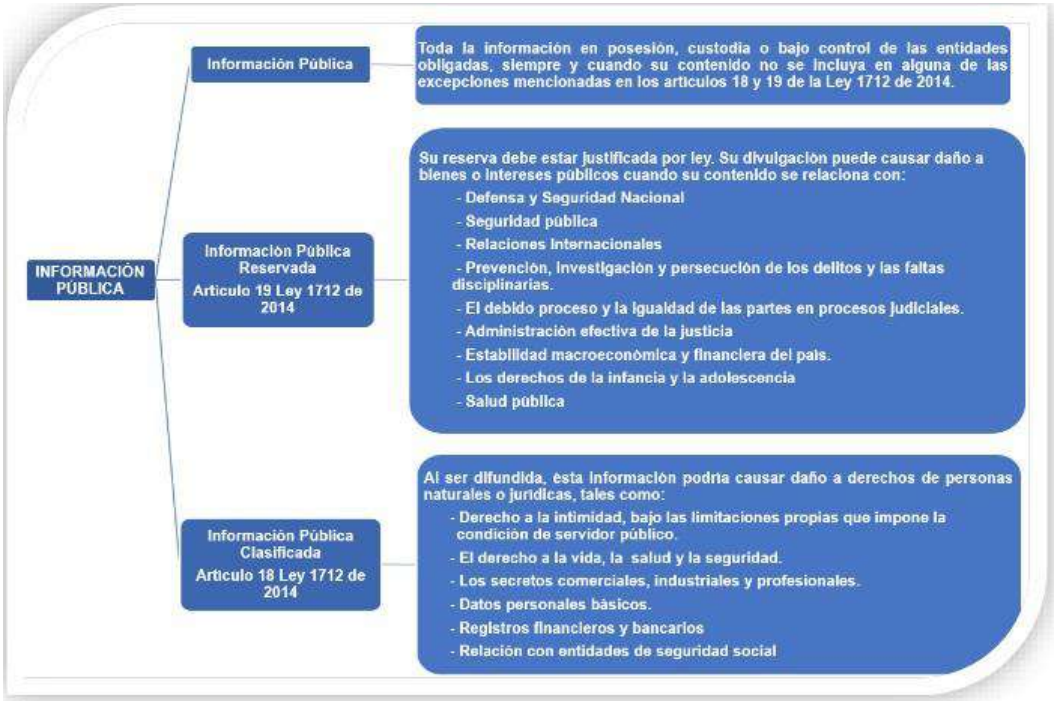

***Fuente.*** *Elaboración propia.*

## *4.1 Criterio de confidencialidad*

Este criterio establece que los datos solo deben ser accedidos por personas autorizadas para proteger adecuadamente la información reservada y clasificada. En Colombia los criterios para asignar la calificación a las categorías de información son aquellos indicados por la Ley de Transparencia y del Derecho de Acceso a la Información (ley 1712 de 2014), la cual fue reglamentada con el Decreto 103 de 2015, en donde se establece que para la publicación de datos abiertos se debe tener en cuenta el título relativo a las “excepciones de acceso a la información”, el cual resalta la procedencia de limitar el acceso a la Información Pública Clasificada y la Información Pública Reservada.

A continuación, se dan a conocer las categorías de la información que describe la Ley y que deben ser tenidos en cuenta para validar el criterio de confidencialidad de la información:

**Ilustración 5. Calificación de la información de acuerdo con sus niveles de seguridad**


***Fuente.*** *Tomado de la Guía para la calificación de acceso a la información producida por el DAPRE, G-GD-02-calificacion-informacion (Ministerio de Tecnologías de la Información y las comunicaciones, 2019).*

El acceso a la información pública exige un ejercicio constante de ponderación entre el derecho a la información y el derecho a la protección de los datos personales, en este contexto, surge la anonimización.

## ¿Qué es anonimizar la información?

El procesamiento y análisis de grandes volúmenes de información puede ofrecer cuantiosos beneficios a la sociedad, siempre y cuando se respete la privacidad de cada individuo o entidad y la protección de sus datos personales. Por tal motivo, se hace necesario establecer e implementar procesos de anonimización de los datos antes de ser disponibles al público en el portal de Datos Abiertos del Gobierno colombiano. El proceso de anonimizar permite identificar y ocultar la información sensible garantizando la divulgación y acceso de la información a los usuarios, sin vulnerar los derechos a la protección de los datos de las personas y entidades.

Para que la anonimización sea exitosa, se debe garantizar que no sea posible identificar de manera directa o indirecta a individuos o entidades. Cabe resaltar, que el avance tecnológico y la información disponible en distintos medios y en particular, a través de Internet, dificultan un anonimato absoluto, principalmente en el transcurrir del tiempo. Aun así, la implementación de procesos de anonimización ofrece mayor seguridad para conservar el anonimato y la privacidad de las personas. Por tal motivo es fundamental que las entidades realicen de una forma adecuada el proceso de anonimización, pues de no ser así, esto podría generar una reducción de la confianza de quienes gestionan la información, y afectar considerablemente la calidad de los datos.

Una vez identificada la información clasificada y/o reservada, se debe tener en cuenta los siguientes elementos para anonimizar los datos a publicar en formatos abiertos:

### Pre-anonimización

Ningún conjunto de datos deberá permitir la identificación directa o indirecta de una persona. Por tal motivo, es recomendable eliminar la información personal con la que se va a trabajar, y tener un especial cuidado con aquellos datos que contienen información sensible, según la normatividad colombiana.

### Ocultamiento, supresión y seudo-anonimización

Se debe identificar en el conjunto de datos la información sensible para su eliminación o sustitución. A continuación, algunos ejemplos de datos sensibles:

- Nombres.
- Fecha de nacimiento, fecha de constitución en Cámara de
- Comercio, (excepto el año).
- Números de teléfono y fax.
- Números de identificación: cédula de ciudadanía, pasaporte,

- tarjeta de identidad.
- Números asociados a la seguridad social, licencias de conducción,
- Número de Identificación Tributaria (NIT), Registro Único
- Tributario (RUT), Registro Único de Proponentes (RUP).
- Registro Único Empresarial (RUES).
- Direcciones de correo electrónico.
- Números de cuentas bancarias.
- Identificadores del vehículo, placa, entre otros.
- Identificadores de dispositivos móviles y números de serie.
- Direcciones de IP.
- Cualquier otro número único de identificación.
- Dirección de domicilio.

Una vez identificadas estas variables, se deben eliminar los campos del conjunto de datos, o reemplazarlos por algún seudónimo.

## *4.2* Criterio de relevancia

Determina los temas importantes, destacados, significativos o con mayor demanda por parte de los usuarios, que se deben publicar en la apertura de datos. Para esto, el Gobierno nacional define en su mapa de ruta algunas entidades, temáticas y criterios de priorización de acuerdo con la Política de Gobierno Digital de MinTIC:

- Temáticas priorizadas en el mapa de ruta.
  - Ordenamiento territorial.
  - Cadena productiva del agro.
  - Calidad y cobertura educativa.
  - Prestación de servicios de salud, salud pública y gestión del riesgo en salud.
  - Movilidad.
  - Seguridad ciudadana.

- Algunos índices internacionales como el Open Data Barometer, el Global Open Data Index y el Our Data Index definen unas temáticas a priorizar, las cuales son:

  - Temáticas priorizadas en el Open Data Barometer.
    - Datos geográficos.
    - Datos de propiedad de la tierra,
    - Microdatos de censos,
    - Presupuesto gubernamental detallado,
    - Gastos gubernamentales,
    - Registro de la compañía,
    - Legislación.
    - Horarios de transporte público.

- Comercio internacional.
- Salud.
- Educación primaria o secundaria.
- Estadísticas de criminalidad.
- Estadísticas del entorno nacional.
- Resultados electorales.
- Contratación pública.

- Temáticas priorizadas en el Global Open Data Index:
  - Presupuesto.
  - Gastos.
  - Adquisiciones.
  - Resultados electorales.
  - Registro de la compañía.
  - Propiedad de la tierra.
  - Mapas nacionales.
  - Límites administrativos.
  - Ubicaciones.
  - Estadísticas nacionales.
  - Proyecto de legislación.
  - Ley nacional.
  - Calidad del aire.
  - Calidad del agua.
  - Predicciones meteorológicas (Discuss.okfn.org, s.f.).

- Temáticas priorizadas en el Our Data Index:
  - Negocios.
  - Registros.
  - Patentes y marcas.
  - Licitaciones públicas.
  - Información geográfica.
  - Legal.
  - Meteorología.
  - Datos sociales.
  - Transporte.

## *4.3* Criterio de actualidad

La actualidad representa la vigencia y validez de los datos. Existen conjuntos de datos que se encuentran desactualizados en sus fechas, cifras periódicas o metadatos; por lo que es fundamental mantener la información actualizada constantemente.

Se debe aplicar la prueba de actualidad a la última versión del conjunto de datos, pues es importante que permanezcan publicados los históricos de datos para la trazabilidad y análisis temporales. Adicional, se debe evaluar que los datos estén siendo publicados oportunamente en relación con la fecha en la que fueron producidos.

La prueba de actualidad permite identificar si el conjunto de datos tiene la edad correcta en un contexto de uso genérico. Dentro de la metadata de cada conjunto de datos se encuentran 2 campos para el análisis de este criterio, fecha y frecuencia de actualización.

Para obtener el puntaje de prueba de actualidad, se debe realizar la diferencia entre la fecha actual menos la de actualización, si la diferencia es mayor a la frecuencia de actualización, el conjunto de datos tendrá 0 puntos, lo cual indica que el criterio de actualidad no es válido; si la diferencia es menor tendrá 10 puntos, lo cual indica que el criterio de actualidad es válido.

**Ecuaciones**

**Puntaje**: Dos ecuaciones

FechaActual – FechaActualizacion > FrecuenciaActualizacion = 0 puntos

FechaActual – FechaActualizacion <= FrecuenciaActualizacion = 10 puntos

En el momento de publicar un conjunto de datos se debe tener en cuenta las siguientes consideraciones:

- Hay datos que pierden su vigencia y utilidad rápidamente, por eso se debe establecer un límite de publicación más ajustado, según las capacidades de la institución.
- Definir un cronograma de publicación según la periodicidad que se estipula en los metadatos, con el fin de que el conjunto de datos se conserve actualizado.
- Implementar procesos de automatización de apertura de datos a través de ETL (Extracción, Transformación y Carga), para ello, revise el manual de creación de ETL.

## *4.4* Criterio de trazabilidad

La trazabilidad permite determinar el flujo de procesos y acciones que se han generado sobre un conjunto de datos a través del tiempo. Este criterio es importante para la interoperabilidad de los conjuntos de datos entre las diferentes instituciones, y para la reutilización de la información por parte de los usuarios.

La trazabilidad es el grado en el que los datos tienen atributos que proporcionan un camino de acceso auditado a estos, o cualquier otro cambio realizado sobre los datos en un contexto de uso específico (ISO/IEC 25012).

La forma de medir el criterio de trazabilidad se determina así:

**Puntaje**

El puntaje dependerá de los metadatos que se ingresaron de los conjuntos de datos durante la apertura; si se ingresa toda la información, el puntaje será de 10, y 0 para el caso de que no se tenga ningún campo.

**Ecuación:**

**Puntaje:** (1 - (campos con información / campos de metadata) * 10)

Se debe garantizar la trazabilidad publicando todos los históricos de los conjuntos de datos que se tengan disponibles.

A manera de ejemplo, se recomienda consolidar todos los datos dispersos, que están agrupados por vigencias o años en un solo conjunto de datos, en donde se incluya una columna con el año de publicación correspondiente.

Ejemplo: conjunto de datos del SECOP I y II, el cual consolida históricamente todos los procesos de compra pública del Estado.

Adicionalmente, la plataforma de Datos Abiertos ofrece la posibilidad de conocer las últimas fechas de actualización, tanto de los datos, como de los metadatos, lo que permite conocer si están actualizados de acuerdo con la frecuencia definida para cada conjunto.

## *4.5* Criterio de conformidad

La conformidad hace referencia al cumplimiento de lineamientos y estándares vigentes en el proceso de apertura de datos. En términos de calidad de datos, el criterio de conformidad está dividido en 2 componentes: metadatos y Lenguaje Común de Intercambio.

