



TIC



DATOS ABIERTOS

Guía de estándares de CALIDAD E
INTEROPERABILIDAD de los
Datos Abiertos

Ministerio de Tecnologías de la Información y las Comunicaciones

Viceministerio de Transformación Digital

Dirección de Gobierno Digital

Julián Molina Gómez

Ministro de Tecnologías de la Información y las Comunicaciones

Yeimi Carina Murcia Yela

Viceministro de Transformación Digital

Lucy Urón Rincón

Director de Gobierno Digital

Natalia Albañil Riaño

Subdirectora de Fortalecimiento de Capacidades Digitales

Equipo Técnico Iniciativa Datos Abiertos

- Luisa Fernanda Medina, Asesora Viceministerio de Transformación Digital Líder iniciativa nacional de datos abiertos
- Andres Felipe Guerra, Consultor Técnico iniciativa Datos Abiertos
- Andres Felipe Yanes, Consultor Lineamientos Datos Abiertos
- Luis Amílcar Bonivento, Consultor de Apropiación Iniciativa Datos Abiertos

CONTROL DE VERSIONES

Versión	Observaciones
Versión 1 Diciembre 2016	Guía para la apertura de datos en Colombia Dirigida a entidades del Estado para el desarrollo de procesos de apertura de datos públicos
Versión 2 Junio 2016	Guía de datos abiertos en Colombia Dirigida a las entidades sujeto de aplicación de la Ley 1712 de 2014 de Transparencia y Acceso a la Información Pública, para la aplicación de orientaciones y buenas prácticas en el desarrollo de estrategias de apertura y reuso de datos abiertos.
Versión 3 Junio 2016	Guía de datos abiertos en Colombia - modificaciones de diagramación
Versión 4 Enero 2019	Actualización temática de la Guía para el uso y aprovechamiento de Datos Abiertos en Colombia
Versión 5 Mayo 2019	Guía para el uso y aprovechamiento de Datos Abiertos en Colombia - modificaciones de diagramación
Versión 6 Septiembre 2019	Actualización proceso Ciclo de Datos alineado al reto de máxima velocidad.
Versión 7 Diciembre 2024	Actualización proceso con cambios en los parámetros de evaluación de Calidad
Versión 8 Mayo 2025	Actualización de normatividad Conpes de Inteligencia Artificial 4144 de 2025

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

Tabla de contenido

1. INTRODUCCIÓN	1
2. LA CALIDAD DE LOS DATOS	3
2.1. IMPORTANCIA DE PUBLICAR DATOS ABIERTOS DE CALIDAD	4
2.2. MEJORAMIENTO DE LA CALIDAD DE LOS DATOS	8
2.3. MARCO DE INTEROPERABILIDAD	9
3. CRITERIOS DE CALIDAD E INTEROPERABILIDAD	14
3.1. HERRAMIENTA DE EVALUACIÓN DE CALIDAD.....	16
3.2. CRITERIO DE CONFIDENCIALIDAD.....	17
3.2.1. DEFINICIÓN Y REGULACIÓN	17
3.2.2. CÁLCULO DEL CRITERIO	19
3.2.3. ANONIMIZAR LOS DATOS.....	21
3.3. CRITERIO DE RELEVANCIA.....	22
3.3.1. DEFINICIÓN Y REFERENTES.....	22
3.3.2. CÁLCULO DEL CRITERIO.....	24
3.4. CRITERIO DE ACTUALIDAD.....	25
3.4.1. CÁLCULO DEL CRITERIO	27
3.5. CRITERIO DE TRAZABILIDAD	28
3.5.1. CÁLCULO DEL CRITERIO	28
3.6. CRITERIO DE CONFORMIDAD	30
3.6.1. CÁLCULO DEL CRITERIO	30
3.7. CRITERIO DE EXACTITUD	32
3.7.1. EXACTITUD SINTÁCTICA.....	32
3.7.2. EXACTITUD SEMÁNTICA.....	34
3.8. CRITERIO DE COMPLETITUD	35
3.8.1. CÁLCULO DEL CRITERIO.....	36
3.9. CRITERIO DE CONSISTENCIA	38
3.9.1. CÁLCULO DEL CRITERIO	38
3.10. CRITERIO DE PRECISIÓN	40
3.10.1. CÁLCULO DEL CRITERIO	40
3.11. CRITERIO DE PORTABILIDAD	41
3.11.1. CÁLCULO DEL CRITERIO	42
3.12. CRITERIO DE CREDIBILIDAD	44
3.12.1. CÁLCULO DEL CRITERIO	44
3.13. CRITERIO DE COMPRENSIBILIDAD	46
3.13.1. CÁLCULO DEL CRITERIO	46
3.14. CRITERIO DE ACCESIBILIDAD	48
3.14.1. CÁLCULO DEL CRITERIO	49
3.15. CRITERIO DE UNICIDAD	50
3.15.1. CÁLCULO DEL CRITERIO	50



TIC

3.16.	CRITERIO DE EFICIENCIA	51
3.16.1.	CÁLCULO DEL CRITERIO	51
3.17.	CRITERIO DE RECUPERABILIDAD	52
3.18.	CRITERIO DE DISPONIBILIDAD	53
4.	SELLOS DE CALIDAD	54
4.1.	SELLO DE CALIDAD 0	55
4.2.	SELLO DE CALIDAD 1	55
4.3.	SELLO DE CALIDAD 2	56
4.4.	SELLO DE CALIDAD 3	57
5.	POTENCIAL DE USO	58
5.1.	MODELO DE PREDICCIÓN DEL POTENCIAL DE USO	59
5.1.1.	EJEMPLOS DE CÁLCULO DEL CTR	60
5.1.2.	INTERPRETACIÓN DEL CTR	60
5.1.3.	APLICACIÓN DEL CTR EN LA EVALUACIÓN	61
5.2.	IMPLEMENTACIÓN DEL MODELO DE PREDICCIÓN	61
5.2.1.	OPTIMIZACIÓN Y EVALUACIÓN DEL MODELO	61
6.	PRINCIPIOS DE LA CALIDAD PARA LA PUBLICACIÓN DE DATOS ABIERTOS	63
6.1.	ATRIBUTOS DE CALIDAD	64
6.2.	EL PERFILEO DE DATOS	65
6.3.	EL PERFILEO DE LA INFORMACIÓN	66
6.4.	EL PERFILEO DE FUNCIÓN PÚBLICA	66
6.5.	INFORME FINAL	67
6.6.	PROCESOS DE CALIDAD	67
6.7.	PLANIFICACIÓN DE LA CALIDAD	68
6.8.	VALIDACIÓN DE LA CALIDAD	69
6.9.	CONTROL Y ASEGURAMIENTO DE LA CALIDAD DE LOS DATOS	70
7.	CALIDAD DE LOS METADATOS	72
7.1.	¿QUÉ SON LOS METADATOS?	73
7.2.	COMO DILIGENCIAR LOS METADATOS	75
7.2.1.	TÍTULO Y DESCRIPCIÓN	75
7.2.2.	ETIQUETA DE LA FILA	75
7.2.3.	CATEGORÍAS Y ETIQUETAS	75
7.2.4.	LICENCIA DE ATRIBUCIÓN	76
7.2.5.	INFORMACIÓN DE CONTACTO	76
7.2.6.	NOMBRE DEL RECURSO	76
7.2.7.	ANEXOS	76
7.2.8.	INFORMACIÓN DE LA ENTIDAD	76
7.2.9.	INFORMACIÓN DE DATOS	77
7.2.10.	URL DE DOCUMENTACIÓN Y NORMATIVA	77
8.	CLASIFICACIÓN Y PRIORIZACIÓN DE ERRORES DE CALIDAD Y ERRORES DE PUBLICACIÓN	78
8.1.	ERRORES DE PUBLICACIÓN	80
8.2.	CLASIFICACIÓN DE DATOS	84
8.3.	COMPLETITUD	85
8.4.	COMPRENSIBILIDAD	86

8.5. CONFIDENCIALIDAD	87
8.6. CONFORMIDAD	88
8.7. DUPLICIDAD	89
8.8. METADATA ERRADA, INCOMPLETA Y/O VACÍA	90
8.9. ERROR SIN FILAS	91
8.10. ERROR POCAS FILAS	92
8.11. ERROR POCAS COLUMNAS	94
8.12. ERROR POCAS COLUMNAS MAL NOMBRADAS	95
8.13. ERROR FALTA CAMPO DE GEOLOCALIZACIÓN DEL CONJUNTO DE DATOS	96
8.14. ENLACE INVALIDO	97
8.15. CONJUNTO O SUBCONJUNTO DE ERRORES	97
8.16. EL CONJUNTO DE DATOS ESTÁ MAL CARGADO	99
8.17. DESACTUALIZADO	100
8.18. ERR017 ENLACE ROTO	101
8.19. EL CONJUNTO DE DATOS PRESENTA AGREGACIONES O TOTALES	102
8.20. ITA – LEY DE TRANSPARENCIA Y DERECHO DE ACCESO A LA INFORMACIÓN PÚBLICA	103
8.21. SUBCONJUNTO DE DATOS MAESTROS	105
8.22. UNICIDAD	106
8.23. POCA REUTILIZACIÓN	107
8.24. REVISIÓN DE ERRORES	108
9. CAMBIOS EN LOS CRITERIOS DE EVALUACIÓN	110
10. GLOSARIO	115
11. REFERENCIAS	118

Lista de tablas

TABLA 1. TIPOS DE ERRORES QUE SE EVALÚAN EN LA ACTUALIDAD	79
TABLA 2. ERRORES DE PUBLICACIÓN	80
TABLA 3. ANÁLISIS DE ERRORES	84
TABLA 4. ANÁLISIS ERRORES COMPLETITUD	85
TABLA 5. ANÁLISIS ERRORES COMPRENSIBILIDAD	86
TABLA 6. ANÁLISIS ERRORES CONFIDENCIALIDAD	87
TABLA 7. ANÁLISIS ERRORES CONFORMIDAD	88
TABLA 8. ANÁLISIS ERRORES DUPLICIDAD	89
TABLA 9. METADATA ERRADA, INCOMPLETA Y / O VACÍA	90
TABLA 10. ERROR SIN FILAS	91
TABLA 11. ERROR POCAS FILAS	92
TABLA 12. ERROR POCAS COLUMNAS	94
TABLA 13. ERROR POCAS COLUMNAS MAL NOMBRADAS	95
TABLA 14. ERROR FALTA DE GEOLOCALIZACIÓN DEL CONJUNTO DE DATOS	96
TABLA 15. ENLACE INVÁLIDO	97
TABLA 16. CONJUNTO O SUBCONJUNTO DE ERRORES	97

TABLA 17. CONJUNTO DE DATOS MAL CARGADO	99
TABLA 18. DESACTUALIZADO	100
TABLA 19. ENLACE ROTO	101
TABLA 20. EL CONJUNTO DE DATOS PRESENTA AGREGACIONES O TOTALES	102
TABLA 21. EL CONJUNTO DE DATOS PRESENTA AGREGACIONES O TOTALES	103
TABLA 22. SUBCONJUNTO DE DATOS MAESTROS.....	105
TABLA 23. UNICIDAD.....	106
TABLA 24. POCA REUTILIZACIÓN.....	107
TABLA 25. REVISIÓN DE ERRORES	108
TABLA 28. TABLA DE CRITERIOS DE EVALUACIÓN DE CONFIDENCIALIDAD.....	111

Lista de ilustraciones

ILUSTRACIÓN 1 ESTÁNDAR UNIVERSAL DE CALIDAD DE DOS CAPAS.....	7
ILUSTRACIÓN 2. CICLO DE VIDA DE LOS DATOS	9
ILUSTRACIÓN 3 ACTORES CLAVE EN EL MARCO DE INTEROPERABILIDAD DE GOBIERNO DIGITAL.....	10
ILUSTRACIÓN 4 MODELO CONCEPTUAL DEL MARCO DE INTEROPERABILIDAD	12
ILUSTRACIÓN 5 CRITERIOS DE CALIDAD E INTEROPERABILIDAD	15
ILUSTRACIÓN 6 CRITERIOS DE CALIDAD	17
ILUSTRACIÓN 7 CALIFICACIÓN DE LA INFORMACIÓN DE ACUERDO CON SUS NIVELES DE SEGURIDAD	18
ILUSTRACIÓN 8 MUESTRA DE CONJUNTO DE DATOS CON INFORMACIÓN CONFIDENCIAL	21
ILUSTRACIÓN 9 MUESTRA DE CONJUNTO DE DATOS CON INFORMACIÓN NO RELEVANTE	25
ILUSTRACIÓN 10 EJEMPLO ÚLTIMA ACTUALIZACIÓN DE LOS DATOS	26
ILUSTRACIÓN 11 FRECUENCIA DE ACTUALIZACIÓN	26
ILUSTRACIÓN 12 MUESTRA DE CONJUNTO DE DATOS CON INFORMACIÓN DESACTUALIZADA.....	27
ILUSTRACIÓN 13 MUESTRA DE CONJUNTO DE DATOS CON ERROR EN EL TITULO	29
ILUSTRACIÓN 14 METADATOS A DILIGENCIAR	30
ILUSTRACIÓN 15 REVISIÓN DE METADATOS PARA EL CRITERIO DE CONFORMIDAD	32
ILUSTRACIÓN 16 REVISIÓN DE METADATOS PARA EL CRITERIO DE EXACTITUD SINTÁCTICA	34
ILUSTRACIÓN 17 ERRORES SEMÁNTICOS.....	35
ILUSTRACIÓN 18 REVISIÓN DEL CRITERIO DE COMPLETITUD	37
ILUSTRACIÓN 19 CONJUNTO DE DATOS CON COLUMNAS VACÍAS	38
ILUSTRACIÓN 20 CONJUNTO DE DATOS CON DATOS INCONSISTENTES.....	40
ILUSTRACIÓN 21 EJEMPLO ERROR PRECISIÓN	41
ILUSTRACIÓN 22 FORMATOS PARA LA PUBLICACIÓN DE DATOS ABIERTOS.....	43
ILUSTRACIÓN 23 EJEMPLO ERROR EN PORTABILIDAD	44
ILUSTRACIÓN 24 METADATOS CREDIBILIDAD.....	46
ILUSTRACIÓN 25 DESCRIPCIONES ADECUADAS PARA LAS COLUMNAS	48
ILUSTRACIÓN 26 METADATOS ENLACE DE LA FUENTE	49
ILUSTRACIÓN 27 METADATOS URL DOCUMENTACIÓN Y URL NORMATIVA	50
ILUSTRACIÓN 28 EVIDENCIA DE FILAS Y COLUMNAS DUPLICADAS	51
ILUSTRACIÓN 29 ERROR EN EFICIENCIA	52
ILUSTRACIÓN 30 ACTIVIDADES LIMPIEZA DE DATOS.....	68
ILUSTRACIÓN 31 METADATOS DEL CONJUNTO DE DATOS.....	74

1. Introducción

En los últimos años, Colombia ha fortalecido su compromiso con la apertura de datos públicos a través de la plataforma nacional www.datos.gov.co, permitiendo a las entidades del Estado publicar información de interés general de forma accesible y reutilizable. Esta práctica responde al derecho fundamental de acceso a la información y representa una herramienta clave para la toma de decisiones basadas en evidencia, la transparencia y la innovación pública.