La conformidad es el grado en el que los datos tienen atributos que se adhieren a estándares, convenciones o normativas vigentes y reglas similares referentes a la calidad de datos en un conjunto y en un contexto genérico.

Este criterio será evaluado a partir del formato de datos. Este tipo de validaciones analiza los valores de los campos y verifica si son correctos.

**Puntaje**

Puntaje de 10**,** si no se encuentra columna a evaluar (correo electrónico, número de teléfono, nombres, coordenadas, direcciones).

Fórmula para columna = 1 - (Registros no válidos) / (Número de registros en la columna) *10

Se sacará un puntaje por columna, el resultado final es el promedio de estos puntajes.

Fórmula = (Puntaje columna1 + puntaje columna 2 + …. + puntaje columna n) /n

En dado caso el valor es decimal, se aproxima a su parte entera más cercana, por ejemplo, si el puntaje es 9,4 el puntaje será 9.

a. Metadatos:

como definición simple se indica que son **"**datos acerca de los datos". El tipo de información que se puede clasificar como Metadatos es de amplio rango. Los Metadatos incluyen información sobre procesos técnicos y de negocio, reglas y restricciones de datos y estructuras de datos lógicas y físicas. Describen los datos en sí (por ejemplo, bases de datos, elementos de datos, modelos de datos), los conceptos que los datos representan (por ejemplo, procesos de negocio, sistemas de aplicación, código de software, infraestructura tecnológica) y las conexiones (relaciones) entre los datos y conceptos. Los Metadatos ayudan a una organización a comprender sus datos, sus sistemas y sus flujos de trabajo. Permiten la evaluación de la calidad de los datos y son parte integral de la gestión de bases de datos y otras (DAMA DMBOK2).

La generación de metadatos no solo es aplicable a la información digital, también debe emplearse en cualquier conjunto de datos independientemente del soporte en el cual se encuentren, ya que ello puede facilitar su localización, y así, agregarle un valor añadido a la información histórica con la que cuenta una entidad. (Estado Peruano, s.f.).

En la siguiente imagen se detallan los metadatos requeridos durante el proceso de publicación de datos abiertos en www.datos.gov.co:

**Ilustración 6. Metadatos del conjunto de datos**


Para acceder a más información acerca del diligenciamiento de los metadatos, remítase al portal de datos abiertos de Colombia: www.datos.gov.co, en la sección de herramientas y observe los videos sobre cómo publicar conjuntos de datos: http://herramientas.datos.gov.co/es/content/publicar-conjuntos-de-datos

***Fuente.*** *Tomado de Portal Datos abiertos. Herramientas*[10]

b. Lenguaje Común de Intercambio de Información

El intercambio de información entre organizaciones puede realizarse a través de diferentes implementaciones técnicas de servicios de interoperabilidad. La interoperabilidad se define como “capacidad de dos o más sistemas o componentes para intercambiar información y utilizar la información intercambiada.” (ISO/25010).

Según el Marco para la Interoperabilidad de la Política de Gobierno Digital, esta se define, como “el ejercicio de colaboración entre organizaciones para intercambiar información y conocimiento en el marco de sus procesos de negocio, con el propósito de facilitar la entrega de servicios en línea a ciudadanos, empresas y a otras entidades”.

[10] *Ministerio de Tecnologías de la Información y las Comunicaciones.2021. Herramientas. https://herramientas.datos.gov.co/herramientas/publicar-conjuntos-de-datos?title=etl*

## *4.6* Criterio de exactitud

La exactitud es el gado en el que los datos representan correctamente el verdadero valor del atributo deseado de un concepto, o evento en un contexto de uso específico (ISO/IEC 25012).

Tiene dos principales aspectos:

- *Exactitud Sintáctica:* Cercanía de los valores de los datos, a un conjunto de valores definidos en un dominio considerado sintácticamente correcto.

- *Exactitud Semántica:* Cercanía de los valores de los datos, a un conjunto de valores definidos en un dominio considerado semánticamente correcto.

  El criterio de "exactitud" al ser estrictamente necesario un contexto de uso específico, se hace imposible el cálculo de este, por ende, se parte del mismo principio de la buena fe, y se otorgará un puntaje siempre de 10.

  **Ecuación:** Sin ecuación
  **Puntaje:** 10

Este criterio evalúa que los datos abiertos de los usuarios estén diligenciados correctamente y que su vez brinden información exacta en sus conjuntos de datos.

Se debe asegurar de que los datos estén diligenciados correctamente para que cumplan con el criterio de exactitud. A continuación, se describen los tipos de datos aportados actualmente por el portal de Datos Abiertos de Colombia. [11]

- Datos numéricos: Son aquellos que contienen números, dinero y porcentajes. Se utilizan para representar cantidades o medidas dentro de un conjunto de datos.

- Fecha y hora: Los campos de fecha y hora deben de ser diligenciados mediante el formato de la norma ISO 8601. Se deben tramitar teniendo en cuenta los lineamientos.[12]

- Ubicación geográfica: Se define utilizando la notación de grados decimales para latitud y longitud en el siguiente formato:
  (xx.xxxx, -yy.yyyy) donde "xx.xxxx" es la latitud y "-yy.yyyy" es la longitud

[11] *Andrew (29 de abril de 2021). Importing, Data Types, and You! Tyler Technologies. https://support.socrata.com/hc/en-us/articles/202949918-Importing-Data-Types-and-You- [13]*
[12] *ISO (s.f.) ISO 8601 Date and Time Format. Popular Standars. https://www.iso.org/iso-8601-date-and-time-format.html*

- Enlaces URL: Solo se aceptan 3 formatos de URL: FTP, HTTP Y HTTPs. No se aceptan guiones bajos en las URL. Se debe ingresar los enlaces a páginas web en uno de los siguientes formatos de entrada:

  1. <a href="http://www.socrata.com/">Socrata</a>
  2. http://www.socrata.com/
  3. Socrata (http://www.socrata.com)

- Email: Todo conjunto de datos debe contener un email de la persona o la entidad encargada de subir la información en la Web. Se debe utilizar uno de los siguientes formatos de entrada para ingresar los datos de correos electrónicos:

  1. <a href="mailto:sam.gibson@socrata.com">Sam Gibson</a>
  2. sam.gibson@socrata.com
  3. Sam Gibson <sam.gibson@socrata.com

- Lógico/ binario: Se utiliza para representar valores binarios y lógicos, tales como: 1/0 y Verdadero/Falso. En este campo se debe tener en cuenta 2 posibles valores que serán representados por uno y cero (1 y 0).

**Tabla 1. Uso del tipo de campo Lógico/ binario**

| Valor binario | Etiqueta 1 | Etiqueta 2 | Etiqueta 3 | Etiqueta 4 |
|---|---|---|---|---|
| 1 | Si | Verdadero | Positivo | Aumento |
| 0 | No | Falso | Negativo | Disminución |

***Fuente.*** *Elaboración propia.*

Es importante tener en cuenta las siguientes consideraciones a la hora de utilizar los diferentes tipos de datos:

- No utilizar más de un tipo de dato en la misma columna.
- En los campos numéricos, adoptar codificación numérica para los casos especiales, por ejemplo: -88: “No aplica”, -89: “No disponible”, -98: “No sabe”, -99: “No responde”, -77: “Otro”.
  Esto con el fin de no combinar tipo de datos de texto con datos numéricos.
- Si el contenido de un campo depende de una etiqueta recurrente, asegurarse de que esta se utilice de manera uniforme en todo el conjunto de datos. Por ejemplo, si se requiere usar la etiqueta SI/NO en un campo, no mezclar etiquetas como Sí, sí, No, no.

## *4.7* Criterio de completitud

La completitud hace referencia a datos completamente diligenciados. Para que un conjunto de datos cumpla a cabalidad con los criterios de calidad establecidos en la presente guía, es fundamental que los datos se presenten completos para el uso de los usuarios.

Grado en el que los datos asociados con una entidad tienen valores para todos los atributos esperados, e instancias de entidades relacionadas en un contexto de uso específico (ISO/IEC 25012). Este criterio se caracteriza principalmente en detectar nulos por cada columna, determinar el total de filas que se tiene para encontrar una relación entre nulos y valores esperados. Esta relación determinará el puntaje asignado a este criterio, que está en un rango entre 1 y 10, donde 10 quiere decir que los datos cumplen con este requerimiento.

Se analizará el número de nulos de cada columna, luego se halla la relación entre el número de nulos y el total de filas del conjunto de datos; con esta relación se obtendrá un valor entre 0 y 1, que representa el porcentaje de nulos de esa columna, es decir, cuando el valor es más cercano a 1, la columna tiene más nulos; si el resultado es 1, quiere decir que la columna está totalmente nula.

Luego de encontrar este valor para cada columna, se procede a normalizar para que arroje un valor entre 0 y 10; se suman estos valores y se divide por el número de columnas que se tengan, de este modo da un promedio de todas las columnas analizadas, y finalmente se obtiene la calificación final del criterio para determinado conjunto de datos.

**Ecuación:**

**Puntaje:** 10-(DatosNulos / CantidadDatos) *10)

Es importante recordar que, de acuerdo con las Políticas de Privacidad y Condiciones de Uso, la Dirección de Gobierno Digital podrá eliminar o cambiar los permisos de privacidad de los conjuntos de datos que presenten algunos de los siguientes errores:

1. Que el conjunto de datos tenga menos de 50 registros (a menos de que la información sea la única existe y/o disponible).
2. Que el conjunto de datos tenga en sus metadatos campos incompletos y/o vacíos.
3. Que el conjunto de datos tenga registros en blanco.
4. Que se carguen archivos en formatos cerrados que no corresponden a conjuntos de datos tales como: formatos .pdf, .doc, .gif, .jpg, .ppt, entre otros.
5. Enlaces a sistemas de información que no permitan la descarga directa de un conjunto de datos en formato abierto: csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI).

6. Que el conjunto de datos contenga datos personales publicados sin contar con la autorización de los titulares de acuerdo con la normatividad vigente: Ley 1581 de 2012 – Protección de Datos Personales.
7. Que el conjunto de datos incluya información clasificada y/o reservada de acuerdo con lo establecido en la Ley 1712 de 2014 - Transparencia y Acceso a la Información Pública.
8. Enlaces rotos de los metadatos y/o conjuntos de datos.
9. Que el conjunto de datos esté compuesto de una única columna.
10. Que el conjunto de datos no cuente con ninguna información en formato de columnas y/o filas.

S debe tener en cuenta las siguientes consideraciones a la hora de diligenciar un registro:

- En los diccionarios de datos o etiquetas de campo tipo texto, incluir siempre opciones como “No aplica”, “Otro”, “No sabe”, “No responde”, “No disponible”, según se requiera.

Esto con el fin de no dejar espacios vacíos cuando no se tenga la información.

- No publicar conjuntos de datos que contengan una sola columna.
- No publicar conjuntos de datos incompletos.
- Si existen datos sobre un mismo tema distribuidos en diferentes fuentes de información, consolidar un único conjunto de datos que represente la información lo más completa posible para los usuarios.
  Cuando los conjuntos de datos provienen de un modelo de datos relacional, tener en cuenta que cada tabla no es un conjunto de datos.

**Tabla 2. Error de completitud**

| ID | cod Dep | Departamento | Población 2017 | Población 2018 | Población 2019 | Población 2020 |
|---|---|---|---|---|---|---|
| 1 | 91 | Amazonas | 77948 | 78830 | 79739 | 80682 |
| 2 | 5 | Antioquia | 66131118 | 6691030 | 6768388 | 6845093 |
| 3 | 88 | Archipiélago de San Andres | 77759 | 78413 | 79060 | 79653 |
| ... | ... | ... | ... | ... | ... | ... |
| 30 | 76 | Valle del Cauca | 4708262 | | | 4853327 |
| 31 | 97 | Vaupés | 44500 | 44928 | 45367 | 45822 |

***Fuente.*** *Elaboración propia.*

## *4.8* Criterio de consistencia

Los datos son consistentes cuando están libres de contradicción y son coherentes respecto a otros datos en el mismo contexto de uso. Esto quiere decir, que se utilice la misma codificación de variables y etiquetas en todos los conjuntos de datos.