Sin embargo, la simple publicación de datos no garantiza su utilidad. Para que los datos abiertos realmente generen valor social, económico y estratégico, es fundamental que cumplan con criterios de **calidad e interoperabilidad**. Estos atributos permiten que la información pueda ser comprendida, integrada y reutilizada por diferentes actores: desde otras entidades públicas hasta ciudadanos, investigadores y empresas.

La presente guía tiene como objetivo **brindar orientaciones prácticas a los líderes de tecnología y responsables de datos** en las entidades públicas del orden nacional y territorial, para asegurar que los conjuntos de datos abiertos cumplan con estándares básicos de calidad e interoperabilidad. Se enfoca en ofrecer un marco claro para planificar, evaluar y mejorar la publicación de datos, alineado con políticas públicas nacionales, la Hoja de Ruta Nacional y Sectoriales de Datos Abiertos Estratégicos y referentes internacionales como la Carta Internacional de Datos Abiertos y la norma ISO/IEC 25012.

Este documento proporciona **criterios de calidad comprensibles y aplicables en el contexto institucional**, con el fin de fortalecer la capacidad de las entidades para publicar datos útiles, confiables y sostenibles en el tiempo.

2. La calidad de los datos

2.1. Importancia de publicar datos abiertos de calidad

En la actualidad, la publicación de datos abiertos no se limita a liberar archivos en línea. Para que esta práctica genere valor público, es indispensable que los datos sean útiles, confiables y reutilizables, lo cual solo se logra si cumplen con estándares mínimos de calidad.

Desde la década de 1950, el concepto de calidad ha evolucionado desde su aplicación a productos físicos hacia bienes intangibles como los datos y el software. Diversos expertos y normas internacionales han aportado definiciones que hoy son base para el trabajo en entidades públicas:

- **Philip Crosby¹:** Calidad es cumplir con los requisitos definidos.
- **Wang & Strong² :** Calidad es que los datos sean aptos para el uso que se espera de ellos.
- **ISO 9001:2000³:** Calidad es el grado en que las características de un producto o servicio cumplen con los requisitos establecidos.
- **ISO 9001:2008 y 2015⁴:** La calidad incluye también la sostenibilidad, la gestión de riesgos y las expectativas de las partes interesadas

En el contexto de la gestión pública, esto significa que los datos deben estar preparados para apoyar decisiones, articular políticas, fomentar innovación y facilitar el control ciudadano.

Normas técnicas que respaldan el concepto de calidad

Este modelo, reafirmado en la norma **ISO/IEC 25012** propone un modelo de calidad de datos que distingue dos dimensiones clave⁵

- **Calidad inherente de los datos:** Qué tan bien los datos, por sí mismos, pueden cumplir su propósito.
- **Calidad dependiente del sistema:** Qué tanto el entorno tecnológico permite mantener y aprovechar esa calidad.

¹ https://es.wikipedia.org/wiki/Philip_Crosby

² <https://www.chospab.es/calidad/archivos/Documentos/NormalInternacionalISO9001.pdf>

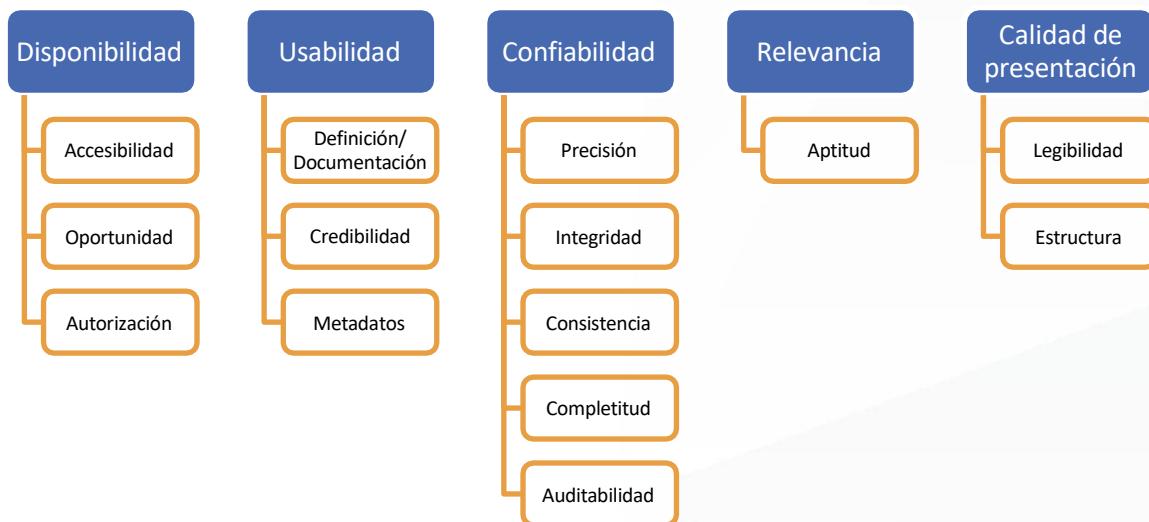
³ <https://www.iso.org/obp/ui/#iso:std:iso:9001:ed-4:v2:es>

⁴ <https://www.iso.org/obp/ui/#iso:std:iso:9001:ed-5:v1:es>

⁵ <https://iso25000.com/index.php/normas-iso-25000/iso-25012>

Este modelo, reafirmado en la norma **UNE-ISO/IEC 25012:2023⁶**, ha sido adoptado por gobiernos y organismos multilaterales como guía para evaluar y mejorar la publicación de datos.

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. & Yangon Z., 2015, Data Science Journal 8 (Cai, 2015)

Dimensiones clave de la calidad de los datos

De forma práctica, las siguientes dimensiones son fundamentales para asegurar que los datos abiertos cumplan su propósito:

- **Precisión:** Representan fielmente la realidad.
- **Integridad:** Medida en que los datos están completos y sin valores faltantes.
- **Validez:** Cumplen con formatos y reglas esperadas.
- **Consistencia:** Son coherentes entre distintas fuentes y períodos.
- **Unicidad:** Ausencia de duplicados en los datos.
- **Actualización:** Están vigentes para su uso.

⁶ <https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0071618>

- **Relevancia:** Son útiles para el análisis y la toma de decisiones⁷

Beneficios de publicar datos con calidad

Contar con datos abiertos de calidad trae beneficios claros para las entidades públicas y sus grupos de interés:

- **Mayor valor público:** La ciudadanía, la academia y las empresas pueden reutilizar los datos para crear soluciones, productos y análisis.
- **Mejor interoperabilidad:** Se reducen redundancias y se facilita el intercambio eficiente de información entre entidades.
- **Impulso a la innovación:** Tecnologías como inteligencia artificial requieren datos confiables para ofrecer resultados precisos.
- **Transparencia y legitimidad:** Datos claros y accesibles fortalecen la confianza en las instituciones.

Importancia para los líderes TIC

Para los CIOs, líderes de TI y responsables de datos, la calidad no debe verse como una tarea técnica adicional, sino como una estrategia organizacional que potencia el valor de la información que ya se produce en la entidad. Esta guía busca facilitar ese camino, proponiendo criterios claros, prácticos y adaptables para fortalecer la calidad e interoperabilidad de los datos abiertos en Colombia.

2.2. Mejoramiento de la calidad de los datos

Calidad como proceso continuo en la gestión pública

La Guía para el Uso y Aprovechamiento de Datos Abiertos en Colombia⁸ resalta que una estrategia de apertura de datos trasciende la simple preparación y publicación de información por parte de las entidades públicas. Esta estrategia exige acciones concretas orientadas a fomentar el uso efectivo de los datos publicados, realizar un seguimiento riguroso y evaluar el valor agregado que generan en su aplicación.

En este marco, la segunda fase del proceso de apertura y uso de los datos, denominada “**Monitoreo de la calidad y el uso**”, se convierte en un eje clave para maximizar el impacto

⁷ IBM Research (s.f.). *Data Quality Dimensions*.

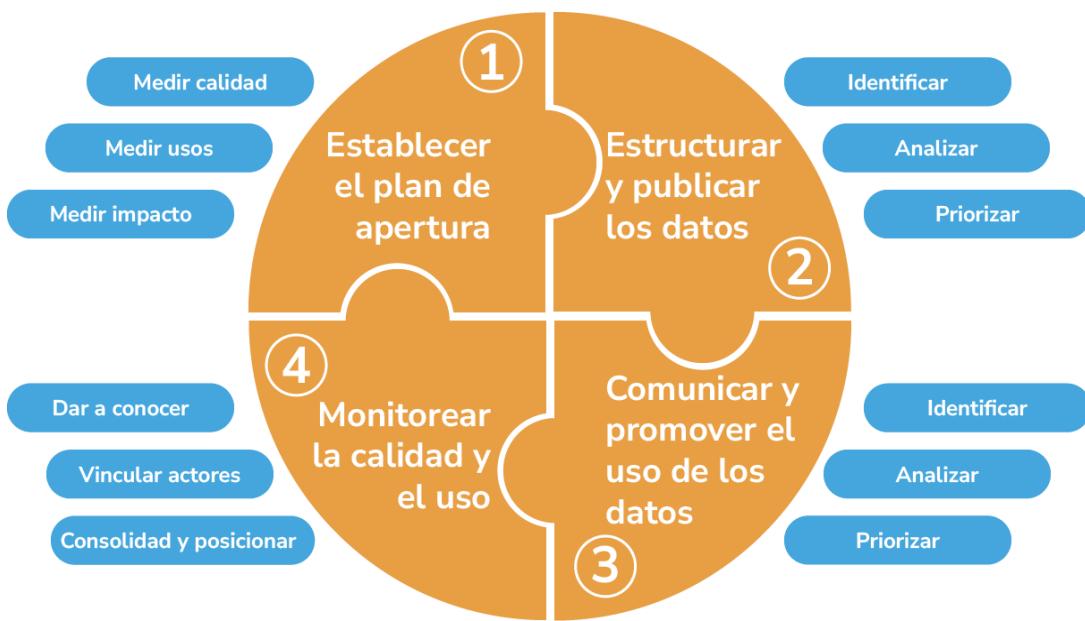
⁸ https://drive.google.com/file/d/1jiuxTrsXyz_rEtkTpG9MR3qC-yTGUhw_/view



TIC

de los datos abiertos. Esta guía desarrolla un enfoque detallado y práctico, diseñado para que las entidades cumplan con los criterios establecidos para estandarizar la calidad y fortalecer la interoperabilidad de los datos.

Ilustración 2. Ciclo de vida de los datos



Fuente. Guía de uso y aprovechamiento de datos en Colombia

Rol de los CIOs y responsables de datos

Para lograr mejoras sostenibles en la calidad de los datos abiertos, los CIOs y líderes de TI deben:

- Integrar la calidad de los datos dentro de los procesos internos de gestión de la información.
- Coordinar con equipos técnicos, jurídicos y de planeación para asegurar que los datos publicados sean útiles y cumplan con la normativa vigente.
- Asegurar que se implementen prácticas de control de calidad, desde la captura hasta la publicación de los datos.
- Promover el uso de herramientas que permitan evaluar y mejorar la calidad, pero sin

depender exclusivamente de soluciones automatizadas.

- **Nota práctica:** La guía nacional de calidad recomienda implementar un enfoque gradual, iniciando con un diagnóstico básico de los conjuntos de datos más consultados y priorizando acciones de mejora según su valor estratégico y nivel de reutilización potencial.

2.3. Marco de interoperabilidad

¿Por qué es clave la interoperabilidad en el contexto de los datos abiertos?

Publicar datos abiertos no solo implica disponibilizarlos en línea, sino también garantizar que puedan ser comprendidos, conectados e integrados por otras entidades públicas, empresas y ciudadanos. Para lograrlo, es esencial que los datos sigan principios de interoperabilidad, es decir, que puedan ser intercambiados y reutilizados entre sistemas, personas y organizaciones sin fricciones técnicas o semánticas.

La interoperabilidad promueve una visión unificada del Estado, donde las entidades trabajan de manera articulada para ofrecer servicios digitales más eficientes, confiables y centrados en el ciudadano. En este sentido, el Marco de Interoperabilidad del MinTIC es la referencia nacional para facilitar el intercambio de información entre entidades públicas, bajo cuatro dominios fundamentales: político-legal, organizacional, semántico y técnico

Objetivos del Marco de Interoperabilidad

Este marco busca:

- **Facilitar la prestación de servicios públicos digitales** basados en datos reutilizables por defecto.
- **Establecer condiciones comunes para el intercambio de información** entre entidades.
- **Promover una visión unificada del Estado**, apoyada en la colaboración digital entre instituciones.
- **Contribuir al uso estratégico de los datos abiertos**, asegurando su integración y aprovechamiento en todo el ciclo de vida de la información.

Aplicación en la calidad de los datos abiertos

La publicación de datos abiertos debe considerar los lineamientos del marco para garantizar que estos:

- Estén alineados con **estándares técnicos y semánticos comunes**.
- Sean interpretables por múltiples actores del Estado.
- Se encuentren disponibles para ser **compartidos, integrados y reutilizados**, sin barreras técnicas o legales.

- **Ejemplo:** Cuando una entidad publica un conjunto de datos sobre contrataciones públicas, debe asegurarse de que los campos como "modalidad de contratación" o "estado del contrato"



TIC

estén definidos según el **Lenguaje Común de Intercambio de Información** del MinTIC, permitiendo que otras entidades los interpreten de forma uniforme.

Cuatro dominios de interoperabilidad para la gestión de datos abiertos

Los CIOs deben comprender que la interoperabilidad va más allá del aspecto técnico. El marco se estructura en los siguientes dominios:

1. **Dominio Político-Legal:** Asegura que el intercambio de información se ajuste al marco jurídico vigente. Las entidades deben revisar y aplicar instrumentos legales (como acuerdos interinstitucionales o decretos) que permitan compartir datos sin afectar la protección de datos personales.
2. **Dominio Organizacional:** Se refiere a los procesos, roles y estructuras internas necesarias para garantizar que los servicios de datos abiertos estén alineados con los objetivos misionales de la entidad y que existan responsables claros del flujo y la calidad de la información.
3. **Dominio Semántico:** Asegura que los datos tengan un significado común entre entidades. Esto implica usar el Lenguaje Común de Intercambio de Información, publicado por el MinTIC, para estructurar y definir datos de forma estandarizada.
4. **Dominio Técnico:** Se enfoca en los mecanismos tecnológicos que habilitan el intercambio digital. Aunque esta guía no aborda aspectos técnicos, es clave que los CIOs garanticen la existencia de canales seguros y eficientes para compartir datos con otras entidades.

Ilustración 4 Modelo conceptual del Marco de Interoperabilidad



☞ Esta guía complementa estos dominios enfocándose especialmente en los aspectos

semánticos y técnicos que inciden directamente en la calidad y reutilización de los datos abiertos.