Ejemplo: Si se utiliza codificación como 1: Mujer, 2: Hombre, para la variable “Sexo”, utilizar siempre estas convenciones en todos los conjuntos de datos. Aplica también, y con mayor rigurosidad, para las codificaciones que están estandarizadas, como los códigos de la división político-administrativa del país.

**Tabla 3. Ejemplo de conjunto de datos con errores de consistencia**

| Id_ persona | Sexo | Edad |
|---|---|---|
| 1 | H | 12 |
| 2 | Mujer | 23 |
| 3 | Hombre | 43 |
| 4 | hombre | 15 |
| 5 | Mujer | 65 |

***Fuente***. Elaboración propia.

**Tabla 4. Ejemplo de conjunto de datos después de corregir errores de consistencia**

| Id_ persona | Sexo | Edad |
|---|---|---|
| 1 | Hombre | 12 |
| 2 | Mujer | 23 |
| 3 | Hombre | 43 |
| 4 | Hombre | 15 |
| 5 | Mujer | 65 |

***Fuente.*** *Elaboración propia.*

La consistencia puede también referirse a la consistencia de formato, por lo cual, es conveniente estandarizar la entrada de los datos donde se cumpla con reglas de contenido y formato.

La consistencia de los datos es cuando estos están libres de contradicción y son coherentes con otros, en un contexto de uso específico. Estos pueden ser analizados en datos que se refieren, tanto a una, como a varias entidades comparables.

Este criterio por su definición es necesario en un contexto específico. Se parte de que los datos están libres de contracción, ya que no se cuenta con un escenario de comparación, ni mucho menos con un contexto de uso específico. Se otorga puntaje siempre de 10.

**Ecuación:** Sin ecuación

**Puntaje:** 10

## *4.9* Criterio de precisión

Los conjuntos de datos se deben publicar con el más alto nivel de desagregación posible, es decir, que siempre desde la fuente primaria de la información.

Se debe evitar hacer agregaciones, ya que puede causar que se sesguen algunas dimensiones asociadas.

**Tabla 5. Criterio precisión**

| ID | Barrio | Sexo | Asistentes |
|---|---|---|---|
| 1 | B1 | Hombre | 10 |
| 2 | B1 | Mujer | 5 |
| 3 | B2 | Hombre | 12 |
| 4 | B2 | Mujer | 4 |
| 5 | B3 | Hombre | 5 |
| 6 | B3 | Mujer | 16 |

*a) Fuente primaria*

| ID | Barrio | Asistentes |
|---|---|---|
| 1 | B1 | 15 |
| 2 | B2 | 16 |
| 3 | B3 | 21 |

*b) Agregación por barrio*

| ID | Sexo | Asistentes |
|---|---|---|
| 1 | Hombre | 27 |
| 2 | Mujer | 25 |

*c) Agregación por sexo*

***Fuente.*** *Elaboración propia.*

En el ejemplo se puede observar la fuente primaria (a) del registro de la asistencia a un evento desagregado por sexo y barrio. Si se hace una agregación por barrio (b), se pierde toda la información relacionada a sexo, e igualmente, si se hace una agregación por sexo (c), se pierde toda la información relacionada con barrio.

## *4.10* Criterio de portabilidad

Se trata de la capacidad del conjunto de datos para ser transferido y adaptarse a diferentes formatos, es decir, ser reutilizado. Es el grado en el que los datos tienen atributos que les permiten ser instalados, reemplazados o eliminados de un sistema a otro, preservando el nivel de calidad en un contexto de uso específico.

Así como el criterio de disponibilidad, que es evaluado en contextos de uso específico, se tiene como alternativa calcular la portabilidad, a partir de 2 criterios anteriores: conformidad y completitud.

El rango del puntaje varía entre 0 y 10.

**Ecuación:**

**Puntaje:** (PuntajeConformidad*0,5) + (PuntajeCompletitud* 0,5)

Este criterio se ve afectado por los puntajes obtenidos anteriormente en conformidad y completitud, por ende, si ambos criterios son bajos, la portabilidad lo será. Este criterio tiene una puntuación de calidad de 9.05/10.

Se refiere además a la facilidad con la que el conjunto de datos se puede procesar, acceder y utilizar fácilmente descargándolo o consumiéndolo a través de interfaces de programación de aplicaciones. Para el caso del portal de Datos Abiertos de Colombia, se cuenta con la API de Socrata que soporta lenguajes de programación como Java, JavaScript, PHP, Ruby, Scala, Swift, .Net, entre otros. [13]

A continuación, se describen los formatos que actualmente soporta el portal de Datos Abiertos de Colombia (www.datos.gov.co) para su descarga. Entre ellos se encuentran: CSV, CSV for Excel, CSV for Excel (Europe), JSON, RDF, RSS, TSV for Excel y XML.

[13] Soda Developers (s.f.). *Libraries & SDKs.* https://dev.socrata.com/libraries/

**Tabla 6. Formatos para la publicación de datos abiertos**

| Formato | Independiente de una aplicación | Formato estructurado de datos | Editable | Observaciones |
|---|---|---|---|---|
| PDF | | | | No se recomienda para datos abiertos |
| DOC | | | X | No se recomienda para datos abiertos |
| TXT | X | | X | Se puede usar para datos abiertos |
| XLS | | X | X | Se puede usar para datos abiertos |
| CVS TVS | X | X | X | Recomendable para datos abiertos |
| JSON | X | X | X | Recomendable para datos abiertos |
| XML </> | X | X | X | Recomendable para datos abiertos |

***Fuente.*** *Elaboración propia.*

## *4.11* Criterio de credibilidad

Por definición de la norma ISO/IEC 25012[14] se tiene que la credibilidad es el grado en el que los datos tienen atributos que se consideran ciertos y creíbles en un contexto de uso específico. Esto se puede estimar con la presencia de información en los metadatos sobre fuentes de información, documentación, normatividad, origen y/o entidad publicadora de los datos.

La credibilidad incluye el concepto de autenticidad (la veracidad de los orígenes de datos, atribuciones, compromisos). A este criterio se le asignará 10 puntos.

[14] *ISO 25.000 (s.f.). ISO/IEC 25012. https://iso25000.com/index.php/normas-iso-25000/iso-25012?limit=5&start=5*

**Ecuación:** Sin ecuación
**Puntaje:** 10

También, en el catálogo de datos se evalúa la credibilidad en la adecuada creación de los usuarios que representan a las diferentes entidades del Estado que publican los datos dentro de la plataforma.

Se debe tener en cuenta que para garantizar el criterio de credibilidad se debe:

- Diligenciar rigurosamente los campos de los metadatos que dan cuenta sobre el origen o autor de los datos.
- Indicar un contacto en los metadatos (nombre y correo) el cual pueda resolver todas las inquietudes que los usuarios tengan sobre el conjunto de datos. Procurar establecer un correo general (no personal) que sea revisado constantemente.
- Crear los usuarios de carga de información de las diferentes instituciones dentro del portal de Datos Abiertos (www.datos.gov.co) Estos deben utilizar una imagen institucional y crear nombres de usuario alusivos a la institución.

## *4.12* Criterio de comprensibilidad

Los datos deben poseer atributos que permitan ser leídos e interpretados por los usuarios. En los metadatos se evalúan atributos como el nombre del conjunto de datos, descripción, diccionario de datos e índice de etiquetas. En el conjunto de datos se evalúa la claridad en los nombres de los campos y las etiquetas.

Comprensibilidad es el grado en el que los datos tienen atributos que permiten ser leídos e interpretados por los usuarios, y son expresados utilizando lenguajes, símbolos y unidades apropiados en un contexto de uso específico. Cierta información sobre la comprensibilidad puede ser expresada mediante metadatos.

Para el criterio es importante analizar en cada columna el tipo de dato, para que de este modo se pueda comprender la información que representa ese dato, por ejemplo, con la información de latitud y longitud, para este caso, se debe analizar la columna para determinar que cada registro tiene el respectivo formato, y además se encuentre dentro del rango permitido.

**Puntaje:**

Este criterio se evaluará con el promedio de los puntajes obtenidos por los criterios de conformidad y trazabilidad.

**Fórmula:** (Puntaje conformidad*0,5) + (puntaje trazabilidad* 0,5)

Este criterio es el promedio de conformidad y trazabilidad. Como estos 2 criterios tienen una puntuación alta (más de 9), el promedio se calcula como el número entero más próximo, por esta razón da como resultado 10 puntos.

Para garantizar el cumplimiento del criterio de comprensibilidad, se debe seguir los pasos que se detallan a continuación:

- Se debe asegurar de que el título del conjunto de datos sea comprensible y completo, respondiendo como mínimo a las siguientes preguntas: ¿qué?, ¿dónde?, ¿cuándo?, ¿vigencia? Si utiliza siglas, describa su significado.

- Tenga en cuenta que en la descripción del conjunto de datos deberá ampliar el contexto del título.

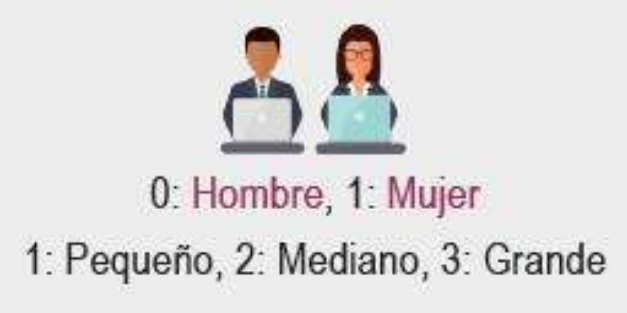
- Defina los nombres de los campos de forma clara y comprensible para cualquier usuario.

- Si los nombres de los campos se nombran con un sistema codificado, por ejemplo, P1, P2, P3 para referirse a las preguntas de una encuesta; incluya la definición de dichos códigos en los metadatos. Ejemplo:

  P1: ¿Cuántos hermanos tiene?
  P2: ¿Cuál es su ocupación?
  P3: ¿Cuál es su estado civil?

- Para etiquetas o valores de los campos codificados, incluya en los metadatos la definición de dichas etiquetas, ejemplo:

**Ilustración 7. Etiquetas o valores de campos codificados**




***Fuente.*** *Elaboración propia.*

- Siempre incluya en los metadatos una descripción narrativa de lo que significa cada campo. Ejemplo:

- *Población2018:* Se refiere a la cantidad de personas estimadas, según las proyecciones de población para cada uno de los departamentos de Colombia en el año 2018.

- Incluya una descripción clara de lo que significa cada registro del conjunto de datos. Ejemplo:

  Cada registro (o fila) de este conjunto de datos representa un departamento de Colombia y el valor de su proyección poblacional para los próximos 3 años.

## *4.13* Criterio de accesibilidad

Los criterios de accesibilidad, eficiencia, recuperabilidad y disponibilidad; son fundamentales para la calidad e interoperabilidad de los datos. Estos criterios son inherentes al funcionamiento de la plataforma de Datos Abiertos (www.datos.gov.co).

La accesibilidad es uno de los principios acordados en la Carta Internacional de Datos Abiertos, la cual pretende facilitar la disposición, búsqueda y descarga de datos para cualquier persona; mediante la distribución de los recursos tecnológicos necesarios para garantizarlo. Los datos deben ser fácilmente visibles y accesibles, poniéndose a disposición, sin barreras burocráticas o administrativas que pueden disuadir a las personas de acceder a los datos. Para ello, la plataforma de Datos Abiertos de Colombia permite:

a. Publicar los datos en un portal central para que los datos abiertos se puedan encontrar fácilmente y estén accesibles en un solo lugar.
b. Liberar los datos en formatos abiertos con el fin de asegurar que estos estén disponibles para el más amplio rango de usuarios; que puedan encontrarse, accederse y utilizarse; proporcionando los datos en múltiples formatos estandarizados, de modo que puedan procesarse por computadoras y utilizarse por personas.
c. Liberar los datos de manera gratuita, sujetos a una licencia abierta y sin restricciones.
d. Liberar los datos sin registro obligatorio, permitiendo a los usuarios escoger y descargarlos, sin requerir que se identifiquen.
e. Hay que asegurar que los datos puedan ser accesibles y usados eficazmente por el más amplio rango de usuarios.