El Marco también incorpora un modelo de madurez, que permite a las entidades evaluar su progreso en la adopción de prácticas interoperables. Este modelo puede usarse para priorizar mejoras en la gestión de datos abiertos, asignar recursos y establecer metas concretas en interoperabilidad.

La gobernanza de la interoperabilidad implica definir responsables dentro de la entidad, formalizar acuerdos de intercambio y asegurar la calidad de los datos compartidos. Esto debe incorporarse al Plan Estratégico de Tecnologías de la Información (PETI) y a los lineamientos de transformación digital institucional.

Recomendaciones prácticas para entidades públicas

- Adopta los lineamientos del Marco de Interoperabilidad como referencia institucional, en especial al planear la publicación de conjuntos de datos abiertos.
- Utiliza el Lenguaje Común del MinTIC para describir y estandarizar los elementos de dato incluidos en tus conjuntos de datos.
- Revisa si tu entidad cuenta con acuerdos interinstitucionales vigentes para el intercambio de información, y ajusta su uso conforme a la ley.
- Evalúa el nivel de madurez de tu entidad usando la herramienta publicada por el MinTIC, e implementa mejoras progresivas.

Interoperabilidad como parte del ciclo de vida de los datos

Publicar datos abiertos de calidad requiere planear su interoperabilidad desde el inicio. Las entidades deben considerar:

1. Revisión normativa sobre el uso y publicación del dato.
2. Estandarización semántica (glosarios, catálogos de datos, lenguaje común).
3. Validación técnica de estructura y formato interoperable.
4. Documentación clara y completa (metadatos y licencias).
5. Monitoreo del uso e intercambio para retroalimentar mejoras.

• Puedes consultar este estándar en el portal oficial:
<http://lenguaje.mintic.gov.co>

Adoptar el Marco de Interoperabilidad no es una tarea técnica aislada. Es una decisión estratégica que permite a las entidades públicas cumplir con los principios de transparencia, eficiencia e innovación mediante el uso inteligente y articulado de la información. En el contexto de los datos abiertos, este marco es una herramienta indispensable para asegurar que la información publicada tenga calidad, utilidad y un verdadero impacto en la ciudadanía.



3. Criterios de Calidad e Interoperabilidad



Para que los datos abiertos realmente generen valor público, deben cumplir con criterios básicos de **calidad e interoperabilidad**, que permitan su comprensión, reutilización y vinculación con otros conjuntos de datos. Estos criterios funcionan como una guía práctica para que las entidades públicas evalúen, mejoren y mantengan la integridad, utilidad y accesibilidad de la información publicada.

A continuación, se presentan los principales criterios que deben ser considerados por las entidades antes, durante y después de publicar conjuntos de datos en el portal nacional datos.gov.co.

3.1. Criterios de calidad de los datos

Los criterios de calidad permiten verificar que los datos cumplan con los requisitos mínimos para ser confiables y útiles. Se recomienda evaluar los siguientes:

Tabla Criterios de calidad e interoperabilidad

#	Criterio	Descripción	Aplicación práctica
1	Accesibilidad	El conjunto puede ser consultado y descargado por quien lo necesite.	Sin requisitos de registro, contraseña ni software especial para acceder.
2	Actualidad	Vigencia y actualización de los datos publicados.	Los datos reflejan el estado más reciente según su periodicidad declarada.
3	Completitud	Todos los campos obligatorios están diligenciados.	Ningún campo crítico queda en blanco (Ej: NIT, fecha, valor).
4	Comprensibilidad	Los datos pueden ser interpretados fácilmente por cualquier usuario.	Encabezados claros, glosarios, campos con nombres comprensibles.
5	Conformidad	Cumplimiento de lineamientos y estándares vigentes.	Uso de plantillas de MinTIC, estándares abiertos y normativas institucionales.
6	Confidencialidad	Los datos solo deben ser accedidos por personal autorizado.	Accesos restringidos en sistemas internos; no publicar datos personales sin anonimizar.
7	Consistencia	Datos coherentes y sin contradicción.	Fechas no desordenadas, estados válidos, formatos coherentes en todos los registros.
8	Credibilidad	Información veraz y confiable para los usuarios.	Fuente oficial declarada y responsable institucional visible.
9	Disponibilidad	Los datos están en línea cuando se necesitan.	El enlace al conjunto funciona 24/7 y no presenta errores o caídas frecuentes.
10	Eficiencia	Plataforma permite análisis y descargas con buen rendimiento.	Portal rápido, sin errores al descargar o consultar.
11	Exactitud	Datos diligenciados correctamente.	No hay errores de escritura, ortografía o campos mal llenados.
12	Portabilidad	Formatos sin restricciones para su reutilización.	Publicación en CSV, JSON o Excel limpio, sin bloques ni macros.

13	Precisión	Nivel de desagregación de los datos es adecuado al original.	Datos publicados a nivel municipio si así fueron recolectados, no agregados por región.
14	Recuperabilidad	Capacidad de restaurar o recuperar datos si se pierden o fallan.	Copias de seguridad del dataset y control de versiones.
15	Relevancia	Los datos publicados deben ser de utilidad para los usuarios.	Conjuntos alineados con demandas ciudadanas o estratégicas del sector.
16	Trazabilidad	Histórico del conjunto de datos: fechas de creación, publicación y actualizaciones.	Registro de versiones del dataset y fechas en metadatos.
17	Unicidad	Detección de registros duplicados e identificación única.	Campos como NIT, ID o número de contrato permiten identificar únicamente registros.

Fuente. Elaboración propia.

3.2. Herramienta de evaluación de calidad

El Portal de Datos Abiertos del Gobierno Colombiano ([Datos.gov.co](https://www.datos.gov.co)) cuenta con una herramienta avanzada de evaluación que analiza cada conjunto de datos en función de los 17 criterios de calidad establecidos. Esta herramienta asigna un puntaje en una escala de 0 a 10, donde un puntaje de 10 indica que el conjunto de datos cumple plenamente con todos los estándares de calidad requeridos.

Esta Herramienta se puede encontrar en el Siguiente enlace:

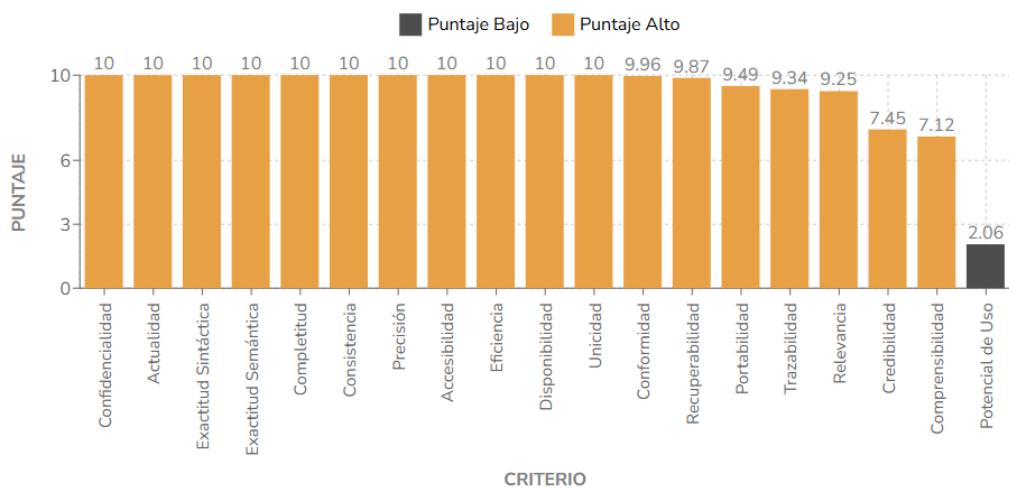
<https://www.datos.gov.co/stories/s/Informe-Calidad-de-datos-2024/ngh6-ckw5>

Este sistema de evaluación no solo mide el nivel de cumplimiento, sino que también permite determinar el potencial de uso y aprovechamiento del conjunto de datos evaluado. De esta manera, las entidades públicas pueden identificar oportunidades de mejora y garantizar que los datos publicados sean relevantes, confiables y útiles para todos los actores del ecosistema digital.

Nota: Antes de cargar un conjunto de datos en el portal, es fundamental que el área jurídica de la entidad lo valide para garantizar que la información a publicar no infrinja la ley adicionalmente que se valide que no exista duplicidad de datos con otras entidades.

Ilustración 6 Criterios de Calidad

RESULTADOS CALIFICACIONES POR CRITERIO



Fuente. Captura de imagen de informe de evaluación de muestra.

3.3. Criterio de confidencialidad

3.3.1. Definición y regulación

La confidencialidad es un principio fundamental en la publicación de datos abiertos. Su objetivo es garantizar que la información compartida no comprometa la privacidad de las personas, la seguridad institucional, ni infrinja la normatividad vigente en materia de protección de datos personales.

Antes de publicar un conjunto de datos en el portal datos.gov.co, las entidades deben asegurarse de que no se incluyan datos sensibles, personales, reservados o sujetos a restricciones legales, conforme a lo dispuesto en la Ley 1581 de 2012, la Ley 1712 de 2014 y otras normas aplicables.

La regulación respecto a la protección de datos la brinda la Ley de Transparencia y del Derecho de Acceso a la Información **ley 1712 de 2014¹¹**, la cual fue reglamentada con el **Decreto 103 de 2015¹²**, 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.

¹¹ <https://iso25000.com/index.php/normas-iso-25000/iso-25012>

¹² <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=56882>

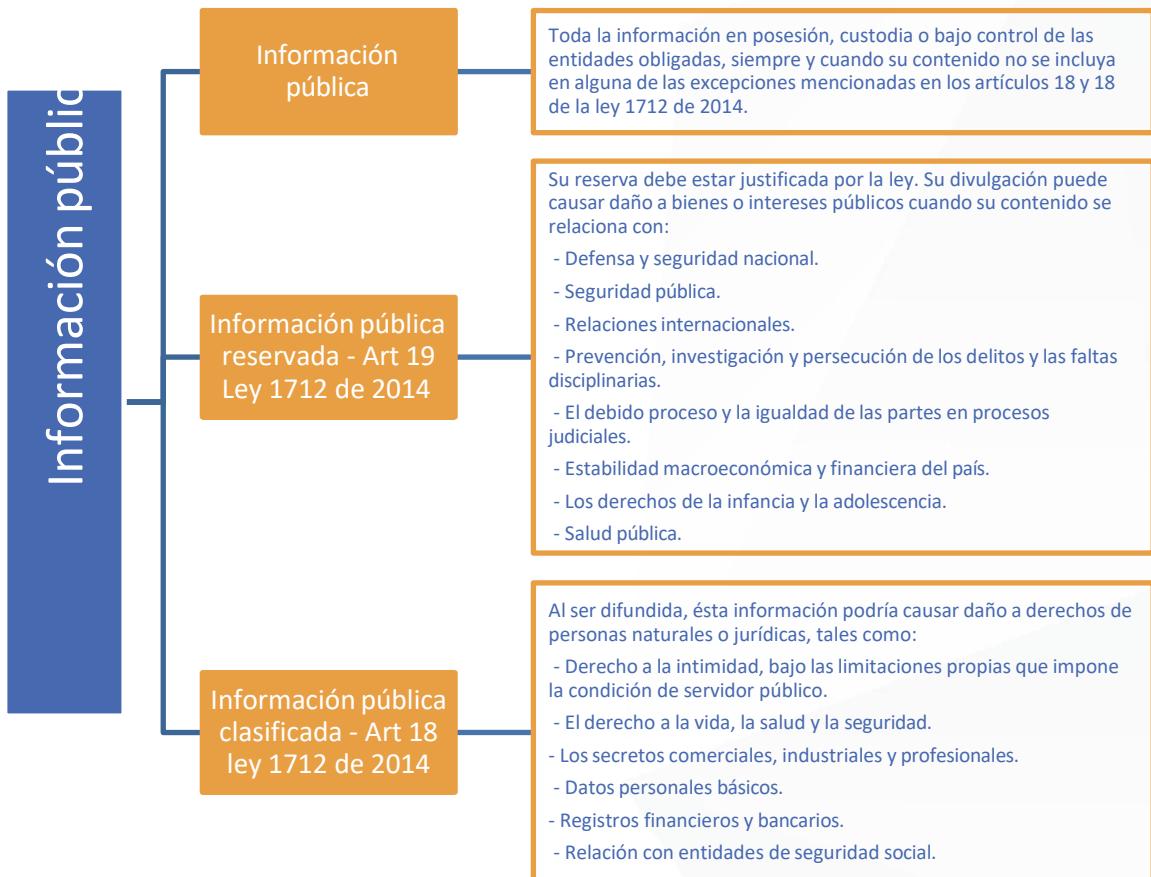
¹³ <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=60556>

Las categorías de la información que describe la Ley se presentan en la siguiente ilustración, y deben ser tenidos en cuenta para validar el criterio de confidencialidad de la información:

Ilustración 7 Calificación de la información de acuerdo con sus niveles de seguridad



Fuente. Elaboración propia.



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)

3.3.2. Cálculo del criterio

El cálculo del criterio de confidencialidad se genera en una escala de 0 a 10 sobre la base de una lista de nombres de columnas consideradas sensibles, entonces, columnas cuyo nombre coincide con palabras clave como tarjeta de identidad, documento de identidad, cuenta bancaria, número de pasaporte, dirección, o teléfono, eran definidos como atributos de alta confidencialidad debido a su potencial de identificación para un individuo. En la versión anterior el cálculo se tomaba como la relación de columnas como confidenciales o sensibles respecto a la totalidad de columnas del conjunto de datos. Para 2024, se incorpora un factor de riesgo específico para cada categoría de datos sensibles, asignando un peso que permite una evaluación más adaptativa y detallada. Se definen entonces tres niveles de riesgo:

- **Riesgo alto (3):** Datos personales tales como 'tarjeta de identidad' y 'documento de

'identidad', o datos financieros, como 'cuenta bancaria' e 'ingresos'. En el sector salud, términos como 'historial médico' y 'diagnóstico'.

- **Riesgo medio (2):** Información de contacto y domicilio para una persona natural como 'dirección de domicilio' ó 'número telefónico' se clasifican con riesgo medio (2).
Nota: Las direcciones y números de contacto de las entidades públicas no se consideran información sensible, ya que son datos de acceso público
- **Riesgo bajo (1):** Datos personales de menor sensibilidad como 'fecha de nacimiento'.

La fórmula final para el cálculo de confidencialidad es:

$$\text{confidencialidad} = \begin{cases} \frac{10}{riesgo_total} & \text{si } numColConfidencial = 0 \\ 10 - \left(\frac{10}{dfColumnas} \right) \cdot \left(\frac{riesgo_total}{numColConfidencial \cdot 3} \right) & \text{si } numColConfidencial > 0 \end{cases}$$

En la que se tienen las siguientes definiciones:

- **numColConfidencial/dfColumnas:** Proporción de columnas confidenciales respecto al total de columnas en el conjunto de datos.
- **riesgo_total:** Resultado de sumar todos los niveles de riesgo de las columnas sensibles.
- **riesgo_total/(numColConfidencial × 3):** Nivel de riesgo promedio, donde el valor máximo de riesgo es 3. Esta fracción pondera el riesgo promedio de las columnas confidenciales.
- **10 – (ajuste_por_riesgo):** Ajusta el puntaje de confidencialidad en una escala de 0 a 10, disminuyendo el puntaje inicial en función del riesgo calculado.

Las siguientes capturas de pantalla de las herramientas disponibles en el portal de datos abiertos, muestran un caso en el que el indicador de confidencialidad es bajo, debido a la existencia de columnas que refieren datos personales y que hacen fácilmente identificables a los usuarios o entidades listados en la siguiente ilustración.



TIC



Ilustración 8 Muestra de conjunto de datos con información confidencial

Persona Natural	Cédula	Nombre	Dirección de Residencia	Correo Electrónico
Persona 1	001-123-24	Carlos Gómez	Calle 15 # 1 - 2	persona1@correo.com
Persona 2	002-246-48	Ana Martínez	Calle 20 # 2 - 3	persona2@correo.com
Persona 3	003-369-72	Luis Fernández	Calle 25 # 3 - 4	persona3@correo.com
Persona 4	004-492-96	Maria López	Calle 30 # 4 - 5	persona4@correo.com
Persona 5	005-615-21	José Torres	Calle 35 # 5 - 6	persona5@correo.com
Persona 6	006-738-45	Laura Rodríguez	Calle 40 # 6 - 7	persona6@correo.com
Persona 7	007-861-69	Pedro Pérez	Calle 45 # 7 - 8	persona7@correo.com
Persona 8	008-984-93	Sofía Sánchez	Calle 50 # 8 - 9	persona8@correo.com

Fuente. Captura de pantalla de datos.gov.co

3.3.3. Anonimizar los datos

El proceso de anonimizar consiste en 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. Si bien el avance tecnológico y la información disponible en distintos medios (particularmente en Internet) dificultan un anonimato absoluto, 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.

3.4. Criterio de relevancia

3.4.1. Definición y referentes

El criterio de relevancia evalúa la medida en que los datos publicados aportan significado y generan valor para la toma de decisiones en contextos específicos, de acuerdo con el sector y propósito de la entidad. En una escala de 0 a 10, este criterio permite entonces determinar qué tan pertinentes y significativos resultan los datos para dar respuesta a la demanda que la misma entidad, otras entidades, o el público general hace de ellos. Como referentes para evaluar este criterio se toma la **Hoja de Ruta Nacional¹³** en el que se pueden encontrar entidades, temáticas y criterios de priorización de acuerdo con la Política de Gobierno Digital de MinTIC.

Otros referentes son de orden internacional y brindan homologación frente a la medición realizada en otros países. En esta categoría aparecen índices como el **Open Data Barometer¹⁴**, el **Global Open Data Index¹⁵** y el **Our Data Index¹⁶**, que plantean los siguientes temas a priorizar:

- Temáticas priorizadas en el Open Data Barometer.

⁴ <https://www.datos.gov.co/stories/s/Hoja-de-Ruta-Nacional-de-Datos-Abiertos-2024-2025/ivxt-5jyc/>

⁵ https://opendatabarometer.org/?_year=2017&indicator=ODB

⁶ <https://opendatacharter.org/resources/global-open-data-index-20162017/>

⁷ https://www.oecd.org/en/publications/2023-oecd-open-useful-and-re-usable-data-ourdata-index_a37f51c3-en.html

- 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.

3.4.2. Cálculo del criterio

El cálculo de este criterio tiene varios pasos que se describen en la sección de cálculos detallados, y en términos generales, el proceso consiste en lo siguiente:

- El primer paso consiste en confirmar si existe una categoría asociada a los datos, como Cultura, Educación, o Turismo, dentro de los metadatos del conjunto de datos. Esto significa que al momento de publicar los datos se asigne un valor válido para el metadato de **“Categoría”**. Si no lo tiene, automáticamente se asigna cero(0) al criterio.
- Si tiene un sector asignado, se realiza un análisis de coincidencia entre las cadenas de texto correspondientes a los campos de descripción y una lista predefinida de categorías, calculando un puntaje ponderado que refleja la correspondencia y calidad de los metadatos respecto a los temas de referencia.
- El análisis de coincidencia no consiste solamente en la comparación de las cadenas de texto, sino en validar a través de técnicas de procesamiento de lenguaje natural la coherencia semántica de los metadatos en bases de datos donde es importante que las categorías y descripciones sean precisas y consistentes con un conjunto de temas predefinidos, asegurando que el puntaje resultante refleje que los metadatos sean relevantes y bien categorizados.
- Adicionalmente, se confirma que exista un número suficiente de filas para que el conjunto sea considerado relevante, pues en la medida que existe un conjunto extenso de registros se pueden hacer comparaciones o análisis en el tiempo respecto al fenómeno medido.

La relevancia final se calcula con el promedio simple de la categoría válida y el resultado de la evaluación de la cantidad de filas.

$$relevancia = \frac{medidaCategoria + medidaFilas}{2}$$

En la ecuación anterior, los valores *medidaCategoria* y *medidaFilas* son el resultado de dos algoritmos que se explican a profundidad en la sección 7, y que realizan el siguiente proceso:

- **medidaCategoria:** Recibe como parámetros la categoría y la descripción del conjunto de

datos y la lista de categorías y temas disponibles, y valida que los valores de los metadatos para "categoría" y "descripción" existen y son válidos, y que los temas asociados corresponden a los temas en 'listaCategorias'. Retorna un valor en el rango de 0 a 10 que representa la calidad ponderada de la medida de correspondencia entre los metadatos y los temas de referencia.

- **medidaFilas:** Recibe el conjunto de datos y valida que posea más registros que el número mínimo (50 filas), que el número de columnas esté en un rango adecuado, y que el porcentaje de celdas no nulas sea bajo. Retorna el puntaje ponderado de la medida en un rango entre 0 y 10.

Para efectos de la mejora de este indicador, asegúrese de que asigna una categoría a su conjunto de datos, y que realiza una descripción que hace referencia a los temas relevantes para su entidad y sector. En la siguiente ilustración se muestra un caso en el que el conjunto de datos, al tener una sola fila, no genera relevancia para el estudio del conjunto de datos.

Ilustración 9 Muestra de conjunto de datos con información no relevante

ID	Nombre	Producto	Cantidad	Precio Unitario	Total	Fecha de Compra	Producto 1	Columns
1	Juan	P1	10	100	1,000	2023 Jan 01 12:00:00 AM	P1	

Fuente. Captura de pantalla de datos.gov.co

3.5. Criterio de actualidad

El criterio de actualidad evalúa en qué medida los datos reflejan información pertinente y vigente en relación con un periodo de referencia definido. La actualidad de los datos es esencial para garantizar decisiones basadas en información precisa y relevante. Este criterio permite identificar si el conjunto de datos está lo suficientemente actualizado en un contexto de uso genérico, y para su cálculo intervienen dos campos que son la “Última Actualización de Datos” y la “Frecuencia de actualización.”

En la Siguiente ilustración se evidencia donde se puede validar correctamente la última

actualización de Datos de un conjunto de datos publicado en el portal de datos abiertos:

Ilustración 10 Ejemplo última actualización de los datos



Fuente. Captura de pantalla de datos.gov.co

Es fundamental tener en cuenta que, en este criterio de actualidad, se considera el campo indicado en la ilustración. En cuanto a la “**Frecuencia de actualización**”, este campo podrá ser completado al momento de diligenciar los metadatos del conjunto de datos que se desea publicar, tal como se muestra en la siguiente ilustración.

Ilustración 11 Frecuencia de Actualización

Editar los metadatos X

* Nombre de la Entidad
 Ministerio de Tecnologías de la Info...

* Área o dependencia
 Gobierno Digital

Información de Datos

* Idioma
 Español

* Cobertura Geográfica
 Nacional

* Frecuencia de Actualización
 Anual

URL Documentación

URL Normativa

Fuente. Captura de pantalla de datos.gov.co

3.5.1. Cálculo del criterio

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. Entonces el puntaje tiene dos posibles valores de 0 o 10 determinado por las siguientes condiciones implementadas en el algoritmo de evaluación de calidad de datos:

$$\text{actualidad} = \begin{cases} 0 & \text{si } (\text{FechaActual} - \text{FechaActualizacion}) > \text{FrecuenciaActualizacion} \\ 10 & \text{si } (\text{FechaActual} - \text{FechaActualizacion}) \leq \text{FrecuenciaActualizacion} \end{cases}$$

Ilustración 12 Muestra de conjunto de datos con información desactualizada

Actualizado 20 de abril de 2024	Última actualización de los datos 4 de noviembre de 2020	Última actualización de metadatos 20 de abril de 2024
--	--	---

Fecha de creación
4 de noviembre de 2020

Fuente. Captura de pantalla de datos.gov.co

Al momento de publicar un conjunto de datos, es importante tener en cuenta las siguientes consideraciones:

- Algunos datos pierden vigencia y utilidad rápidamente, por lo que se debe establecer un límite de publicación acorde con las capacidades de la institución.
- Definir un cronograma de publicación basado en la periodicidad indicada en los metadatos, para garantizar que los datos se mantengan actualizados.
- Implementar procesos de automatización en la apertura de datos mediante ETL (Extracción, Transformación y Carga).
- Recuerde que la “**Frecuencia de actualización**” puede ajustarse en los metadatos del conjunto de datos que desee publicar o que ya esté publicado.

3.6. Criterio de trazabilidad

La trazabilidad se mide a través de los atributos que proporcionan un camino de acceso auditado al conjunto de datos y a cualquier otro cambio realizado sobre los datos en un contexto de uso específico **ISO 9001: 2008¹⁷**. La trazabilidad permite determinar el flujo de procesos y acciones que se han generado sobre un conjunto de datos a través del tiempo, lo que en la práctica significa que los conjuntos de datos se deben poder actualizar, y debe existir información de cada una de las actualizaciones. 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.

Los algoritmos que calculan este criterio analizan los metadatos completados dentro de un conjunto de datos, evaluando la proporción de atributos que han sido diligenciados en comparación con aquellos que están incompletos al momento de su apertura. La penalización es más grande en la medida en que aumenta la proporción de campos faltantes.

Una función adicional dentro del cálculo evalúa la existencia de campos como fecha de última actualización, correo electrónico del publicador y propietario, de manera que se pueda validar que los cambios realizados al conjunto de datos estén correctamente documentados y se pueda identificar plenamente la información relacionada con la actualización.

3.6.1. Cálculo del criterio

La función comienza calculando la proporción de campos faltantes (`missing_proportion`), dividiendo el número de campos nulos entre el total de atributos. Luego, aplica una penalización cuadrática sobre esta proporción (`penalty = missing_proportion2`), lo que aumenta significativamente el impacto de los campos no diligenciados en el puntaje final. Este enfoque no lineal incrementa la penalización a medida que crece la proporción de metadatos incompletos, destacando la importancia de completar todos los campos.

El puntaje final se obtiene restando esta penalización de 1 y multiplicando el resultado por 10. Un conjunto de datos completamente diligenciado recibe un puntaje de 10, mientras que el puntaje disminuye progresivamente conforme aumenta el número de campos faltantes. En caso de que el total de atributos sea igual a cero, el puntaje se asigna automáticamente como cero.

$$\text{trazabilidad} = \text{medidaPropMetaDiligenciados} \cdot 75\% + \\ \text{medidaMetaAccesoAuditado} \cdot 20\% +$$

⁸ <https://www.iso.org/obp/ui/#iso:std:iso:9001:ed-4:v2:es>

medidaTituloSinFecha · 5%

Las siguientes métricas que están presentes la ponderación para calcular la trazabilidad:

- **medidaPropMetaDiligenciados:** Calcula el puntaje de la proporción de metadatos diligenciados, aplicando una penalización no lineal para campos no diligenciados.
- **medidaMetaAccesoAuditado:** Calcula el puntaje de los metadatos de acceso auditado, incluyendo pesos y penalizaciones por combinaciones críticas de campos faltantes.
- **medidaTituloSinFecha:** Identifica si en el título se encuentran referencias a fechas o vigencias del conjunto de datos, y aplica una penalización en caso de encontrarlo. Si el título no incluye referencias a fechas o limitaciones temporales, no aplica ninguna penalización.

En las siguientes visualizaciones de conjuntos de datos, se puede evidenciar que hacen falta datos para identificar la entidad que ha realizado la publicación, y que se refieren a los datos en un periodo específico de tiempo.

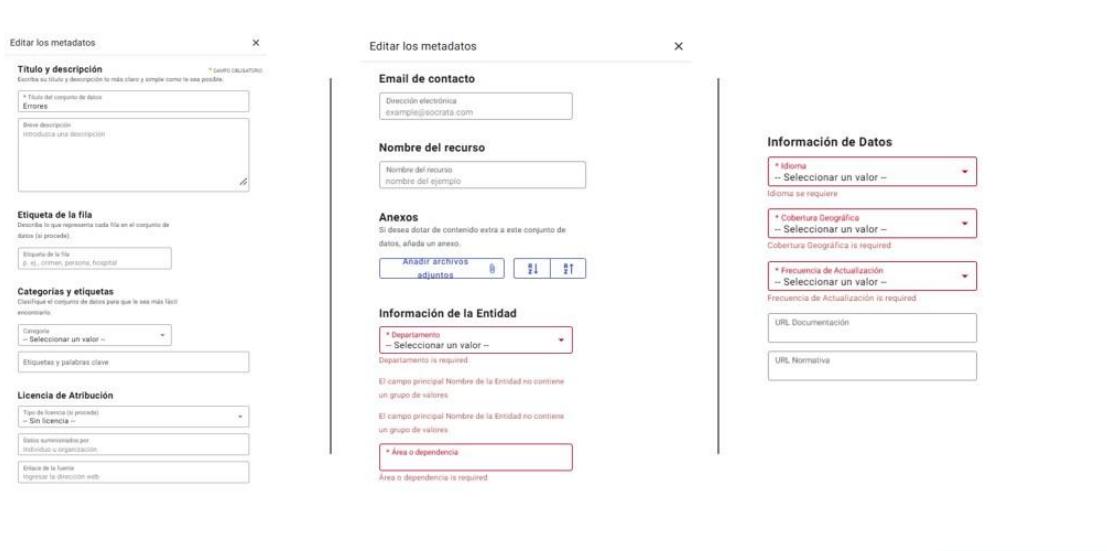
Para asegurar que no se apliquen penalizaciones en la evaluación de este criterio, asegúrese de que su conjunto de datos tenga todos los metadatos diligenciados desde el momento de su apertura, y que haga referencia a un fenómeno al que se le puede hacer seguimiento en el tiempo, por lo tanto, sea susceptible de ser actualizado.