## *4.14* Criterio de eficiencia

Este criterio está relacionado con la plataforma y su capacidad de análisis y descargas de los datos con unos niveles de desempeño y tiempos esperados.

La actual plataforma de Datos Abiertos garantiza que los datos puedan ser procesados y proporcionados con los niveles de rendimiento esperados. Además, cuenta con protocolos que permiten una fácil conexión de los datos dispuestos en la plataforma, para ser utilizados en otras aplicaciones directamente y en tiempo real, como pueden ser las herramientas de Business Intelligence, analítica e incluso Excel.

## *4.15* Criterio de recuperabilidad

Programas, software, plataformas digitales y aplicaciones que permiten mantener y preservar un nivel específico de operaciones y calidad, incluso en caso de fallos, manteniendo la recuperabilidad de los datos.

El portal de Datos Abiertos de Colombia se reserva el derecho de remover el contenido en cualquier momento, sin previo aviso, si este está relacionado con difamación, calumnia, injuria, falsedad, obscenidad y pornografía dentro del servicio.

Por lo tanto, es importante que cada entidad tenga copias de seguridad de toda la información subida al portal, y no lo tome como la fuente primaria de almacenamiento.

## *4.16* Criterio de disponibilidad

Este criterio de calidad garantiza que los usuarios autorizados tengan acceso a la información y a otros activos asociados en el lugar, momento y forma en que es requerido. Una plataforma segura debe mantener la información disponible para los usuarios.

Disponibilidad es el grado en que los datos tienen atributos que permiten ser obtenidos por usuarios y/o aplicaciones autorizadas, en un contexto de uso específico.

Aunque la definición no da margen para hacer cambios, debido a la especificidad que tiene cada conjunto de datos, se plantea modificar el cálculo de dicho criterio para otorgar el puntaje basado en dos condiciones: accesibilidad y el criterio de actualidad.

**Ecuación:**

La accesibilidad otorga 0 o 5 puntos basándose en 2 condiciones:

0 – Si el conjunto de datos es privado.
5 – Si el conjunto de datos es público.

Por otra parte, la actualidad otorga los puntos restantes para dar completitud a los 10 puntos.

0 – Si el conjunto de datos está desactualizado (puntaje de 0 obtenido en actualidad).
5 – Si el conjunto de datos está actualizado (puntaje de 10 obtenido en actualidad).

**Puntajes resultantes:**

0 = Conjunto de datos privado y desactualizado.
5 = Conjunto de datos privado y actualizado, o conjunto de datos público y desactualizado.
10 = Conjunto de datos público y actualizado.

Este criterio depende de la accesibilidad y actualidad, obteniendo en promedio una puntuación de calidad en el criterio de disponibilidad, de los conjuntos de datos de 6.41/10, donde la baja puntuación del criterio de actualidad tiende a arrastrar el criterio a 5, casi nunca es 0, ya que la mayoría de los criterios son públicos.

La plataforma de Datos Abiertos de Colombia cuenta con herramientas que garantizan que los sistemas de información de las entidades estatales puedan integrarse automatizando la publicación de sus datos en dicho portal, a través de procesos de: Extracción, Transformación y Carga (ETL)[15]. De igual manera, permite la consolidación e integración de datos de otros portales a través de procesos como la federación.[16]

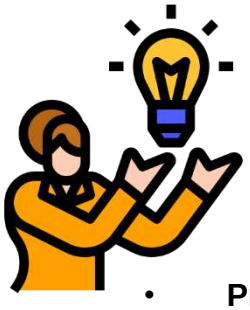
## *4.17* Criterio de unicidad

El criterio de unicidad permite obtener la cantidad de datos duplicados, sin tener un contexto de uso específico del conjunto de datos. Este, se calcula creando una llave única con todas las columnas del conjunto de datos.

**Ecuación: Puntaje:** 10 – ((FilasTotalesRepetidas / FilasTotalesConjuntoDatos) *10).

[15] *Portal de Datos Abiertos. (s.f.). Manual Creación de un ETL. https://herramientas.datos.gov.co/sites/default/files/CO_417_MANUAL_ETL_0.pdf*
[16] *Portal de Datos Abiertos. (s.f.). Manual Federación. https://herramientas.datos.gov.co/sites/default/files/CO_417_MANUAL_FEDERAR_0.pdf*

# 05. Principios de la calidad para la publicación de datos abiertos



Los datos son considerados abiertos si se cumple con los siguientes principios de calidad:

- **Completos:** Todos los datos públicos deben estar disponibles. Los datos públicos no contemplan datos privados ni limitaciones de seguridad o privilegios.
- **Primarios:** Los datos deben ser recolectados en la fuente de origen, con el nivel de desagregación más alto posible, no en forma agregada ni modificada.
- **Oportunos:** Los datos deben estar disponibles tan rápido como sea necesario, para garantizar el valor de estos.
- **Accesibles:** Los datos deben estar disponibles para el rango más amplio de usuarios, y para el rango más amplio de propósitos.
- **Procesables por máquinas:** Los datos deben estar estructurados razonablemente para permitir un procesamiento automático.
- **No discriminatorios:** Los datos deben estar disponibles para cualquier persona, sin requerir un registro.
- **No propietarios:** Los datos deben estar disponibles en un formato sobre el cual ninguna entidad tiene un control exclusivo.
- **Libres de licencias:** Los datos no deben estar sujetos a ningún derecho de autor, patente, marca registrada o regulación de acuerdo de secreto. Se podrán permitir restricciones razonables de privacidad, seguridad o privilegio.

## 5.1 Atributos de calidad

Perfilado y reporte alrededor de atributos de calidad del dato.

El reporte de perfilado se utiliza para presentar avance en calidad contabilizando las instancias de atributo encontradas.

Para realizar el perfilamiento de datos se utiliza:

### 5.1.1 Informe de perfilado

Este informe incluye perfilado de datos, de información y de la Función Pública (en el sector público).

### 5.1.2 El perfilado de datos

El perfilado de datos es una medición de datos nulos, duplicados, posibles campos en desuso, máscaras de formato del dato, validaciones de formato de datos, etc. En este paso se detectan, entre otros, codificaciones incorrectas (por ejemplo, en departamentos del Estado colombiano), tipologías no previstas (por ejemplo, más de 2 tipos de sexo: hombre, mujer, masculino, femenino, etc.).

Este perfilado debe estar apoyado por una herramienta informática especializada para tal fin (software libre como Open Refine permite realizar esta tarea).

### 5.1.3 El perfilado de la información

Este perfilado de información va un paso más allá que el perfilado de datos. Por ejemplo, si en el perfilado de datos se valida que una fecha sea correcta, en el perfilado de información se iría más allá, interpretando la fecha (el contenido), tal vez como una fecha de nacimiento, y por lo tanto, obteniendo nuevos análisis.

### 5.1.4 El perfilado de Función Pública

El perfilado de Función Pública permite implementar la lógica compleja, por ejemplo, sobre resultados en campañas finalizadas en el sector público para stakeholders, o sobre el uso de reglas cruzadas para validación en cálculos de tarificación o facturación de servicios, etc. Este perfilado de Función Pública se apoya tanto, en los datos, como en sus metadatos funcionales: interpretaciones en el sector público del dato, lógicas de validación y umbrales de calidad.

### 5.1.5 Informe final

El informe final muestra la identificación de iniciativas preventivas para limpieza de datos en los sistemas fuente y para detección de posibles inconsistencias en sus repositorios, así como, la detección de fallas de calidad durante la ejecución del perfilado de datos, con el fin de aplicarlas en la fase de diseño de la migración de datos.

## 5.2 Procesos de calidad

Los procesos de calidad de los datos están ligados a la limpieza de estos y siempre están asociados a objetivos de prevención de fallas de calidad en el proceso de captura de datos.

Estos procesos están definidos por las características y reglas de negocio de la organización: ¿para qué se están capturando datos?, ¿qué tipo de datos se requieren? ¿Como se requiere la estructura de los datos?, otras. Según las respuestas se debe definir diversos procesos de limpieza o ninguno.

La limpieza de los datos incluye 5 actividades principales:

**Ilustración 8. Actividades limpieza de datos**


***Fuente.*** *Elaboración propia.*

## 5.3 Planificación de la calidad

La planificación de la calidad consiste en establecer una estrategia en la entidad en la cual se definan roles, responsabilidades, programas de calidad, secuencias de ejecución e interacción de procesos y uso de herramientas para la gestión de la calidad de los metadatos y datos.

**Fases:**

a. Planeación:
   - Diagnóstico de la calidad de los metadatos y datos.
   - Establecimiento del plan estratégico de calidad de metadatos y datos.

b. Desarrollo
   - Estimación de ajustes y nuevos desarrollos solicitados de Data Quality.
   - Realizar pruebas unitarias y funcionales de Data Quality.

c. Ejecución
   - Ejecución de programa de automatización (mensual o quincenal).
   - Comunicación con la entidad mediante correos (mensual o quincenal).
   - Monitoreo del correo de datosabiertos@mintic.gov.co

d. Seguimiento
   - Reunión periódica de seguimiento.
   - Informe DataQ sobre el avance en la mejora de la calidad (mensual o quincenal).

## 5.4 Validación de la calidad

La validación de la calidad de datos durante proceso de apertura por parte de las entidades consiste en realizar una identificación básica y manual de características de calidad de los datos de forma periódica.

**Semanal:**

- El conjunto de datos contiene menos de 50 registros.
- El enlace del conjunto de datos externo no permite la descarga directa de un conjunto de datos en formatos válidos: csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI).
- El enlace que se encuentra en el campo de la metadata de la URL de documentación y/o URL normativa, direcciona a una página no disponible.
- La metadata está incompleta y/o vacía.
- Conjunto de datos con menos de tres columnas.
- Carga de archivos no admitidos como datos abiertos.

**Quincenal:**

- Ajustes automáticos municipio, departamento, entidad, formato de fechas.
- El conjunto de datos no se encuentra actualizado a la fecha, de acuerdo con el campo de la metadata, “frecuencia de actualización” y la última fecha de actualización del conjunto de datos.
- El enlace del conjunto de datos externo direcciona a una página no disponible.
- Se requiere modificar el nombre del usuario, por el nombre de la entidad.
- Conjunto con columnas de solo texto.
- Conjunto de datos no tiene filas con información.

## 5.5 Control y aseguramiento de la calidad de los datos

Se puede resumir las actividades de control y aseguramiento de la calidad de los datos en 5 pasos que involucran algunas de las acciones descritas anteriormente, como son:

- **Descubrimiento de datos:** Proceso de búsqueda, recopilación, organización y notificación de metadatos.

- **Perfilado de datos**: Proceso de analizar los datos en detalle, comparándolos con sus metadatos, calculando estadísticas de datos, e informando de las medidas de calidad que se deben aplicar en cada momento.

- **Reglas de calidad de datos:** Se orientarán a optimizar el nivel de calidad de los activos informacionales de la organización y, para ello, se basarán en los requisitos de negocio aplicables, las reglas comerciales y técnicas a las que deben adherirse los datos.

- **Monitorización de la calidad de los datos**: La mejora continua requiere de un esfuerzo de seguimiento que permita comparar los logros con los umbrales de error definidos, la creación y almacenamiento de excepciones de calidad de datos y la generación de notificaciones asociadas para establecer mejoras continuas.

- **"Reporting" de calidad de datos:** Está relacionado con los procedimientos y herramientas empleadas para informar, detallar excepciones y actualizar las medidas de calidad de datos en curso.

- **Limpieza de datos:** Se ocupa de la corrección en curso de las excepciones y problemas de calidad de datos según son notificadas.

# 06. Calidad de los metadatos

## 6.1 ¿Que son los metadatos?

Los metadatos describen otros datos que explican y facilitan la información importante acerca de estos, lo que permite el acceso alos recursos de información, la interpretación y la reutilización de los datos.