Ilustración 13 Muestra de conjunto de datos con error en el título

Tipo	Nombre	Acciones	Último actualizado	Categoría	Propietario	Destinatarios
	Erros Año 2024 Erros	...	December 21, 2024	Ciencia, Tecnología e Innovación	 Ministerio de Tecnologías de la Información y las Comunicaciones-MINTIC	 Público

Fuente. Captura de pantalla de datos.gov.co

Ilustración 14 Metadatos a diligenciar



The image shows three separate windows of a metadata editing tool:

- Titulo y descripción:** Fields include 'Título del conjunto de datos' (Title of the dataset), 'Breve descripción' (Brief description), and 'Etiqueta de la fila' (Label for the row).
- Email de contacto:** Fields include 'Dirección electrónica' (Electronic address) and 'Nombre del recurso' (Resource name).
- Información de Datos:** Fields include 'Cobertura Geográfica' (Geographic coverage), 'Frecuencia de Actualización' (Frequency of update), 'URL Documentación' (Documentation URL), and 'URL Normativa' (Normative URL).

Fuente. Captura de pantalla de datos.gov.co

3.7. Criterio de conformidad

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 se mide a partir de la relación entre los valores correctos y los incorrectos en un conjunto de datos, y para su evaluación se aplica una función de penalización exponencial. Para mejorar la precisión y sensibilidad del cálculo de conformidad, se han implementado ajustes en la función de evaluación de columnas y en la función de criterio de conformidad.

La función de evaluación de columnas revisa aquellas consideradas como relevantes (departamento, municipio, año, latitud, longitud, y correo electrónico) y ha sido optimizada en esta versión para mejorar tanto su eficiencia como su precisión en la validación, a través de algoritmos que usan listas reutilizables de referencias previamente configuradas, que validan de manera especial los valores de coordenadas geográficas, y manejan de manera especial los valores faltantes y errores en el tipo de los datos, por ejemplo, la existencia de letras en campos donde se esperan solamente números.

3.7.1. Cálculo del criterio

El cálculo se realiza a partir de la siguiente fórmula:

$$conformidad = e^{-5 \cdot proporcionErrores}$$

$$proporcionErrores = \frac{numValoresIncorrectos}{totalValoresValidados}$$

La función de criterio de conformidad calcula la proporción de errores, definida como el cociente entre el número de valores incorrectos y el total de valores validados. Se entiende como valores incorrectos aquellos valores que no son posibles dentro de un listado de valores previamente especificados. Por ejemplo, si en el campo de “departamento” se encuentra un texto que no corresponde con ninguno de los departamentos de Colombia, será clasificado como un valor incorrecto.

La penalización exponencial implementada en este criterio se refiere a qué tan sensible es esta penalización a la cantidad de errores. Si hay pocos errores la penalización será moderada, mientras que incrementa notablemente si la cantidad de errores es mayor. Los cambios realizados en la evaluación de este criterio traen las siguientes ventajas:

- **Mayor precisión:** La función de penalización exponencial permite que el puntaje de conformidad sea más sensible a pequeños cambios en la proporción de errores. Esto asegura que los conjuntos de datos con baja calidad reciban un puntaje acorde con el nivel de errores presentes.
- **Penalización escalable y severa:** La penalización se incrementa de forma más severa cuando los errores son numerosos, lo cual es ideal para evaluar la calidad de grandes conjuntos de datos en donde la conformidad es crítica.
- **Robustez en el manejo de datos vacíos:** Se ha incorporado lógica adicional para manejar casos donde el conjunto de datos está vacío (sin filas o sin valores validados), retornando un puntaje de 0 para estos casos y evitando errores en el cálculo.
- **Consistencia en la escala de evaluación:** La fórmula asegura que el puntaje siempre se mantenga dentro del rango esperado de 0 a 10, proporcionando una métrica de conformidad clara y estándar.

Para asegurar un puntaje alto en este criterio, asegúrese de ingresar correctamente el tipo de datos solicitado en el formato correspondiente, evitando ingresar datos aleatorios o de relleno. En las siguientes ilustraciones se evidencian campos sin diligenciar, lo que lleva a una penalización en el cálculo del puntaje. Se evidencia también el tipo de datos esperado en el campo, y que se penaliza también en el caso en que esté vacío.

Ilustración 15 Revisión de metadatos para el criterio de conformidad

Departamento	Municipio	Año	Latitud	Longitud	Correo Electrónico
Antioquia	Medellín	2020	6.2442	-75.5812	juan@example.com
Bogotá		2021	4.711	-74.0721	
Desconocido	Bogotá				maria@
Cundinamarca	Soacha	Desconocido	4.589	-74.083	sofia@
Error	Error	2023	Invalid	Desconocido	
Atlántico	Barranquilla	XXXX			carlos@example
Bolívar	Cartagena	2019	10.391	-75.482	pedro@example.com
Invalid	Popayán	2025	2.444	-76.606	error@
Cauca	Pasto		Error	Error	
Desconocido		2021	-1.208	-77.281	luisa@

Fuente. Captura de pantalla de datos.gov.co

3.8. Criterio de exactitud

La exactitud es el grado en el que los datos representan correctamente el verdadero valor del atributo deseado de un concepto, o evento en un contexto de uso específico, de acuerdo con la ISO/IEC 25012:2008¹⁸ En el caso de los datos abiertos en Colombia, implica la implementación de algoritmos que realicen procesamiento de texto en español de manera que se puedan interpretar correctamente los siguientes aspectos:

- **Exactitud Sintáctica:** Evalúa la cercanía de los valores de los datos, a un conjunto de valores definidos en un dominio considerado sintácticamente correcto. Esto se refiere a la corrección formal y estructural de los datos, es decir, que cumplan con las reglas gramaticales, de formato y estructura predefinidas.
- **Exactitud Semántica:** Cercanía de los valores de los datos, a un conjunto de valores definidos en un dominio considerado semánticamente correcto. Esto significa hacer una medición de qué tan bien se encuentran representados los datos en el sistema de información.

3.8.1. Exactitud sintáctica

La función original de cálculo de exactitud sintáctica fue diseñada para evaluar la calidad de los

⁹ <https://www.iso.org/standard/35736.html>

datos textuales categóricos, identificando similitudes entre valores únicos de una columna y reportando aquellos con posibles errores. Sin embargo, para mejorar su precisión en el contexto de datos en español, se implementaron varios cambios significativos.

- **Normalización Avanzada del Texto en español:** Se utilizaron librerías especializadas para asegurar que variaciones menores, como diferencias de tildes, mayúsculas y puntuación, no se interpongan en la detección de valores similares. También se eliminan todos los signos de puntuación y caracteres especiales, dejando solo caracteres alfanuméricos. Este proceso asegura que palabras como “acción” y “Accion” se consideren equivalentes y se detecten como similares.
- **Lematización en español:** La lematización convierte cada palabra a su forma base o lema, lo que ayuda a identificar equivalencias entre términos con diferentes variaciones morfológicas, como “caminaré”, “caminando” y “cami-na”, todos los cuales se reducen a la forma base “caminar”. Esta técnica es particularmente útil en español, donde la morfología es compleja y cambia en función de tiempo, género y número.
- **Comparación Avanzada de Cadenas:** Para realizar la comparación entre valores únicos en las columnas de texto, se utilizaron algoritmos que permiten comparar conjuntos de palabras sin importar su orden, proporcionando una mayor flexibilidad y precisión en la detección de similitudes en frases. Esto es particularmente útil en datos en español, donde el orden de las palabras puede variar en expresiones comunes. Para ajustar el umbral de similitud, se usa el parámetro porcentajeSimilitudSintactica, que se multiplica por 100 para adaptarse al rango de comparación utilizado en rapidfuzz (0 a 100). Se debe asegurar de que los datos estén diligenciados correctamente para que cumplan con el criterio de exactitud.

La función de evaluación de criterios de exactitud sintáctica también fue modificada para proporcionar una evaluación más sensible y precisa de la calidad sintáctica de los datos, en esta mejora se introdujo una penalización cuadrática, lo que permite reflejar con mayor precisión el impacto de los errores en la calidad de los datos. La fórmula es la siguiente:

$$\text{exactitudSintactica} = 10 \times \left(1 - \frac{\text{numColValoresUnicosSimilares}^2}{\text{dfColumnas}} \right)$$

donde:

- **numColValoresUnicosSimilares:** es el número de columnas con valores únicos similares.
- **dfColumnas:** representa el total de columnas de texto en el conjunto de datos.

Ilustración 16 Revisión de metadatos para el criterio de exactitud sintáctica

Vista previa de la tabla

[Ver datos](#) [Crear visualización](#)

NOMBRE PRESTADOR	DIRECCION	BARRIO	CLASE PRESTADOR	LOCALIZACION
Carmen teresa cárdenas Rodríguez	CARRERA 3 # 13-12 LOCAL 56	EL TREBOL	Profesional Independiente	(4.711437°, -74.223211°)
Carola Andrea Ortegón Diaz	kr. 3 a no. 13 -12 centro comer...	el trebol	profesional independiente	(4.71115°, -74.22295°)
Carola Andrea Ortegón Diaz	diagonal 19 no. 3 c -05	villa nueva	profesional independiente	(4.7131°, -74.22055°)
catalina Alejandra Rodriguez ...	calle 5 # 02-73 cs 302	centro	profesional independiente	(4.70697°, -74.2292°)
catalina Alejandra Rodriguez ...	calle10 n: 16b -15	poblado	Profesional Independiente	(4.7217°, -74.2292°)
centro de intervención idear s...	Cra 2 N° 3 - 16 oficina 101	Centro	Profesional Independiente	(4.70538°, -74.22952°)

Fuente. Captura de pantalla de datos.gov.co

3.8.2. Exactitud semántica

La exactitud semántica evalúa qué tan correctamente los datos textuales son interpretados dentro de su contexto, aspecto fundamental en la integración semántica y el procesamiento de datos. El algoritmo para este criterio tiene como objetivo evaluar la similitud semántica entre el título y la descripción de columnas de tipo texto y sus respectivos valores en un conjunto de datos. La fórmula es la siguiente:

$$\text{exactitudSintactica} = 10 - \left(1 - \left(\frac{\text{numColNoSimSemantica}}{\text{dfColumnas}} \right)^2 \right)$$

donde:

- **numColNoSimSemantica:** Cuenta la cantidad de columnas con similitud semántica menor a 0.4.
- **dfColumnas:** Cantidad de columnas que posee el conjunto de datos.

Esta versión del algoritmo introduce un cálculo exponencial para la reducción del puntaje de exactitud semántica, permitiendo que el puntaje descienda más drásticamente cuando la proporción de columnas con baja similitud semántica es alta, y de manera menos abrupta cuando dicha proporción es baja.

En la siguiente ilustración se evidencia un ejemplo práctico de errores semánticos en tres columnas clave: "Información Ingresos Mensuales", "Temperatura Grados Celsius", y "Código del Producto Vendido". Estos errores destacan la importancia de mantener la coherencia entre los valores registrados y su contexto definido.

En la columna "Información Ingresos Mensuales", se observan valores que no son numéricos (por ejemplo, letras o palabras), lo que representa un error semántico dado que los ingresos deben ser exclusivamente valores numéricos. En "Temperatura Grados Celsius", se incluyen valores en grados Fahrenheit, que aunque son sintácticamente correctos, no son adecuados en el contexto definido de grados Celsius. Finalmente, en "Código del Producto Vendido", se presentan nombres de productos en lugar de códigos alfanuméricos, lo que rompe con la lógica esperada para esta columna.

Ilustración 17 Errores Semánticos

Información Ingresos Mensuales	Temperatura Grados Celsius	Código del Producto Vendido
3200000 (Anual)	95	Camisa
4500000	32.5	A100
5000000	100.0	Z200
5000	30	Ropa
3200000	101.0	B350
	35.5	X500
4000000	98.0	Zapatos
error	33	C400
3600000	180	D600
2900000	31	Accesorios

Fuente. Captura de pantalla de datos.gov.co

3.9. Criterio de completitud

El criterio de completitud mide la proporción de valores necesarios disponibles en los datos, reflejando su capacidad para cumplir con los requisitos de análisis. Para que un conjunto de datos cumpla a cabalidad con los criterios de calidad establecidos en esta guía, es fundamental que los datos se presenten completos para el uso de los usuarios.

Para obtener un puntaje alto en este criterio, es importante incluir opciones como "No aplica", "Otro", "No sabe", "No responde" o "No disponible" en los diccionarios de datos o etiquetas de campo de tipo texto. Esto ayuda a evitar espacios vacíos cuando no se cuenta con la información necesaria. Además, se recomienda evitar publicar conjuntos de datos con una sola columna o que estén incompletos. Si existen datos relacionados con un mismo tema, pero están distribuidos en diferentes fuentes, es mejor consolidarlos en un único conjunto de datos que sea lo más completo

posible para facilitar su uso por los usuarios.

3.9.1. Cálculo del criterio

El cálculo se realiza a partir de la siguiente fórmula:

$$\text{completitud} = \frac{\text{medidaCompletitudDatos} + \text{medidaCompletitudCol} + \text{medidaColNoVacias}}{3}$$

donde:

- **medidaCompletitudDatos:** Corresponde al cálculo del puntaje de completitud en función de la proporción de celdas no nulas en el conjunto de datos. En esta versión de la guía se ha introducido un factor de penalización exponencial que intensifica la caída del puntaje en casos de alto porcentaje de celdas:

$$\text{medidaCompletitudDatos} = 10 \times \left(1 - \left(\frac{\text{totalNulos}}{\text{totalCeldas}} \right)^{\text{penalizacionFactor}} \right)$$

en el que *penalizacionFactor* se ha establecido en 1.5. Esta modificación permite una caída más pronunciada del puntaje cuando hay una alta proporción de celdas nulas, mejorando la sensibilidad de la función en situaciones de incompletitud elevada.

- **medidaCompletitudCol:** calcula el puntaje de completitud basado en la proporción de columnas con altos valores diligenciados, que incorpora un factor de penalización de valor 2, aplicado exponencialmente sobre la proporción de columnas nulas:

$$\text{medidaCompletitudCol} = 10 \times \left(1 - \left(\frac{\text{numColPorcNulos}}{\text{dfColumnas}} \right)^{\text{penalizacionFactor}} \right)$$

en el que **numColPorcNulos** es la cantidad de columnas que poseen más de un porcentaje de datos nulos en el conjunto de datos y **dfColumnas** es el total de columnas que tiene el conjunto de datos.

- **medidaColNoVacias:** Corresponde al cálculo del puntaje de completitud en función en función de la proporción de columnas presentes respecto al total de columnas esperadas según metadatos:

$$\text{medidaColNoVacias} = 10 \times \left(\frac{\text{dfColumnas}}{\text{totColMetadatos}} \right)$$



TIC



en el que **numColPorcNulos** es la cantidad de columnas que poseen más de un porcentaje de datos nulos en el conjunto de datos y **dfColumnas** es el total de columnas que tiene el conjunto de datos.

Un ejemplo de error de completitud se observa en la ilustración siguiente, donde una columna presenta celdas vacías, lo que afecta la calidad del conjunto de datos

Ilustración 18 Revisión del criterio de completitud

Vista previa de la tabla								Ver datos	Crear visualización
CONJUNTO	TELEFONO	DIRECCION	CORREO	LATITUD	LONGITUD	LOCALIZACI...	Nueva colum...		
El Trébol MZ 2		Carrera 2 No. 12a...	eltrebol.manzana...	(4.710486,	-74.222351)	(4.710486,-74.22...	POINT (-74.2223...		
El Trébol MZ 3		Calle 12a No. 2b ...	eltrebolmz3@gm...	(4.709940,	-74.219804)	(4.709940,-74.21...	POINT (-74.2198...		
Conjunto Residen...		Calle 10 No. 4e - ..	alejandrogonro69...	(4.707663,	-74.221717)	(4.707663,-74.22...	POINT (-74.2217...		
El Trébol MZ 6		Carrera 3e No 11-..	eltrebolmz6@hot...	(4.706633,	-74.222466)	(4.706633,-74.22...	POINT (-74.2224...		
El trébol MZ 7		Calle 10 No. 4e - ..	eltrebolmanzana...	(4.705994,	-74.222891)	(4.705994,-74.22...	POINT (-74.2228...		
El trébol MZ 8	8297330	Carrera 1b Este N...	creltrebols8@gm...	(4.706588,	-74.225523)	(4.706588,-74.22...	POINT (-74.2255...		
El trébol MZ 9		Calle 8a No. 1b-E...	trebolmz9@gmail...	(4.707123,	-74.224569)	(4.707123,-74.22...	POINT (-74.2245...		
Quintas El trébol ...		Calle 8a No. 1b E...	admonquintastre...	(4.708143,	-74.224914)	(4.708143,-74.22...	POINT (-74.2249...		
Quintas El trébol ...		Calle 8a No. 2-34	quintasmz11@g...	(4.709465,	-74.222460)	(4.709465,-74.22...	POINT (-74.2224...		
Quintas El trébol ...		Calle 10 No. 1a - ..	quintaseltrebol...	(4.708502,	-74.226956)	(4.708502,-74.22...	POINT (-74.2269...		

Fuente. Captura de pantalla de datos.gov.co

La siguiente captura de pantalla en el portal de datos abiertos muestra un conjunto de datos que tiene tanto celdas vacías dentro de una columna, como columnas completas sin ningún valor diligenciado, lo que impacta significativamente en el puntaje. Previo a la publicación del conjunto de datos valide que sí tenga suficientes datos y retire aquellas columnas sobre las que no tiene datos para publicar.

Ilustración 19 Conjunto de datos con columnas vacías

Categoría de la Inf...	Nombre de la Inf...	Descripción	Área Responsable	Dueño de la Infor...	Responsable de l...	Idioma	Clasificación de l...	Fundamento Legal
Intranet	Información (Datos)	Office (Word-Excel-PP)	Intranet		Generar	Pública	Asesor de Presidencia	Presidente
Correo Electrónico	Sistema de Informació	BD	Intranet		Almacenar	Pública Clasificada	Asistente Administrati	Secretario General
Archivo Físico	Infraestructura	Correo Electrónico	Archivo Físico		Procesar	Pública Reservada	Auxiliar de Oficina Gra	Vicepresidente de Of
PC	Instalaciones	PDF	Correo Electrónico		Consultar		Conductor de Preside	Vicepresidente Técn
Sist de Info (SISE - SIA	Recurso Humano	Archivo Físico	Carpetas Compartida		Recibir		Gerente	Vicepresidente Adm
Log de Auditoria	Cintas	Medios de Alm Ext			Publicar		Gerente de Área Grad	Vicepresidente de Im
Medios de almac Ext	Imagen				Archivar		Gerente Sucursal Tipo	Vicepresidente de Py
Recurso Compartido	Audios				Enviar		Gerente Sucursal Tipo	Jefe de la OGIR
BD Misionales	Videos				Verificar		Gerente Sucursal Tipo	Jefe de la OTI

Fuente. Captura de pantalla de datos.gov.co

3.10. Criterio de consistencia

La consistencia asegura que la información sea coherente y libre de conflictos entre diferentes fuentes y estructuras, fortaleciendo su utilidad en análisis integrados. Se entiende como consistencia cuando los datos 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, y se refiere también 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, como fechas

El criterio será evaluado con una escala más alta si, por ejemplo, para el campo “Departamento” el valor utilizado es CUNDINAMARCA en todos los casos. Por el contrario, si en un conjunto de datos aparecen filas donde el campo “Departamento” tiene CUNDINAMARCA, cmarca, o C/MARCA, que se pueden interpretar como el mismo departamento, pero que en el conjunto de datos tiene múltiples textos refiriéndose a lo mismo. Aplica con mayor rigurosidad, para las codificaciones que están estandarizadas, como los códigos de la división político-administrativa del país.

3.10.1. Cálculo del criterio

El cálculo se realiza a partir de la siguiente fórmula:

$$consistencia = \frac{exactitudSintactica + medidaConsistenciaCar + atributonombresColDuplicadas}{3}$$

donde:

- **exactitudSintactica:** Corresponde al criterio de exactitud sintáctica explicado en la sección 2.7 y que analiza la validez del campo respecto a su formato y valores esperados.
- **medidaConsistenciaCar:** este valor proporciona un puntaje de consistencia robusto no solo basado en la longitud de los textos, sino también en la similitud de sus contenidos que refleja el nivel de coherencia entre los valores textuales, valores de texto con una cantidad de caracteres muy corta en casos donde se esperan descripciones o textos más extensos, serán penalizados pues no estarán cumpliendo con su propósito de brindar la suficiente información. El algoritmo para calcular este valor evalúa la longitud y similitud de los textos en cada columna para determinar si existe suficiente variabilidad y consistencia en el contenido textual. Se aplica también un factor de penalización de 2 para amplificar la penalización en casos de mayor inconsistencia.
- **atributonombresColDuplicadas:** Este puntaje es calculado según la cantidad de columnas duplicadas en el conjunto de datos, donde un mayor número de columnas duplicadas disminuye el puntaje.

Para tener un puntaje alto en este criterio, se recomienda validar cuáles son los campos comunes entre conjuntos de datos y cuáles son los valores posibles, de manera que sólo exista una versión o dato válido para cada opción. Realizar tablas con datos válidos y usar herramientas de validación de datos (dentro del mismo Excel, por ejemplo) previo a la publicación de los datos, ayudará a identificar múltiples valores que se refieren al mismo concepto. La siguiente ilustración muestra un conjunto de datos en el que los valores de talla registrada no son consistentes con valores esperados respecto a lo que indica la descripción.

Ilustración 20 Conjunto de datos con datos inconsistentes

ID	Departamento	Fecha Registro	Código Producto
1	CUNDINAMARCA	2023-01-15	P001
2	cmarca	15/01/2023	P001
3	C/MARCA	2023-01-15	P1
4	CUNDINAMARCA	15-Jan-2023	P001
5	Cundinamarca	2023/01/15	001P
6	cundinamrca	2023-01-15	P001
7	CUNDINAMARCA	15-01-2023	001-P
8	c/Marca	2023-01-15	P001

Fuente. Captura de pantalla de datos.gov.co

3.11. Criterio de precisión

La precisión evalúa la exactitud de los datos respecto a las entidades del mundo real, siendo clave para el desarrollo de sistemas basados en datos según lo estipulado por la norma **ISO/IEC 25012:2008**¹⁹ El mecanismo de evaluación original calculaba la proporción de columnas en un conjunto de datos que contienen más de un valor único y esto se utilizaba como criterio para estimar la calidad de la variabilidad de los datos.

El algoritmo fue actualizado de manera que incluye criterios adicionales en la evaluación de la calidad de los datos:

- **Varianza:** se considera la variabilidad de los valores en cada columna, utilizando un umbral mínimo de varianza (varianza_columna). Esto ayuda a determinar si una columna tiene una distribución de valores diversa.
- **Valores únicos:** se establece un mínimo de valores únicos (min_unicos) para asegurar que cada columna tenga suficiente diversidad en sus datos.

3.11.1. Cálculo del criterio

¹⁰ <https://www.iso.org/standard/35736.html>

El cálculo se realiza a partir de la siguiente fórmula:

$$consistencia = \frac{columnas_cumplen_criterios}{Total\ de\ columnas}$$

donde:

- **columnas_cumplen_criterios:** Corresponde al valor mínimo de las columnas que cumplen con el criterio de varianza mínima (columnas_varianza_suficiente) y las columnas que cumplen con el criterio de valores únicos.
- **dfColumnas:** es el total de columnas que tiene el conjunto de datos.

Cada columna se evalúa de manera independiente en función de estos criterios. Para los valores únicos, si el número de valores es mayor o igual al umbral *min_unicos* (2), la columna cumple el criterio de diversidad. Para la varianza, la función intenta convertir cada columna a un formato numérico, y si su varianza es mayor o igual al umbral (0.1), la columna cumple el criterio de variabilidad. El cálculo de varianza se realiza solo para las columnas que pueden convertirse numéricamente

A continuación, se presenta un ejemplo que resalta la importancia de evaluar la diversidad y variabilidad en las columnas de un conjunto de datos. Por ejemplo, si se define un umbral de al menos 3 valores únicos y una varianza mínima de 0.1 para las columnas numéricas, algunas columnas no cumplen con estos criterios, evidenciando problemas en la calidad de la información.

Ilustración 21 Ejemplo error precisión

ID	Edad (años)	Ingreso Mensual (\$)	Ciudad de Residencia	Estado Civil
1	25	3,000	Bogotá	Soltero
2	25	3,000	Bogotá	Soltero
3	40	4,500	Medellín	Casado
4	35	3,200	Cali	Soltero
5	35	3,200	Cali	Soltero
6	29	2,800	Bogotá	Casado
7	29	2,800	Bogotá	Casado
8	40	4,700	Medellín	Casado

Fuente. Captura de pantalla de datos.gov.co

3.12. Criterio de portabilidad

La portabilidad mide la facilidad de transferencia e interoperabilidad de los datos en diferentes sistemas, un atributo clave para la interoperabilidad tecnológica según la norma **ISO/IEC 25010:2008²⁰**. 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. La evaluación de este criterio incluye el análisis de tres aspectos:

- **Presencia de caracteres especiales:** La función calcula el porcentaje de columnas que no contienen caracteres especiales, ya que estos pueden afectar la interoperabilidad y el procesamiento de los datos. Según el porcentaje de columnas "limpias" de caracteres especiales, se asignan puntos de manera progresiva:
 - Si más del 90 % de las columnas están libres de caracteres especiales, se asignan 3 puntos.
 - Si el porcentaje está entre el 60 % y el 90 %, se otorgan 2 puntos.
 - Entre el 30 % y el 60 % se asigna 1 punto.
 - Si menos del 30 % de las columnas están libres de caracteres especiales, no se otorgan puntos.
- **Ausencia de valores nulos:** Para asegurar la consistencia, se evalúa si el conjunto de datos está libre de valores nulos. En caso afirmativo, se asignan 4 puntos, ya que la ausencia de valores nulos es fundamental para la portabilidad y para reducir posibles errores en el análisis de los datos.
- **Tamaño del conjunto de datos:** Se evalúa el tamaño del conjunto en megabytes. Conjuntos de datos de menor tamaño son generalmente más fáciles de manejar y transportar. La función asigna puntos según el tamaño:
 - Conjuntos menores a 500 MB obtienen 3 puntos.
 - Conjuntos entre 500 MB y 1 GB obtienen 1 punto.
 - Conjuntos mayores a 1 GB no reciben puntos.

3.12.1. Cálculo del criterio

El cálculo se realiza a partir de la siguiente fórmula en la que se ponderan los tres criterios descritos anteriormente:

$$\text{total_portabilidad} = \text{portabilidad} \cdot 50\% + \\ \text{conformidad} \cdot 25\% + \\ \text{completitud} \cdot 25\%$$

Este criterio se refiere además a la facilidad con la que el conjunto de datos se puede procesar,

²⁰ <https://www.iso.org/standard/35736.html>

acceder y utilizar fácilmente descargándolo o consumiéndose 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.

tras

Ilustración 22 Formatos para la publicación de datos abiertos



Fuente. Elaboración propia

En la siguiente ilustración se evidencian principales errores identificados se encuentran la presencia de caracteres especiales en la columna "Nombre del Archivo", como #, \$, @, %, &, y *. Estos caracteres pueden generar problemas de interoperabilidad, ya que no siempre son compatibles con los nombres de archivo aceptados por ciertos sistemas operativos. Además, se detectaron valores nulos en las columnas "Tipo de Archivo" y "Tamaño (MB)"

Ilustración 23 Ejemplo Error en Portabilidad

ID	Nombre del Archivo	Tipo de Archivo	Tamaño (MB)
1	ventas_anuales	CSV	50
2	datos#2023	Null	Null
3	clientes\$lista	JSON	100
4	productos_info	CSV	45
5	empleados	XML	30
6	ingresos@2022	Null	60
7	ventas%data	JSON	85
8	transacciones	CSV	40

Fuente. Captura de pantalla de datos.gov.co

3.13. Criterio de credibilidad

Por definición de la norma **ISO/IEC 25012:14²¹** 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.

Para efectos de esta guía, el cálculo criterio de credibilidad se basa en tres componentes clave: la proporción de columnas con descripciones válidas, la validez de los metadatos asociados al publicador y las descripciones completas de los datos.

3.13.1. Cálculo del criterio

El cálculo se realiza a partir de la siguiente fórmula en la que se ponderan los tres criterios

²¹ <https://iso25000.com/index.php/normas-iso-25000/iso-25012>

mencionados previamente:

$$\begin{aligned} \text{credibilidad} = & \text{ medidaMetadatosCompletos} \cdot 70\% + \\ & \text{ medidaPublicadorValido} \cdot 5\% + \\ & \text{ medidaColDescValida} \cdot 25\% \end{aligned}$$

Donde:

- **medidaMetadatosCompletos:** La función evalúa la presencia de información en los metadatos sobre las fuentes de información, documentación, normatividad, origen y/o entidad publicadora de los datos, entre otros aspectos.
- **medidaPublicadorValido:** La función evalúa si en el conjunto de datos se puede identificar un publicador y su correo electrónico.
- **medidaColDescValida:** Esta función calcula el puntaje de acuerdo con la proporción de columnas con descripciones válidas. Aplica una penalización cuadrática para reflejar una mayor sensibilidad a las descripciones faltantes, y considera un control adicional para manejar casos en los que el número de columnas válidas sea negativo debido a inconsistencias en los datos.

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.

En las siguientes imágenes se puede identificar la sección en la que se debe realizar el respectivo diligenciamiento en los metadatos para cumplir con este criterio.

Ilustración 24 Metadatos Credibilidad

Editar los metadatos X

Licencia de Atribución

Tipo de licencia (si procede)
-- Sin licencia --

Datos suministrados por
MinTIC

Enlace de la fuente
<https://www.datos.gov.co/>

Email de contacto

Dirección electrónica
datosabiertos@mintic.gov.co

Nombre del recurso

Nombre del recurso
nombre del ejemplo

Fuente. Captura de pantalla de datos.gov.co

3.14. Criterio de Comprensibilidad

El criterio de comprensibilidad evalúa la calidad de las descripciones y etiquetas en los datos, asegurando que sean claras, completas y progresivamente mejores. Antes, las evaluaciones eran rígidas: las descripciones o etiquetas cumplían un estándar mínimo y recibían un puntaje máximo o eran descartadas por completo. Ahora, se introdujeron métodos más flexibles: para las descripciones, una función exponencial asigna puntajes proporcionales a su longitud, incluso si no alcanzan el tamaño ideal. Para las etiquetas, una función logarítmica ajusta el puntaje según su extensión, fomentando mejoras continuas. Estos cambios eliminan evaluaciones binarias y promueven una mejora incremental, incentivando la creación de metadatos más comprensibles y útiles.