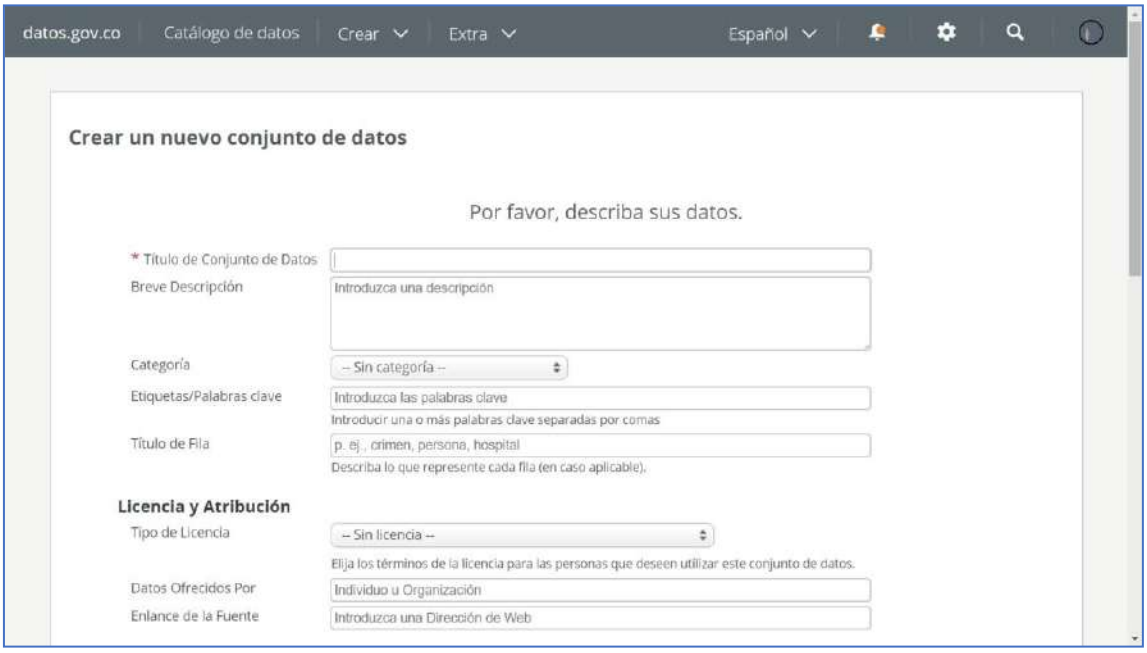
Una buena gestión de los metadatos debe conducir a una buena calidad de los datos, ya que tenerlos y confiar en ellos puede ayudar a identificar información deficiente, incorrecta y faltante.

El portal de Datos Abiertos de Colombia adoptó el estándar DCAT[17] para la definición de los metadatos de los conjuntos de datos, cuyo uso básico es hacer posible la interoperabilidad entre conjuntos de datos albergados en diferentes portales.

En la siguiente imagen se relacionan algunos de los metadatos requeridos durante el proceso de publicación de datos abiertos:

[17] *Ministerio de Tecnologías de la Información y las Comunicaciones*. 2016. Concepto Incorporación Estándar DCAT informhttp://lenguaje.mintic.gov.co/sites/default/files/archivos/concepto_incorporacion_dcat.pdf

**Ilustración 9. Metadatos del conjunto de datos**




***Fuente.*** *Tomado de Portal Datos abiertos. Herramientas.*

Se debe asegurar que los datos estén protegidos con alguna de las licencias para el uso de intercambio de información. Se recomienda proteger los conjuntos de datos a través de licencias abiertas estándar. Esta, permite proteger la estructura de los conjuntos de datos, sus contenidos, y garantizar que los usuarios de los datos atribuyan la titularidad de sus derechos a los creadores cuando se haga reúso de estos. [18]

La licencia Creative Commons, recomendada en los términos y condiciones de uso del portal de Datos Abiertos (www.datos.gov.co), es atribución – compartir igual (Attribution-ShareAlike 4.0 International) CC BY-SA 4.0.

Para acceder a más información acerca del diligenciamiento de los metadatos, se debe remitir al portal de Datos Abiertos de Colombia (www.datos.gov.co), en la sección de herramientas y observar los videos sobre cómo publicar conjuntos de datos[19].

[18] Portal de Datos Abiertos (s.f.). *Guía Apertura de Datos.* https://herramientas.datos.gov.co/sites/default/files/Guia%20de%20Datos%20Abiertos%20de%20Colombia.pdf
[19] Portal de Datos Abiertos (s.f.). *Publicar Conjuntos de Datos.* https://herramientas.datos.gov.co/herramientas/publicar-conjuntos-de-datos

# 07. Clasificación y priorización de errores de calidad y errores de publicación

## 7.1 Errores de Calidad

La siguiente tabla [20]contiene información importante de los tipos de errores que se evalúan en la actualidad, sobre los cuales, permite al negocio tomar decisiones del proceso actual de validación de conjuntos de datos, brinda la información del rechazo a partir de un conjunto de errores, y el usuario (entidad publicadora) identificará qué corregir para una futura carga.

**Tabla 7. Tipos de errores que se evalúan en la actualidad**

| Código del error | Categoría de error | Tipo de error | Descripción error |
|---|---|---|---|
| Clasificación de datos | Clusterizar datos | Clasificación de conjuntos de datos por periodos o categorías. | Actualmente los conjuntos de datos están siendo publicados por periodos y/o algún tipo de clasificación. |
| Completitud | Datos incompletos | Columnas o campos con información nula o vacía. | Grado en que los datos asociados con una entidad tienen valores para todos los atributos esperados e instancias de entidades relacionadas en un contexto de uso específico. |
| Comprensibilidad | Comprensibilidad de los datos | Conjunto de datos no permite ser interpretado. | Grado en el que los datos tienen atributos que permiten ser leídos e interpretados por los usuarios, y son expresados utilizando lenguajes, símbolos y unidades apropiados en un contexto de uso específico. Cierta información sobre la comprensibilidad puede ser expresada mediante metadatos. |
| Confidencialidad | Confidencialidad de los datos | Tratamiento de datos personales. | Grado en el que los datos tienen atributos que aseguran que estos son solo accedidos e interpretados por usuarios autorizados en un contexto de uso específico. La confidencialidad es un aspecto de la seguridad de la información (junto con la |

[20] Portal de Datos Abiertos. (6 de septiembre de 2021). *Errores Frecuentes de Calidad Conjuntos Abiertos de Datos del Portal*. Ministerio de Tecnologías de la Información y las Comunicaciones. https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/Errores-Frecuentes-de-Calidad-Conjuntos-de-datos-a/xbc7-65j4

| | | | |
|---|---|---|---|
| | | | disponibilidad y la integridad) definida como en: ISO/IEC 13335-1:2004. |
| Conformidad | Conformidad en los datos | Valores de los datos, de acuerdo con sus formatos no pueden ser utilizados. | Los datos que están en los campos de la tabla deben estar en un formato estándar y legible. |
| Duplicidad | Información duplicada | La información se está duplicando y ya está siendo publicada en el portal de Datos Abiertos por otra entidad. | La información se está duplicando y ya está siendo publicada en el portal de Datos Abiertos por otra entidad que puede pertenecer a la plataforma, departamento, municipio, ciudad, región, área o dependencia. |

***Fuente.*** *Elaboración propia.*

# 7.2 Errores de publicación

**Tabla 8. Errores de publicación.**

| Código del error | Categoría de error | Tipo de error | Descripción error |
|---|---|---|---|
| ERR001 | Metadata errada, incompleta y/o vacía. | Título y descripción mal nombrados. | El título y descripción del conjunto de datos está incompleto, presenta siglas, caracteres especiales o información que no es clara para los usuarios del conjunto de datos. |
| ERR002 | Metadata errada, incompleta y/o vacía. | La metadata del conjunto de datos está vacía o incompleta. | La metadata está incompleta o vacía, por lo cual se requiere completar los siguientes campos: {Incluir campos} |
| ERR003 | Metadata errada, incompleta y/o vacía. | El campo de la metadata correspondiente al nombre de usuario debe vincular la entidad publicadora. | El campo de la metadata correspondiente al nombre de usuario contiene un nombre particular de un funcionario, se recomienda vincular el nombre de la entidad, |
| ERR004 | Error sin filas | Conjunto de datos no tiene filas con información. | Actualmente el conjunto de datos no tiene registros, lo cual impide el máximo aprovechamiento de la información por parte de los ciudadanos, por lo anterior, se solicita complementar la estructura y los datos para que genere mayor valor al ciudadano. |
| ERR005 | Error pocas filas | Conjunto de datos, no es una base de datos o presenta poca información para reutilización. | El conjunto de datos tiene muy pocos registros (menos de 50) lo que no permitiría la reutilización por parte de los ciudadanos, para realizar un producto o servicio. |

| | | | |
|---|---|---|---|
| ERR005_01 | Error pocas filas y agregado | Conjunto de datos, no es una base de datos o presenta poca información para reutilización, y adicionalmente el conjunto de datos presenta agregaciones o totales. | El conjunto de datos tiene muy pocos registros (Menos de 50) lo que no permitiría la reutilización por parte de los ciudadanos para realizar un producto o servicio. Adicionalmente el conjunto de datos tiene registros agregados y totales, le recordamos que los datos abiertos deben publicarse en su máximo nivel de desagregación y completitud con el fin de maximizar el uso de los datos por parte de los ciudadanos |
| ERR005_02 | Error pocas filas turismo | Conjunto de datos, no es una base de datos o presenta poca información para reutilización. | El conjunto de datos tiene muy pocos registros (menos de 50), lo que no permitiría la reutilización por parte de los ciudadanos para realizar un producto o servicio. |
| ERR005_02 | Error pocas filas y clasificado por periodos | Conjunto de datos, no es una base de datos o presenta poca información para reutilización, y adicionalmente está publicado por periodos (año, mes). | El conjunto de datos tiene muy pocos registros (menos de 50) lo que no permitiría la reutilización por parte de los ciudadanos, para realizar un producto o servicio. Adicionalmente se publica por periodos, es decir por años, por ejemplo. |
| ERR007 | Error filas vacías | Conjunto de datos con campos vacíos y/o basura. | El conjunto de datos presenta campos vacíos; en las columnas XX el campo fecha XX no presenta formato tipo fecha en todos sus campos. |
| ERR008 | Error columnas | El conjunto de datos tiene una sola columna. | Actualmente la estructura del conjunto de datos cuenta con una única columna o un único campo de datos, lo cual impide el máximo aprovechamiento del portal de Datos Abiertos. |
| ERR008_1 | Error pocas columnas | Conjunto de datos presenta muy pocas columnas. | El conjunto de datos tiene muy pocas columnas (menos de 3), lo que no permitiría la reutilización por parte de los ciudadanos para realizar un producto o servicio. |
| ERR008_2 | Error columnas mal nombradas | Conjunto de datos presenta columnas mal nombradas Unnamed Column | |
| ERR009 | Error columnas | Error falta campo de geolocalización del conjunto de datos. | El conjunto de datos presenta campo de dirección, el cual no presenta estandarización. Siempre que exista un campo de dirección es necesario incluir campos de geolocalización (latitud y longitud), con el fin de que los usuarios puedan reutilizar el conjunto de datos para realizar mapas de ubicación. |

| | | | |
|---|---|---|---|
| ERR010 | Enlace inválido | El conjunto de datos enlaza a una dirección a un archivo en formato inválido (PDF). | El enlace del conjunto de datos externos no permite la descarga directa de un conjunto de datos en formatos válidos: csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI).<br>NO es abierto, está en cerrado como PDF, DOC, PPT. |
| ERR011 | Completitud del conjuntode datos | Conjunto de datos clasificado por periodos/<br>por tipologías. | Actualmente los conjuntos de datos están siendo publicados por periodos y/o algún tipo de clasificación, por lo cual se recomienda con el fin de que los datos abiertos cumplan con el principio de completitud. |
| ERR012_ 01 | Subconjunto de dato maestro - contratación | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro SECOP Integrado https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-Integrado/rpmr-utcd/data de la entidad nacional Colombia Compra Eficiente. |
| ERR012_ 02 | Subconjunto de dato maestro - ICFES | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacionalICFES. |
| ERR012_ 02 | Subconjunto de dato maestro - MinEducación | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional Ministerio de Educación. |
| ERR012_ 03 | Subconjunto de dato maestro - ZonasWIFI | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional XX. |
| ERR012_ 04 | Subconjunto de dato maestro - COVID-19 | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional Instituto Nacional de Salud (INS). |
| ERR012_ 05 | Subconjunto de dato maestro - trámites y servicios | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional Departamento Administrativo de la Función Pública (DAFP). |
| ERR012_ 06 | Subconjunto de dato maestro - DNP | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional Departamento Nacional de Planeación (DNP). |
| ERR013 | Publicación | El conjunto de datos está mal cargado. | El conjunto de datos presenta errores de calidad en la publicación. |

| | | | |
|---|---|---|---|
| ERR015 | Desactualizado | El conjunto de datos esta desactualizado. | El conjunto de datos fue creado xxx y la última fecha de actualización fue el xxx, de acuerdo con la información la periodicidad de actualización del conjunto es xxx. por lo cual el conjunto de datos no es reutilizable por los usuarios del portal. |
| ERR017 | Enlace roto | El conjunto de datos enlaza a una dirección inválida. | El enlace del conjunto de datos externos no permite la descarga directa de un conjunto de datos en formatos válidos: csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI). |
| ERR018 | Datos agregados o totalizados | El conjunto de datos presenta agregaciones o totales. | El conjunto de datos tiene registros agregados y totales. Se recuerda que los datos abiertos deben publicarse en su máximo nivel de desagregación y completitud, con el fin de maximizar el uso de los datos por parte de los ciudadanos. |
| ITA_1 | Ley de Transparencia y Derecho de Acceso a la Información Pública | Los datos no hacen parte de los conjuntos de datos como mínimos a publicar en el portal de Datos Abiertos, como instrumentos de gestión de información pública. | Información de interés. |
| ITA_2 | Errores activos de información | Error en conjunto de datos de registro de activos de información, | El conjunto de datos de activos de información no cumple con la estructura establecida y completa. |
| ITA_3 | Ley de Transparencia y Derecho de Acceso a la Información Pública | El conjunto de datos hace parte de la información a publicar en el portal propio de la entidad. | De acuerdo con la Ley 1712 de 2014 y a la Resolución 1519 de 2020, el conjunto de datos que se está publicando hace parte de la información mínima a publicar en el sitio web propio de la entidad en las sesiones:<br>1. Información de la entidad<br>2. Normativa<br>3. Contratación<br>4. Planeación, presupuesto e informes<br>5. Trámites<br>6. Participa<br>7. Información específica para grupos de interés. |
| Subconjunto | Subconjunto de datos maestros | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de una entidad nacional que consolida esta información. |
| Unicidad | Datos duplicados | Los registros del conjunto de datos se encuentran duplicados. | La medida en que todos los valores distintos de un elemento de datos aparecen solo una vez. |
| Uso y aprovechamiento | Poca reutilización | La información contenida en el conjunto de datos puede no estar | El conjunto de datos puede ser no reutilizado por parte de los ciudadanos para realizar un producto o servicio, en cuanto al contexto particular. |

| | |
|---|---|
| | considerada como dato abierto. |

***Fuente.*** *Elaboración propia.*

Como se puede apreciar con el siguiente ejemplo, el error ERR005_01 es una combinación de los errores ERR005 y ERR018.

- **ERR005**: Conjunto de datos, no es una base de datos, o presenta poca información para reutilización.
- **ERR005_01**: Conjunto de datos, no es una base de datos, o presenta poca información para reutilización, y adicionalmente el conjunto de datos presenta agregaciones o totales.
- **ERR018**: El conjunto de datos presenta agregaciones o totales.

Con la depuración y la especificidad de errores anteriores, se garantiza que los que han quedado, se utilicen para la parte de validación de calidad de datos basado en la norma ISO 25012.

## 7.3 Análisis de errores

**Tabla 9. Análisis de errores**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Clasificación de datos | Clusterizar datos | Clasificación de conjuntos de datos por periodos o categorías | Actualmente los conjuntos de datos están siendo publicados por periodos y/o algún tipo de clasificación. | Se recomienda utilizar columnas para la clasificación de periodos ( años, meses, días); utilizar columnas de agrupación por temáticas o tópicos, y unificar los datasets relacionados. |

***Fuente.*** *Elaboración propia.*

Como solución, se recomienda, utilizar las columnas de periodos y/o algún tipo de clasificación, y unificar los datasets relacionados.

## 7.3.1 Completitud

**Tabla 10. Análisis errores completitud**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Completitud | Datos incompletos | Columnas o campos con información nula o vacía | Grado en que los datos asociados con una entidad tienen valores para todos los atributos esperados e instancias de entidades relacionadas en un contexto de uso específico. | Se recomienda completar los campos vacíos o nulos de las columnas que presentan esta inconsistencia. De no ser un campo que no de valor o aporte al contexto del conjunto de datos, se recomienda eliminarla. |

***Fuente.*** *Elaboración propia.*

Si la sumatoria de las columnas nulas o vacías representan más del 5 % de columnas totales, se recomienda rechazar el conjunto.

## 7.3.2 Comprensibilidad

**Tabla 11. Análisis errores comprensibilidad**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Comprensibilidad | Comprensibilidad de los datos | Conjunto de datos, no permite ser interpretado. | Grado en el que los datos tienen atributos que permiten ser leídos e interpretados por los usuarios, y son expresados utilizando lenguajes, símbolos y unidades apropiados en un contexto de uso específico. Cierta información sobre la comprensibilidad puede ser expresada mediante metadatos. | Se recomienda mejorar los metadatos, incluyendo información o campos para poder clasificar los registros. Los datos deben encontrarse en formatos que permitan el procesamiento automático, con el más alto nivel de detalle posible, no en forma agregada, ni modificada. Así mismo, se requiere también un contexto para darles sentido, propósito y uso. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

## 7.3.3 Confidencialidad

**Tabla 12. Análisis errores confidencialidad**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Confidencialidad | Confidencialidad de los datos | Tratamiento de datos personales. | Grado en el que los datos tienen atributos que aseguran que estos son solo accedidos e interpretados por usuarios autorizados en un contexto de uso específico. La confidencialidad es un aspecto de la seguridad de la información (junto con la disponibilidad y la integridad) definida como en ISO/IEC 13335-1:2004. | Anonimizacion de la información. Se recomienda no publicar |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

## 7.3.4 Conformidad

**Tabla 13. Análisis errores conformidad**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Conformidad | Conformidad en los datos | Valores de los datos de acuerdo con sus formatos no pueden ser utilizados. | Los datos que están en los campos de la tabla deben estar en un formato estándar y legible. | Dar formato correcto para la utilización de las columnas del Dataset. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

## 7.3.5 Duplicidad

**Tabla 14. Análisis errores duplicidad**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Duplicidad | Información duplicada | La información se está duplicando y ya está siendo publicada en el portal de Datos Abiertos por otra entidad. | La información se está duplicando y ya está siendo publicada en el portal de Datos Abiertos por otra entidad que puede pertenecer a el mismo departamento, municipio, ciudad, región, área o dependencia. | Contacte la entidad que suministra la misma información, consolide y establezca el rol de la pertinencia de publicación. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

## 7.3.6 Metadata errada, incompleta y/o vacía

**Tabla 15. Metadata errada, incompleta y / o vacía**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR001 | Metadata errada, incompleta y/o vacía. | Título y descripción mal nombrados. | El título y descripción del conjunto de datos está incompleto, presenta siglas, caracteres especiales o información que no es clara para los usuarios del conjunto de datos. | Se recomienda actualizar el título y la descripción siguiendo las recomendaciones de calidad de las herramientas de apoyo. |
| ERR002 | Metadata errada, incompleta y/o vacía. | La metadata del conjunto de datos está vacía o incompleta. | La metadata está incompleta o vacía por lo cual se requiere completar los siguientes campos: {Incluir campos} | Se recomienda actualizar o incluir la información relacionada para los siguientes campos { Incluir campos} en la metadata de su conjunto de datos. |
| ERR003 | Metadata errada, incompleta y/o vacía. | El campo de la metadata correspondiente al nombre de usuario debe vincular la entidad publicadora. | El campo de la metadata correspondiente al nombre de usuario contiene un nombre particular de un funcionario. Se recomienda vincular el nombre de la entidad. | Se recomienda, con el fin de darle mayor reconocimiento a la entidad, actualizar la información del usuario, en el campo "Nombre de usuario" en donde se debe diligenciar en nombre de la entidad. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

## 7.3.7 Error sin filas

**Tabla 16. Erros sin filas**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR004 | Error sin Filas | Conjunto de datos, no tiene filas con información. | Actualmente el conjunto de datos no tiene registros, lo cual impide el máximo aprovechamiento de la información por parte de los ciudadanos, por lo anterior, se solicita complementar la estructura y los datos para que genere mayor valor al ciudadano. | Se recomienda "eliminar" el conjunto de datos y "crear" uno que contenga columnas y filas que complementen lo actualmente publicado o disponga la tabla en su sitio web.  <br><br>En caso de que la información solo sea de interés particular de la entidad se recomienda disponer esta información en la página web. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

## 7.3.8 Error pocas filas

**Tabla 17. Error pocas filas**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR005 | Error Poca Filas | Conjunto de datos, no es una base de datos, o presenta poca información para reutilización. | El conjunto de datos tiene muy pocos registros (menos de 50), lo que no permitiría la reutilización por parte de los ciudadanos para realizar un producto o servicio. | Dado a que el conjunto de datos tiene menos de 50 registros, se recomienda aumentar el número de estos, con el fin de que sea de utilidad para los usuarios. Así mismo se debe agregar nuevas columnas y consolidar con otras bases de datos complementarias o históricas. |
| ERR005_01 | Error Poca Filas y Agregado | Conjunto de datos no es una base de datos o presenta poca información para reutilización. Adicionalmente el conjunto de datos presenta agregaciones o totales. | El conjunto de datos tiene muy pocos registros (menos de 50), lo que no permitiría la reutilización por parte de los ciudadanos para realizar un producto o servicio. Adicionalmente el conjunto de datos tiene registros agregados y totales. SE recuerda que los datos abiertos deben publicarse en su máximo nivel de desagregación y completitud, con el fin de maximizar el uso de los datos por parte de los ciudadanos. | Se recomienda mejorar la completitud del conjunto de datos, con el fin de que sea de utilidad para los usuarios, disponer la fuente de datos detallada con la cual se generó este reporte y hacer uso de la funcionalidad de vistas filtradas del portal para generar reportes<br><br>En caso de que la información solo sea de interés particular de la entidad, se recomienda disponer esta información en la página web. |

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| **ERR005_02** | Error pocas filas turismo | Conjunto de datos, no es una base de datos o presenta poca información para reutilización. | El conjunto de datos tiene muy pocos registros (menos de 50), lo que no permitiría la reutilización por parte de los ciudadanos para realizar un producto o servicio. | Se recomienda para que se disponga una base de datos consolidada, integrar los conjuntos de datos relacionados con turismo en un solo, denominado "Oferta Turística del Municipio de xxx", de esta manera se |

| | | | | |
|---|---|---|---|---|
| | | | | consolidarían hoteles, comercio, restaurantes, puntos de interés, parques, entre otros, con el fin de mejorar la completitud del conjunto de datos, para que sea de utilidad para los usuarios, añadiéndole nuevas columnas que brinden más información al ciudadano como: datos de geolocalización, tipo de establecimiento, ciudad, departamento, sitio web, correo, horario de atención, etc. |
| **ERR005_02** | Error pocas filas y clasificado por periodos | Conjunto de datos, no es una base de datos o presenta poca información para reutilización y adicionalmente esta publicado por periodos (año, mes). | El conjunto de datos tiene muy pocos registros (menos de 50), lo que no permitiría la reutilización por parte de los ciudadanos para realizar un producto o servicio. Adicionalmente se publica por periodos, es decir, por años, por ejemplo. | Se recomienda mejorar la completitud del conjunto de datos, con el fin de que sea de utilidad para los usuarios, consolidando toda la información en un solo conjunto de datos, en vez de publicar varias bases por periodos (años), con el fin de disponer una información con el mayor nivel de completitud posible y en su máximo nivel de desagregación e “incluir” una columna de periodo( ejemplo año, fecha, semestre, etc.) al conjunto de datos, se consoliden todos los conjuntos de datos en uno solo y se vaya incluyendo los registros a medida que se van generando nuevos datos en los diferentes años. En caso de que lainformación solo sea de interés particular de la entidad, se recomienda disponer está en la página web. |
| **ERR007** | Error filas vacías | Conjunto de datos con campos vacíos y/o basura. | El conjunto de datos presenta campos vacíos  en las columnas XX, el campo fecha XX, no presenta formato tipo fecha en todos sus campos. | Se recomienda hacer uso de herramientas de calidad de datos para estandarizar los campos e identificar los errores presentados. |

***Fuente.*** *Elaboración propia.*

Error pocas filas es de los más populares, no obstante, hay excepciones donde hay conjunto de datos que se pueden aprobar con menos de 50 registros, identificando si el universo completo de los datos lo componen menos de 50 datos, por ejemplo, departamentos de Colombia, municipios de Cauca, etc.