3.14.1. Cálculo del criterio

El cálculo se realiza a partir de la siguiente fórmula donde se tienen en cuenta los siguientes criterios

- **medidaDescExt:** La función evalúa si la descripción tiene un mínimo número de caracteres y se puede considerar lo suficientemente extensa.

$$puntajeMedidaDescExt = 10 \times (1 - \exp(-0,05 \times length))$$

- **medidaEtiquetaFila:** La función evalúa si la descripción tiene un mínimo número de caracteres y se puede considerar lo suficientemente extensa.

$$puntajeMedidaEtiquetaFila = 10 \times \frac{\log(1 + (length - 2))}{\log(1 + (max_length - 2))}$$

Para obtener un puntaje alto en el criterio de comprensibilidad, es importante seguir estas recomendaciones:

- Asegúrese de que el título del conjunto de datos sea claro, completo y fácil de entender.
- Proporcione una descripción detallada que amplíe el contexto del título, explicando siglas o términos complejos para facilitar la comprensión de la información en el conjunto de datos.
- Asigne nombres claros y descriptivos a los campos, evitando ambigüedades.
- Si utiliza un sistema codificado para nombrar los campos, como $P1, P2, P3$ para preguntas de una encuesta, incluya en los metadatos una definición clara de estos códigos.
- Añada una descripción narrativa para cada campo en los metadatos, con un mínimo de 15 caracteres, evitando simplemente repetir el nombre de la columna.
- En el campo "Etiqueta de la fila", incluya una descripción clara de lo que representa cada registro o fila dentro del conjunto de datos.

Ilustración 25 Descripciones adecuadas para las columnas

 Fecha de Generación de la información (AAAA/MM/DD)	Fecha en que se genera la información	fecha_de_generaci_n_de_la	<u>Marca de tiempo variable</u>
 Nombre del responsable de producción de la información (Propietario)	Entidad responsable de producir la información	nombre_del_responsable	<u>Texto</u>
 Nombre del responsable de la información (Custodio)	Entidad responsable de custodiar la información	nombre_del_responsable_de	<u>Texto</u>
 Objetivo legítimo de la excepción	Norma que legitima la excepción	objetivo_leg_timo_de_la	<u>Texto</u>
 Fundamento constitucional o legal	Norma constitucional	fundamento_constitucional	<u>Texto</u>
 Fundamento jurídico de la excepción	Norma jurídica	fundamento_jur_dico_de_la	<u>Texto</u>
 Excepción Total o Parcial	Si la excepción es total o parcial	excepci_n_total_o_parcial	<u>Texto</u>
 Fecha de la calificación (AAAA/MM/DD)	Fecha de clasificación del activo	fecha_de_la_calificaci_n	<u>Marca de tiempo variable</u>

Fuente. Captura de pantalla de datos.gov.co

Nota: Asegúrese de que las siglas utilizadas sean claras y fáciles de entender, y revise que no existan errores ortográficos, ya que estos pueden afectar la comprensión y credibilidad del conjunto de datos.

3.15. Criterio de accesibilidad

El criterio de accesibilidad evalúa si los datos incluyen metadatos que permiten a los usuarios encontrar y acceder fácilmente a la información, específicamente a través de etiquetas de búsqueda y vínculos de atribución. 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, por lo que este criterio es fundamental para medir la facilidad con la que los datos pueden ser utilizados y consultados. Por su naturaleza, la plataforma de Datos Abiertos de Colombia permite:

- Publicar los datos en un portal central para que los datos abiertos se puedan encontrar fácilmente y estén accesibles en un solo lugar.
- 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.

- Liberar los datos de manera gratuita, sujetos a una licencia abierta y sin restricciones.
- Liberar los datos sin registro obligatorio, permitiendo a los usuarios escoger y descargarlos, sin requerir que se identifiquen.
- Permite asegurar que los datos puedan ser accesibles y usados eficazmente por el más amplio rango de usuarios.

3.15.1. Cálculo del criterio

El cálculo se realiza a partir de validar la cantidad de etiquetas y la cantidad de vínculos de atribución. Para cada uno de ellos se valida que al menos exista uno para asignar un puntaje de 5, de manera que si tiene una etiqueta y un vínculo de atribución, se tiene el puntaje máximo de 10. En caso contrario, se asigna un puntaje de 0.

$$\begin{aligned} \text{accesibilidad} &= \text{puntaje_tags} + \text{puntaje_link} \\ 0 &\quad \text{si cantidadEtiquetas} = 0 \\ &\quad \{ 5 \quad \text{si cantidadEtiquetas} > 0 \\ \text{puntaje_link} &= \{ 0 \quad \text{si cantidadVinculos} = 0 \\ &\quad 5 \quad \text{si cantidadVinculos} > 0 \end{aligned}$$

En la siguiente ilustración se evidencian los campos a diligenciar en los cuales se podrán agregar links permiten a los usuarios encontrar y acceder fácilmente a la información, específicamente a través de etiquetas de búsqueda y vínculos de atribución

Ilustración 26 Metadatos enlace de la fuente

Editar los metadatos X

Lienzos y paquetes web

Licencia de Atribución

Tipo de licencia (si procede)
-- Sin licencia --

Datos suministrados por
MinTIC

Enlace de la fuente
Ingresar la dirección web

Fuente. Captura de pantalla de datos.gov.co

Ilustración 27 Metadatos URL documentación y URL normativa

Editar los metadatos X

* Nombre de la Entidad
Ministerio de Tecnologías de la Info...

* Área o dependencia
Gobierno Digital

Información de Datos

* Idioma
Español

* Cobertura Geográfica
Nacional

* Frecuencia de Actualización
Anual

URL Documentación

URL Normativa

Fuente. Captura de pantalla de datos.gov.co

3.16. Criterio de unicidad

El criterio de unicidad mide el grado en que los datos no contienen duplicados, tanto a nivel de filas como en columnas clave. Esto es fundamental para garantizar la integridad de los datos y evitar redundancias que pueden afectar la calidad y precisión de los análisis.

3.16.1. Cálculo del criterio

El cálculo actual del criterio introduce un parámetro adicional, *nivel_riesgo*, que permite ajustar la penalización de los duplicados según la importancia o riesgo que representan. La función calcula la proporción de filas únicas y la eleva a la potencia de *nivel_riesgo*, de modo que un nivel de riesgo alto genera una penalización mayor, útil cuando los duplicados son críticos. En cambio, un nivel de riesgo bajo suaviza la penalización, lo que es adecuado cuando los duplicados son menos problemáticos. Este nuevo enfoque mejora la precisión del cálculo, permitiendo adaptarse mejor a distintos contextos de calidad de datos.

$$\text{unicidad} = \frac{\text{proporcionFilasDuplicadas} + \text{proporcionColumnasDuplicadas}}{2}$$

La siguiente ilustración muestra ejemplos de errores de unicidad, como valores duplicados en la columna ID y registros redundantes en Nombre. Además, incluye columnas duplicadas ("Producto1" y "Fecha de Compra 1"), lo que evidencia cómo la falta de unicidad puede afectar la integridad y calidad de los datos.

Ilustración 28 Evidencia de filas y columnas duplicadas

ID	Nombre	Producto	Fecha de Compra	Producto 1	Fecha de Compra 1
1	Juan	P1	2023 Jan 01 12:00:00 AM	P1	2023 Jan 01 12:00:00 AM
2	Maria	P2	2023 Jan 02 12:00:00 AM	P2	2023 Jan 02 12:00:00 AM
3	Carlos	P3	2023 Jan 03 12:00:00 AM	P3	2023 Jan 03 12:00:00 AM
4	Ana	P4	2023 Jan 04 12:00:00 AM	P4	2023 Jan 04 12:00:00 AM
5	Juan	P1	2023 Jan 01 12:00:00 AM	P1	2023 Jan 01 12:00:00 AM
1	Juan	P1	2023 Jan 01 12:00:00 AM	P1	2023 Jan 01 12:00:00 AM
6	Sofia	P5	2023 Jan 05 12:00:00 AM	P5	2023 Jan 05 12:00:00 AM
2	Maria	P2	2023 Jan 02 12:00:00 AM	P2	2023 Jan 02 12:00:00 AM
7	Luis	P6	2023 Jan 06 12:00:00 AM	P6	2023 Jan 06 12:00:00 AM

Fuente. Captura de pantalla de datos.gov.co

3.17. 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, y en la versión anterior de la guía se asignaba por defecto el máximo valor. Para la versión actualizada, la medida de la eficiencia se ha relacionado con los elementos propios del conjunto de datos que impactan los recursos necesarios para el procesamiento y el análisis de los datos.

3.17.1. Cálculo del criterio

El cálculo actual del criterio se realiza como el promedio de tres medidas derivadas de criterios previos, la medida de completitud en datos, la medida de columnas no duplicadas y la medida filas no duplicadas.

$$\text{eficiencia} = \frac{\text{completitud} + \text{medidaFilasDuplicadas} + \text{medidaColumnasDuplicadas}}{3}$$

En la siguiente ilustración se evidencia errores relacionados con el criterio de eficiencia, como filas y columnas duplicadas. Por ejemplo, la columna "Producto 1" repite información ya presente en "Producto", lo que genera redundancia y afecta el análisis de los datos. Además, se observan registros repetidos en columnas clave como "Nombre" y "Fecha de Compra", lo que puede comprometer la integridad y precisión del procesamiento de los datos. Estos errores destacan la necesidad de eliminar duplicados y garantizar la completitud para optimizar la calidad del conjunto de datos.

Ilustración 29 Error en eficiencia

ID	Nombre	Producto	Cantidad	Precio Unitario	Total	Fecha de Compra	Producto 1
1	Juan	P1	10	100	1,000	2023 Jan 01 12:00:00 AM	P1
2	Maria	P2	5	200	1,000	2023 Jan 02 12:00:00 AM	P2
3	Carlos	P3	8	150	1,200	2023 Jan 03 12:00:00 AM	P3
4	Ana	P4	2	300	600	2023 Jan 04 12:00:00 AM	P4
5	Luis	P5	7	250	1,750	2023 Jan 05 12:00:00 AM	P5
6	Sofia	P6	4	400	1,600	2023 Jan 06 12:00:00 AM	P6
7	Elena	P7	6	350	2,100	2023 Jan 07 12:00:00 AM	P7

Fuente. Captura de pantalla de datos.gov.co

3.18. Criterio de recuperabilidad

Comprendida previamente como la relación con los programas, software, plataformas digitales y aplicaciones que permiten mantener y preservar un nivel específico de operaciones y calidad de los datos, se ha definido como el promedio de la medida metadatos completos, la medida de accesibilidad, y la medida de metadatos con acceso auditado. Al combinar estas medidas que se refieren a la gestión de actualizaciones de los datos por parte de la entidad y la buena gestión de

los metadatos, se logra una evaluación de la recuperabilidad que refleja la facilidad de acceso y comprensión de la información en función de la completitud y accesibilidad de los metadatos.

$$\text{recuperabilidad} = \frac{\text{accesibilidad} + \text{medidaMetadatosCompletos} + \text{metadatosAuditados}}{3}$$

Para cumplir con el criterio de recuperabilidad en un conjunto de datos, es necesario garantizar que los metadatos estén completos, asegúrese de que los metadatos incluyan toda la información necesaria para describir, identificar y contextualizar el conjunto de datos. Esto incluye títulos, descripciones, etiquetas claras y definiciones de campos.

3.19. Criterio de disponibilidad

Comprendida previamente como la medida de la garantía para que los usuarios autorizados tengan acceso a la información y a otros activos asociados en su contexto de uso, se ha migrado su definición al promedio las medidas de actualidad accesibilidad. Este enfoque es útil para medir la vigencia y accesibilidad simultáneamente, garantizando que los datos estén disponibles y actualizados para el usuario y contando con la operación adecuada de la infraestructura y sistemas de información que soporta la plataforma de datos abiertos.

$$\text{disponibilidad} = \frac{\text{accesibilidad} + \text{actualidad}}{2}$$

Para cumplir con el criterio de disponibilidad, es crucial garantizar que se cumplan tanto el **criterio de accesibilidad** como el **criterio de actualidad**, ya que juntos aseguran que los datos sean útiles y confiables para los usuarios.

1. **Accesibilidad:** Este criterio evalúa si los datos son fáciles de encontrar, acceder y utilizar. Es fundamental que los datos incluyan metadatos claros, etiquetas de búsqueda y estén disponibles en formatos abiertos. Además, deben publicarse en un portal central, sin barreras administrativas o burocráticas, y ofrecerse de manera gratuita y sin registro obligatorio, permitiendo el acceso al mayor número posible de usuarios.
2. **Actualidad:** Evalúa si los datos reflejan información pertinente y vigente, tomando en cuenta su última actualización y la frecuencia con la que se actualizan. Este criterio garantiza que los datos proporcionen información relevante y precisa para apoyar decisiones basadas en datos recientes.

Cumplir con estos dos criterios permite garantizar la **disponibilidad**, asegurando que los datos estén accesibles, actualizados y listos para ser consultados por cualquier usuario en cualquier momento, maximizando su utilidad y pertinencia.

4. Sellos de calidad

Los sellos de calidad brindan una visión global de la calidad de los datos, y reflejan el detalle que la entidad ha puesto en cumplir cada uno de los criterios definidos en la sección 2 de esta guía. Para obtener los sellos de calidad se consideran diversos factores: la calificación de los criterios, la calificación del incentivo de uso mediante un modelo de aprendizaje automático XGBoost, la validez de la licencia y del enlace de los datos, así como la calificación final de la calidad de los mismos. Los Sellos de Calidad se asignan en categorías que van desde la más baja, la categoría cero, hasta la más alta, la categoría tres.

4.1. Sello de Calidad 0

El nivel cero de los Sellos de Calidad se asigna cuando el conjunto de datos no cumple con los requisitos mínimos establecidos para obtener un sello superior. En efecto, esto puede ocurrir si alguno de los criterios evaluados no alcanza los valores umbrales necesarios. En el proceso de evaluación, se analizan diferentes aspectos vitales, como la actualidad, consistencia, completitud, trazabilidad, credibilidad, comprensibilidad y portabilidad de los datos. Además, se verifica la validez de la licencia y la calificación final. Si alguno de estos aspectos no cumple con los criterios mínimos, se asigna el nivel cero, indicando que los datos no alcanzan un nivel aceptable de calidad. Específicamente, para que los datos puedan obtener un sello de calidad uno (1), deben cumplir con los siguientes requisitos: la calificación de actualidad debe ser igual a 10, la consistencia, completitud y trazabilidad deben ser al menos 8, la credibilidad y comprensibilidad deben ser al menos 7, y la portabilidad debe ser al menos 5. Además, la validez de la licencia debe ser confirmada (valor igual a 1) y la calificación final debe ser al menos 7. Si alguno de estos criterios no se cumple, el nivel asignado será cero, indicando que los datos no alcanzan los estándares mínimos requeridos.