## 7.3.9 Error pocas columnas

**Tabla 18. Error pocas columnas**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR008 | Error Columnas | El conjunto de datos tiene una sola columna | Actualmente la estructura del conjunto de datos cuenta con una única columna o un único campo de datos, lo cual, impide el máximo aprovechamiento del portal de datos abiertos. | Se recomienda eliminar el conjunto de datos y crear uno que contenga columnas y filas que complementen lo actualmente publicado, o disponga la tabla en el sitio web.  <br><br>Si la información a publicar es solo de interés particular de la entidad, se debe disponer de la tabla en el sitio web. |
| ERR008_1 | Error Poca Columnas | Conjunto de datos presenta muy pocas columnas | El conjunto de datos tiene muy pocas columnas (menos de 3), lo que no permitiría la reutilización por parte de los ciudadanos para realizar un producto o servicio. | Se recomienda mejorar la completitud del conjunto de datos agregando nuevas columnas, con el fin de que sea de utilidad para los usuarios del conjunto de datos. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

## 7.3.10 Error pocas columnas mal nombradas

**Tabla 19. Error pocas columnas mal nombradas**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR008_2 | Error columnas mal nombradas | Conjunto de datos presenta columnas mal nombradas Unnamed Column | | Se recomienda hacer uso de herramientas de calidad de datos para estandarizar los campos e identificar los errores presentados. |

***Fuente.*** *Elaboración propia.*

Se recomienda cambiar los textos para dejarlo más genérico y que comunique generalidades del error, y no especificaciones, por lo tanto, se recomienda:

- Tipo de error: Conjunto de datos presenta columnas mal nombradas.
- Descripción Error: Existen columnas que en su nombre tiene títulos genéricos que no permiten identificar los atributos que contiene el conjunto de datos.
- Solución: Se recomienda hacer uso de herramientas de calidad de datos para estandarizar los campos e identificar los errores presentados nombrando columnas de manera genérica, por ejemplo: Unnamed Column, Unnamed, colum1, colum2, dato1, dato, etc.

## 7.3.11 Error falta campo de geolocalización del conjunto de datos

**Tabla 20. Error falta de geolocalización del conjunto de datos**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR009 | Error columnas | Error falta campo de geolocalización del conjunto de datos | El conjunto de datos cuenta con el campo de dirección, el cual, no presenta estandarización. Siempre que exista un campo de dirección, es necesario incluir campos de geolocalización (latitud y longitud), con el fin de que los usuarios puedan reutilizar el conjunto de datos para realizar mapas de ubicación. | Se recomienda incluir campos de geolocalización (latitud y longitud), con el fin de que los usuarios puedan reutilizar el conjunto de datos para realizar mapas de ubicación. |

**Fuente.** Elaboración propia.

Es una regla muy exigente y se recomienda cambiarla para solo los conjuntos de datos que se utilizan en mapa o en forma geográfica. Para decidir lo anterior, es necesario que se analice si el conjunto de datos se utiliza en trabajos de análisis geográfico, o si, por lo contrario, se ejecutan análisis estadísticos. Por lo anterior, se recomienda cambiar así:

- Descripción: El conjunto de datos presenta datos de ubicación (dirección), y según la revisión, se concluye que el conjunto de datos se utiliza como mapa o capa geográfica, por lo tanto, es necesario incluir columnas de geolocalización (latitud y longitud), y registrar las coordenadas en el sistema de coordenadas Magna Sirgas o WGS84.
- Solución: Se recomienda incluir columnas de geolocalización (latitud y longitud) y registrar las coordenadas en el sistema Magna Sirgas o WGS84, con el fin de que los usuarios puedan reutilizar el conjunto de datos para realizar mapas de ubicación.

## 7.3.12. Enlace inválido

**Tabla 21. Enlace inválido**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR010 | Enlace inválido | El conjunto de datos enlaza a una dirección, o a un archivo en formato inválido (PDF) | El enlace del conjunto de datos externos no permite la descarga directa de un conjunto de datos en formatos válidos: csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI) <br><br>NO es un dato abierto esta en un formato cerrado como PDF, .DOC, .PPT | Se recomienda corregir el enlace del conjunto de datos externo, que permita la descarga directa de alguno de los formatos estándares abiertos soportados.<br><br> Se recuerda que los datos abiertos son estructurados en filas y columnas, dispuestos en formatos abiertos como:  csv, xls, xlsx, json, kml, kmz  y zip (shapefile de ESRI) y deben cumplir con los principios de completitud,  fuente primaria y debe presentar un alto nivel de desagregación. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

# 7.3.13 Conjunto o subconjunto de errores

### Tabla 22. Conjunto o subconjunto de errores

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| **ERR011** | Completitud del Conjunto de datos | Conjunto de datos clasificado por periodos/ por tipologías | Actualmente los conjuntos de datos están siendo publicados por periodos y/o algún tipo de clasificación, por lo cual se recomienda con el fin de que los datos abiertos cumplan con el principio de completitud. | Se recomienda crear un solo conjunto de datos denominado “Atenciones” e incluir una columna de periodo (ejemplo año, fecha, semestre, etc.) al conjunto de datos; se consoliden todos los conjuntos de datos en uno solo, y se vaya incluyendo los registros a medida que se van generando nuevos datos en los diferentes años. Este conjunto de datos es de gran importancia, por lo cual se recomienda brindarles a los ciudadanos la información de manera consolidada, que facilite el uso y análisisde esta información histórica. Adicionalmente que garantice la actualización permanente del conjunto de datos consolidados. |
| **ERR012 _01** | Subconjunto de dato maestro - Contratación | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional | Este conjunto de datos es un subconjunto de un dato abierto maestro SECOP Integrado https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-Integrado/rpmr-utcd/data de la entidad nacional Colombia Compra Eficiente | Se recomienda, para evitar datos duplicados y desactualizados, en el Portal Nacional de Datos Abiertos eliminar este conjunto de datos publicado y crear una vista filtrada del conjunto de datos maestro dela entidad nacional, y filtrar por la ciudado departamento. |
| **ERR012 _02** | Subconjunto de dato maestro - ICFES | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional ICFES. | Se recomienda para evitar datos duplicados y desactualizados en el Portal Nacional de Datos Abiertos eliminar este conjunto de datos publicado y crear una vista filtrada del conjunto de datos maestro dela entidad nacional, y filtrar por la ciudad o departamento. |
| **ERR012 _02** | Subconjunto de dato maestro - MinEducación | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad Nacional Ministerio de Educación. | Se recomienda, para evitar datos duplicados y desactualizados en el Portal Nacional de Datos Abiertos eliminar este conjunto de datos publicado y crear una vista filtrada del conjunto de datos maestro dela entidad nacional, y filtrar por la ciudad o departamento. |
| **ERR012 _03** | Subconjunto de dato maestro - ZonasWIFI | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional XX. | Se recomienda para evitar datos duplicados y desactualizados en el Portal Nacional de Datos Abiertos, eliminar este conjunto de datos publicado y crear una vista filtrada del conjunto de datos maestro dela entidad nacional, y filtrar por la ciudad o departamento. |
| **ERR012 _04** | Subconjunto de dato maestro - COVID-19 | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional Instituto Nacional de Salud (INS). | Se recomienda, para evitar datos duplicados y desactualizados en el Portal Nacional de Datos Abiertos, eliminar su conjunto de datos, y crear una vista filtrada del conjunto de datos maestro del INS https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data, y |

| | | | | |
|---|---|---|---|---|
| | | | | filtrar por su municipio o departamento. Se adjunta el tutorial donde se explica el paso a paso de cómo construir esta vista de casos positivos del COVID. |
| **ERR012_05** | Subconjunto de dato maestro - Trámites y Servicios | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional Departamento Administrativo de la Función Pública-DAFP. | Se recomienda, para evitar datos duplicados y desactualizados en el Portal Nacional de Datos Abiertos, eliminar su conjunto de datos y crear una vista filtrada del conjunto de datos maestro del DAFP https://www.datos.gov.co/Funci-n-p-blica/Detalle-de-la-informaci-n-registrada-en-el-SUIT-po/48fq-mxnm, y filtrar por su municipio o departamento. Se adjunta el tutorial donde se explica el paso a paso de cómo construir esta vista. |
| **ERR012_06** | Subconjunto de dato maestro - DNP | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de la entidad nacional Departamento Nacional de Planeación (DNP). | Se recomienda, para evitar datos duplicados y desactualizados en el Portal Nacional de Datos Abiertos, eliminar su conjunto de datos, y crear una vista filtrada del conjunto de datos maestro del DNP. Se adjunta el tutorial donde se explica el paso a paso de cómo construir esta vista. |

***Fuente.*** *Elaboración propia.*

No se tienen observaciones a los errores.

## 7.3.14 El conjunto de datos está mal cargado

**Tabla 23. Conjunto de datos mal cargado**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| **ERR013** | Publicación | El conjunto de datos está mal cargado. | El conjunto de datos presenta errores de calidad en la publicación. | Se recomienda eliminar el conjunto de datos y crear correctamente el conjunto. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error, solo que este aplica únicamente en files o ref.

## 7.3.15 Desactualizado

**Tabla 24. Desactualizado**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR015 | Desactualizado | El conjunto de datos esta desactualizado. | El conjunto de datos fue creado xxx, y la última fecha de actualización fue el xxx, de acuerdo con la información, la periodicidad de actualización del conjunto es xxx, por lo cual, el conjunto de datos no es reutilizable por los usuarios del portal. | Se recomienda actualizar el conjunto de dato y mejorar la completitud, con el fin de que sea de utilidad para los usuarios del conjunto de datos<br><br> En caso de que la información solo sea de interés particular de la entidad, se recomienda eliminar el conjunto de datos, y disponer esta información en la página web de la misma. |

***Fuente.*** *Elaboración propia.*

No hay observaciones del error.

## 7.3.16 ERR017 Enlace roto

**Tabla 25. Enlace roto**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR017 | Enlace roto | El conjunto de datos enlaza a una dirección inválida. | El enlace del conjunto de datos externos no permite la descarga directa de un conjunto de datos en formatos válidos: csv, xls, xlsx, json, kml, kmz  y zip (shapefile de ESRI) | Se recomienda corregir el enlace del conjunto de datos externo, que permita la descarga directa de alguno de los formatos estándares abiertos soportados.  Se recuerda que los datos abiertos son datos estructurados en filas y columnas, dispuestos en formatos abiertos como  csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI) y deben cumplir con los principios de completitud,  fuente  primaria y debe presentar un alto nivel de desagregación. |

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| **ERR017** | Enlace roto | El conjuntode datos enlazaauna dirección inválida. | El enlace del conjunto de datos externos no permite la descarga directa de un conjunto de datos en formatos válidos: csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI) | Se recomienda corregir el enlace del conjunto de datos externo, de manera que permita la descarga directa en alguno de los formatos estándares abiertos soportados. Se recuerda que los datos abiertos son datos estructurados en filas y columnas, dispuestos en formatos abiertos como csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI) y deben cumplir con los principios de completitud, fuente primaria y presentar un alto nivel de desagregación. |

***Fuente.*** *Elaboración propia.*

Se sugiere eliminar el “ERR017 Enlace roto”, ya que está duplicado con: ERR010 Enlace inválido.

## 7.3.17 El conjunto de datos presenta agregaciones o totales

**Tabla 26. El conjunto de datos presenta agregaciones o totales**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| ERR018 | Datos agregados o totalizados | El conjunto de datos presenta agregaciones o totales. | El conjunto de datos tiene registros agregados y totales, se recuerda que los datos abiertos deben publicarse en su máximo nivel de desagregación y completitud, con el fin de maximizar el uso de los datos por parte de los ciudadanos. | Se recomienda disponer de la fuente de datos detallada, con la cual se generó este reporte, y hacer uso de la funcionalidad de vistas filtradas del portal para generar reportes. |

**Fuente.** Elaboración propia.

No hay observaciones del error.

# 7.3.18 ITA - Ley de Transparencia y Derecho de Acceso a la Información Pública

**Tabla 27. El conjunto de datos presenta agregaciones o totales**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| **ITA_1** | Ley de Transparencia y Derecho de Acceso a la Información Pública | Los datos no hacen parte de los conjuntos de datos como mínimos a publicar en el portal de Datos Abiertos *como* instrumentos de gestión de información pública. | Información de interés | Cómo mínimo, el índice de información pública reservada y clasificada, y los registros de activos de información, deben estar publicados en datos abiertos. La publicación de estos datos, independientemente del formato del archivo en el que se encuentren (Word, Excel, CSV), debe estar disponible de forma accesible y reutilizable. |
| **ITA_2** | Errores activos de información | Error en conjunto de datos de registro de activos de información. | El conjunto de datos de activos de información no cumple con la estructura establecida y completa. | El portal cuenta con una herramienta de activos de información que facilitará el diligenciamiento correcto y completo de la información, tal como lo establece la Ley 1712 de 2014. |
| **ITA_3** | Ley de Transparencia y Derecho de Acceso a la Información Pública | El conjunto de datos hace parte de la información a publicar en el portal propio de la entidad. | De acuerdo con la Ley 1712 de 2014 y a la Resolución 1519 de 2020, el conjunto de datos que se está publicando hace parte de la información mínima a publicar en el sitio web propio de la entidad en las sesiones;<br>1. Información de la entidad.<br>2. Normativa.<br>3. Contratación.<br>4. Planeación, presupuesto e informes.<br>5. Trámites.<br>6. Participa.<br>8. Información específica para grupos de interés. | Se recomienda publicar en el sitio propio de acuerdo con la normativa en cuanto a los numerales;<br>1. Información de la entidad. 2. Normativa. 3. Contratación. 4. Planeación, presupuesto e informes. 5. Trámites. 6. Participa. 8. Información específica para grupos de interés<br>y publicar en datos abiertos lo que dicta el numeral 7. Datos abiertos. |

***Fuente.*** *Elaboración propia.*

No hay observaciones al error, no obstante, se encuentra reiterativo que el ITA_3 ocurre con mucha frecuencia en conjuntos de datos de tipo: recurso externos y files, por lo tanto, se sugiere que en las capacitaciones sea explicativo en todos estos errores.

## 7.3.19 Subconjunto de datos maestros

**Tabla 28. Subconjunto de datos maestros**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Subconjunto | Subconjunto de datos maestros | Dato es un subconjunto, de un conjunto de dato maestro, de una entidad nacional. | Este conjunto de datos es un subconjunto de un dato abierto maestro de una entidad nacional que consolida esta información. | Se recomienda, para evitar datos duplicados y desactualizados en el Portal Nacional de Datos Abiertos, eliminar su conjunto de datos, y crear una vista filtrada del conjunto de datos maestro del de la entidad productora. |

***Fuente.*** *Elaboración propia.*

Se encuentra que el “ERR012_03”, es muy similar a este, por lo tanto, se sugiere unirlo en uno solo con “Subconjunto”.

## 7.3.20 Unicidad

**Tabla 29. Unicidad**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Unicidad | Datos duplicados | Los registros del conjunto de datos, se encuentran duplicados. | La medida en que todos los valores distintos de un elemento de datos aparecen solo una vez. | Ajustar el conjunto de datos de manera que la información no se perciba con duplicidades, o registros duplicados. |

**Fuente.** Elaboración propia.

Se sugiere que en el código de error no esté la palabra *unicidad* porque puede haber confusiones, y que se cambie la descripción del error por: “Se encuentra

columnas que existen un mismo valor en todos los registros, o se encontraron filas idénticas".

## 7.3.21 Poca reutilización

**Tabla 30. Poca reutilización**

| Código del error | Categoría de error | Tipo de error | Descripción error | Solución |
|---|---|---|---|---|
| Uso y Aprovechamiento | Poca reutilización | La información contenida en el conjunto de datos puede no estar considerada como dato abierto. | El conjunto de datos puede ser no reutilizado por parte de los ciudadanos para realizar un producto o servicio, en cuanto al contexto particular. | Se recomienda actualizar el conjunto de dato y mejorar la pertinencia, con el fin de que sea de utilidad para los usuarios del conjunto de datos. En caso de que la información solo sea de interés particular de la entidad, se recomienda eliminar el conjunto de datos, y disponer esta información en la página web de esta. |

***Fuente.*** *Elaboración propia.*

No hay observaciones al error.

## 7.3 Revisión de errores

**Tabla 31. Revisión de errores**

| Calidad de datos | Errores frecuentes |
|---|---|
| Exactitud | Hay errores frecuentes relacionados con la exactitud de los datos, no obstante, faltan fórmulas para determinar la Exactitud Sintáctica y Semántica. |
| Completitud | Existen varias reglas de completitud que para el concepto de este análisis son suficientes, ya que se analiza la completitud, tanto en columnas, filas y volumen de información. Se recomienda establecer fórmulas y niveles de aprobación. |
| Consistencia | Hay reglas de subconjuntos y conjuntos globales que suplen este ítem de evaluación, no obstante, se podría utilizar IA y Machine Learning para analizar la consistencia del dato, utilizando modelos propios elaborados, con información de datos abiertos. |
| Credibilidad | Se debe trabajar en implementar reglas y técnicas de IA o Machine Learning con las que se revise lo creíble o cierto de los datos. |
| Actualidad | Hay reglas como "Desactualizado", pero no obstante hay que determinar hasta qué año anterior se va a recibir datos, por ejemplo, determinar ¿si llegan datos del 2016 se suben, o no?, o de qué clase subir y de cuáles no. |

***Fuente.*** *Elaboración propia.*

# 08. Glosario

**Anonimización:** Es un proceso aplicado a un conjunto de datos con el fin de impedir la identificación, individualización y caracterización de los individuos sujetos en la producción de los datos. La anonimización se realiza por razones de seguridad para evitar posibles fines maliciosos que se le pueda dar a los datos publicados.

**API:** una Interfaz de Programación de Aplicaciones (Application Programming Interface) es un conjunto de protocolos, servicios y métodos de comunicación entre varios componentes de software. En las plataformas de datos abiertos, las API's proveen un servicio de consumo de datos, que puede ser accedido por cualquier desarrollador que requiera dichos datos para alimentar una aplicación o sistema informático. Las colecciones pueden tener un enfoque temático u orgánico, dependiendo de si están organizadas por un tema específico (clima, educación, transporte, etc.) o si están organizadas por la entidad que los publica, como por ejemplo "División de Transporte" de un municipio.[21]

**Catálogo de datos**: Es la parte central de un portal de datos abiertos y contiene un listado de todas las tablas publicadas con una descripción del contenido de la base de datos, el nombre de la agencia responsable, la frecuencia de actualización, el número de veces que se ha visitado, la información técnica para conectarla con aplicaciones informáticas y un espacio para los comentarios de los usuarios. Algunos portales incorporan también un área de valoración de la calidad de la base de datos.[22]

**Conjunto de datos -Dataset-:** Es una colección de registros discretos representados en estructuras de datos generalmente tabulares (filas y columnas), que pueden ser accedidos y utilizados individualmente o en combinación.

**Datos abiertos -Open Data-:** Son datos digitales que son puestos a disposición con las características técnicas y jurídicas necesarias para que puedan ser usados, reutilizados y redistribuidos libremente por cualquier persona, en cualquier momento y en cualquier lugar.

---

[21] Moneo. A. BID. (20 de noviembre de 2014). *¿Qué es un Portal de Datos Abiertos y para qué Sirve?* BID. https://blogs.iadb.org/conocimiento-abierto/es/que-es-un-portal-de-datos-abiertos-y-para-que-sirve/

[22] Moneo. A. BID. (20 de noviembre de 2014). *Ibid.* BID. https://blogs.iadb.org/conocimiento-abierto/es/que-es-un-portal-de-datos-abiertos-y-para-que-sirve/

**Distribuciones de los datos:** Representa una forma accesible de datos en un catálogo de datos: un archivo descargable, una fuente RSS o un servicio web que proporciona los datos.

**Interoperabilidad:** Es la "capacidad de 2 o más sistemas, o componentes para intercambiar información y utilizar la información intercambiada": ISO 25010.

**Lenguaje Común de Intercambio de Información:** Es el estándar definido por el Estado colombiano para el intercambio de información entre sus dependencias y entidades.

**Metadatos-Metadata:** es información adicional que describe, explica y facilita datos importantes acerca del conjunto de datos. Estos permiten el acceso a los recursos de información y la reutilización de esta. Los metadatos son fundamentales para la interoperabilidad de la información y para el cruce de datos en las instituciones públicas y privadas. La estandarización de los campos de metadatos que se deben de diligenciar en un dataset posibilita a los usuarios comprender mejor el conjunto de datos y hacer mucho más sencillo ese cruce de información entre las entidades.

**Visualizaciones de datos:** La sección de visualizaciones es un área donde se pueden publicar gráficos o vistas de datos que hayan creado los administradores o los usuarios, si la aplicación lo permite. Algunas plataformas vienen con módulos integrados para generar visualizaciones básicas y, en la mayoría de los casos, estas plataformas se conectan con otras herramientas externas que permiten hacer visualizaciones más vistosas y complejas.[23].

[23] Moneo. A. BID. (20 de noviembre de 2014). *¿Qué es un Portal de Datos Abiertos y para qué Sirve?* BID. https://blogs.iadb.org/conocimiento-abierto/es/que-es-un-portal-de-datos-abiertos-y-para-que-sirve/

# Referencias


Cai, L. a. (2015). *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.* Retrieved from Data Science Journal, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002

Crosby, P. B. (1988). *a organización permanentemente exitosa /por Philip B. Crosby y traducción de Roberto Haas.* Mexico: McGraw-Hill.

Estado Peruano. (s.f.). *Infrastructura de datos especiales*. Retrieved from Plataforma digital única del Estado Peruano: https://www.geoidep.gob.pe/conoce-las-ides/metadatos/que-son-los-metadatos

Ministerio de Tecnologías de la Información y las Comunicaciones. (2019, agosto). *Marco de interoperabilidad.* Retrieved from http://lenguaje.mintic.gov.co/sites/default/files/archivos/marco_de_interoperabilidad_para_gobierno_digital.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones. (2010). *Marco para la Interoperabilidad del Gobierno en Línea.* Retrieved from https://www.mintic.gov.co/arquitecturati/630/articles-9375_marco_interoperabilidad_pdf.pdf

Ministerio de Tecnologías de la Información y las comunicaciones. (2019). *Guía para el uso y aprovechamiento de Datos Abiertos en Colombia.* Retrieved from http://es.presidencia.gov.co/dapre/DocumentosSIGEPRE/G-GD-02-calificacion-informacion.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones. (2019, agosto). *Lenguaje Común de Intercambio de Información.* Retrieved from Marco de interoperabilidad para Gobierno Digital: http://lenguaje.mintic.gov.co/sites/default/files/archivos/marco_de_interoperabilidad_para_gobierno_digital.pdf

Norma ISO 9000. (2000). *ISO 2000.* Retrieved from GESTIONDECALIDADMPN: https://gestiondecalidadmpn.files.wordpress.com/2012/02/iso-9000-2000-sistemas-de-gestic3b3n-de-la-calidad-conceptos-y-vocabulario.pdf

Presidencia de la República de Colombia. (2019, septiembre). *Presidencia de la República de Colombia.* Retrieved from https://dapre.presidencia.gov.co/dapre/DocumentosSIGEPRE/G-GD-02-calificacion-informacion.pdf

Strong, R. Y. (1996). *Beyond Accuracy: What Data Quality Means to Data Consumers.* M.E. Sharpe, Inc.