4.2. Sello de Calidad 1

El nivel de calidad de calidad 1 corresponderá a aquellos valores que cumplan con los siguientes criterios mínimos.

- **Actualidad:** Los datos deben reflejar la realidad de manera inmediata. Es necesaria una calificación de 10, lo que implica actualizaciones frecuentes y procesos para garantizar que la información se mantenga vigente.
- **Consistencia y Completitud:** Los datos deben mantener una calificación de consistencia 8 y una completitud de al menos 80 %, garantizando que sean fiables y estén alineados con las fuentes originales.

- **Metadatos Documentados:** Debe alcanzarse un 80% o más de metadatos completos, con una calificación mínima de 7 en trazabilidad, credibilidad y comprensibilidad. Para lograrlo, es recomendable elaborar diccionarios de datos, descripciones claras de las variables, historial de modificaciones y contexto del origen de la información.
- **Licencia de Uso:** La licencia debe estar claramente definida en los metadatos. Al menos debe indicarse el tipo o nombre de la licencia empleada, permitiendo a los usuarios conocer las condiciones de uso, distribución y adaptación del conjunto de datos.
- **Calificación Final:** El promedio de las calificaciones debe ser igual o superior a 7, reflejando un cumplimiento global de los criterios básicos.

4.3. Sello de Calidad 2

Además de lo mencionado anteriormente, un conjunto de datos obtendrá el Sello de Calidad 2 si cumple con los requisitos del Sello de Calidad 1 y, adicionalmente, presenta un valor superior en el incentivo de uso, junto con las siguientes consideraciones. En ese caso, el sello asignado será el nivel 2, como se explica a continuación.

- **Formatos Estándar (Conformidad):** Los datos deben alinearse con formatos y estándares reconocidos, facilitando la interoperabilidad con otras herramientas y sistemas. Se exige una calificación 9 en conformidad, lo que puede lograrse adoptando estándares internacionales (por ejemplo, esquemas JSON/CSV/RDF bien estructurados) y validando la calidad técnica mediante scripts o servicios externos.
- **Incentivo de Uso:** Un modelo de aprendizaje automático que analice la interacción de los usuarios (vistas, descargas, comentarios) apoyándose en pesos calculados por un modelo y la valoración otorgada por expertos. Para que un conjunto de datos sea considerado valioso por la comunidad, debe obtener una calificación mínima de 7. Además, proporcionar ejemplos prácticos, manuales de uso, casos de estudio y notificaciones sobre actualizaciones incrementa significativamente el interés de los usuarios.
- **Calificación Final:** La calificación final promedio debe ser de al menos 8, reflejando la existencia de procesos sistemáticos de mejora. Esto incluye monitoreo regular, encuestas de satisfacción, análisis de brechas y planes de acción concretos para elevar la calidad del conjunto de datos a lo largo del tiempo.

4.4. Sello de Calidad 3

Finalmente, el Sello de calidad es la máxima distinción. Para ello, será para aquellos conjuntos de datos que garanticen haber cumplido previamente con todos los criterios de los dos primeros niveles y adicionalmente, tener lo siguiente.

- **Documentación completa (URL Válida):** Se debe proporcionar un enlace a documentación detallada, ofreciendo descripciones técnicas, glosarios, pautas de actualización, métodos de recolección, fuentes primarias y casos de uso especializados. Esto impulsa la transparencia y la comprensión a profundidad del conjunto de datos.
- **Incentivo de uso avanzado:** Se incrementa el requisito a una calificación mínima de 9, lo cual indica un reconocimiento amplio y sostenido por la comunidad de usuarios. Para lograrlo, es aconsejable fomentar el intercambio de experiencias, crear foros de discusión, participar en eventos o conferencias, y mantener una comunicación activa con la comunidad.
- **Mejora continua avanzada:** La calificación final ponderada debe alcanzar al menos 9. Esto supone procesos de perfeccionamiento muy maduros, con auditorías periódicas, retroalimentación de expertos, adopción dinámica de nuevas normativas, estándares internacionales y actualización proactiva ante cambios tecnológicos o regulatorios.

5. Potencial de Uso

El 'potencial de uso' de un conjunto de datos se refiere a qué tan probable es que la información disponible sea utilizada de manera efectiva por los usuarios. Esta medida no solo considera la cantidad de veces que los datos son vistos o descargados, sino también cómo esas interacciones reflejan su relevancia, utilidad e impacto. Sin embargo, este criterio no aplica para los conjuntos de datos obligatorios, ya que su publicación responde a requerimientos normativos y no a su potencial uso. En un mundo donde los datos son el nuevo recurso clave para tomar decisiones, identificar el potencial de uso en los conjuntos de datos no obligatorios ayuda a priorizar la calidad y optimizar la información disponible para la comunidad.

Para evaluar este potencial, se utiliza una métrica conocida como **CTR (Click Through Rate)**²², que conecta dos indicadores esenciales:

- **Número de vistas:** la cantidad de veces que un conjunto de datos es consultado.
- **Número de descargas:** la frecuencia con la que los usuarios deciden descargar esa información para su uso.

Este enfoque permite identificar qué tan atractiva y valiosa resulta la información publicada, ya sea para investigaciones académicas, análisis empresariales o el diseño de políticas públicas. Así, el "potencial de uso" se convierte en una guía fundamental para mejorar la calidad y relevancia de los datos abiertos.

5.1. Modelo de Predicción del Potencial de Uso

El modelo de predicción del potencial de uso utiliza como principal métrica el **CTR (Click Through Rate)**, que mide la relación entre el número de descargas y el número de vistas de un conjunto de datos. Este indicador permite identificar qué tan relevante o útil resulta un conjunto de datos para los usuarios del portal.

¿Cómo se calcula el CTR?

La fórmula utilizada es la siguiente:

$$\text{CTR} = \text{Número de Descargas (D)} / \text{Número de Vistas (V)}$$

Donde:

D: Número de descargas del conjunto de datos.

¹³ <https://www.appsflyer.com/es/glossary/ctr/>

V: Número de vistas del conjunto de datos.

5.1.1. Ejemplos de Cálculo del CTR

Ejemplo 1:

Un conjunto de datos titulado "*Indicadores de Salud 2023*" recibe 2,000 vistas y 400 descargas. Aplicando la fórmula:

$$\text{CTR} = 400 / 2000 = 0.2$$

Esto significa que el 20 % de las personas que visualizaron el conjunto también lo descargaron.

Ejemplo 2:

Otro conjunto de datos recibe 200 vistas y 300 descargas. Al aplicar la fórmula:

$$\text{CTR} = 300 / 200 = 1.5$$

En este caso, el CTR supera el valor máximo lógico de 1, lo cual no tiene sentido práctico. Para solucionar esto, el modelo ajusta el CTR a un valor máximo de 1 (o 10 en la escala final), garantizando la coherencia de los resultados.

5.1.2. Interpretación del CTR

El CTR refleja la proporción de descargas respecto a las vistas, independientemente de la popularidad del conjunto. Un CTR alto sugiere un mayor impacto y utilidad percibida. Por ejemplo:

- Un conjunto con 2,000 vistas y 400 descargas (CTR = 20 %) podría considerarse exitoso si está dirigido a un público reducido pero interesado.
- Un conjunto con 10,000 vistas y 500 descargas (CTR = 5 %) podría indicar problemas de relevancia o calidad, a pesar de su alta visibilidad.

5.1.3. Aplicación del CTR en la Evaluación

El CTR calculado es un insumo clave para el modelo de predicción del potencial de uso. Este valor se escala a una calificación de 0 a 10:

- Si el CTR excede 1 (más descargas que vistas), se ajusta a 1 para evitar incoherencias.
- La calificación potencial se calcula como:

$$\text{Calificación Potencial de Uso} = \text{CTR} \times 10$$

Por ejemplo, un conjunto con un CTR de 0.8 recibirá una calificación de 8, mientras que uno con un CTR cercano a 1 obtendrá la máxima calificación de 10.

5.2. Implementación del Modelo de Predicción

Para estimar el potencial de uso de los conjuntos de datos, se implementó un modelo basado en el **algoritmo XGBoost²³**, conocido por su eficiencia y capacidad de manejo de datos complejos.

Proceso de Entrenamiento

- **Datos de entrenamiento:**
Se utilizaron registros históricos de conjuntos de datos publicados hasta el 1 de abril de 2023.
- **Variables de entrada:**
Se emplearon 17 criterios de calidad como predictores, excluyendo información relacionada con la entidad o sector para evitar sesgos.
- **División de datos:**
Conjunto de entrenamiento: 5,735 registros.
Conjunto de prueba: 1,434 registros.

5.2.1. Optimización y Evaluación del Modelo

²⁴ <https://www.themachinelearners.com/xgboost-python/>

El modelo fue optimizado para minimizar el **error cuadrático medio (RMSE)**²⁴, logrando un resultado final de **0.2050** en el conjunto de prueba. Además, se utilizó la **técnica SHAP (SHapley Additive exPlanations)**²⁵ para interpretar los resultados y determinar la importancia de cada criterio en las predicciones.

Criterios con mayor impacto:

- Consistencia de los datos.
- Exactitud sintáctica y semántica.
- Credibilidad de la información.

Criterios con menor impacto:

- Disponibilidad y eficiencia, debido a su uniformidad en la mayoría de los registros.

Aplicación Temporal del Modelo

El modelo de predicción del potencial de uso se aplica exclusivamente a conjuntos de datos publicados en el último año. Para conjuntos más antiguos, se utiliza directamente el valor real del CTR para evaluar su calidad

¹⁵ <https://arize.com/blog-course/root-mean-square-error-rmse-what-you-need-to-know/>

¹⁶ <https://www.sciencedirect.com/topics/computer-science/shapley-additive-explanation>



6. Principios de la calidad para la publicación de datos abiertos

La calidad de los datos abiertos está definida por estándares internacionales que garantizan su utilidad, accesibilidad y transparencia. Para que los datos sean considerados verdaderamente abiertos, deben cumplir con los siguientes principios:

- **Completitud:** Toda la información pública debe estar disponible de manera íntegra, excepto aquella protegida por razones de privacidad, seguridad o privilegios legales, promoviendo un acceso amplio y sin restricciones indebidas.
- **Origen confiable y primario:** Los datos deben ser recolectados directamente de su fuente original, manteniendo el máximo nivel de detalle y evitando modificaciones o agregaciones que puedan alterar su valor original.
- **Actualización oportuna:** Los datos deben ser publicados en tiempos adecuados para garantizar su relevancia y utilidad, permitiendo la toma de decisiones basadas en información actualizada.
- **Acceso universal e inclusivo:** La información debe estar disponible para el mayor número de usuarios, sin barreras técnicas o administrativas, asegurando que puedan ser utilizados para diversos fines y por diferentes sectores.
- **Estructura procesable por máquinas:** Los datos deben estar organizados de manera que faciliten su análisis y procesamiento automático, optimizando su uso en aplicaciones tecnológicas y herramientas de análisis.
- **Acceso equitativo y sin discriminación:** Los datos deben ser accesibles para cualquier persona interesada, sin requisitos de registro, permisos o barreras que limiten su acceso.
- **Formatos abiertos y no propietarios:** La información debe publicarse en formatos libres, garantizando que ningún actor tenga control exclusivo sobre ellos, fomentando la reutilización y la interoperabilidad.
- **Exención de restricciones legales:** Los datos deben estar libres de derechos de autor, patentes, marcas registradas o acuerdos de confidencialidad, salvo excepciones justificadas en privacidad, seguridad o privilegios específicos.

6.1. Atributos de Calidad

El perfilamiento y reporte de atributos de calidad es una herramienta estratégica clave para las entidades públicas, orientada a monitorear y mejorar continuamente la gestión de la información. Este proceso permite identificar el avance en la calidad de los datos mediante la contabilización detallada de las instancias de atributos encontrados.

El perfilado de datos tiene como propósito evaluar y garantizar la precisión, completitud, consistencia y relevancia de la información gestionada. Esto contribuye a tomar decisiones basadas en datos confiables y alineadas con los estándares de gobernanza de la información.

Para realizar el perfilamiento de datos, se emplean técnicas y herramientas avanzadas que permiten:

- Analizar la estructura y contenido de los datos.
- Identificar inconsistencias, duplicados o valores fuera de rango.
- Generar indicadores de calidad de los atributos evaluados.
- Establecer lineamientos para mejorar la integridad y precisión de la información.

6.2. El perfilado de datos

El perfilado de datos es un proceso esencial para evaluar y mejorar la calidad de la información. Este análisis incluye la identificación y medición de elementos críticos como:

- **Datos nulos:** Detección de campos incompletos que afectan la integridad de la información.
- **Registros duplicados:** Identificación de duplicidades que impactan la confiabilidad de los datos.
- **Campos en desuso:** Determinación de datos obsoletos o irrelevantes para optimizar la gestión.
- **Máscaras y validaciones de formato:** Verificación del cumplimiento de estándares en los formatos de datos, asegurando consistencia y precisión.

En este proceso, también se detectan errores comunes como:

- **Codificaciones incorrectas:** Por ejemplo, inconsistencias en nombres de departamentos en registros del Estado colombiano.
- **Tipologías no previstas:** Como variaciones en categorías predefinidas, por ejemplo, más de dos opciones para el campo sexo (hombre, mujer, masculino, femenino, etc.).

Para garantizar un análisis eficiente y confiable, es fundamental apoyarse en herramientas tecnológicas especializadas. Soluciones de software libre como **Open Refine** ofrecen capacidades

avanzadas para realizar estas tareas de manera eficiente, permitiendo a las entidades públicas optimizar sus procesos de depuración y validación de datos.

6.3. El perfilado de la información

El perfilado de información eleva el análisis a un nivel superior, enfocándose no solo en la calidad técnica de los datos, sino también en su interpretación y contexto.

Por ejemplo, mientras que en el perfilado de datos se verifica que una fecha sea técnicamente válida, en el perfilado de información se analiza el significado de esa fecha. Podría interpretarse como una fecha de nacimiento, permitiendo realizar análisis más profundos, como la segmentación por grupos etarios o el cálculo de antigüedad en un registro.

6.4. El perfilado de Función Pública

El perfilado de la Función Pública permite implementar análisis complejos y específicos que optimizan la toma de decisiones y mejoran los procesos en el sector público.

Por ejemplo:

- **Análisis de campañas finalizadas:** Evaluación de resultados en iniciativas públicas, proporcionando información clave para stakeholders y facilitando la planificación de futuras acciones.
- **Validación con reglas cruzadas:** Aplicación de lógicas avanzadas para garantizar precisión en cálculos críticos, como la tarificación y facturación de servicios públicos.

Este perfilado no solo se fundamenta en los datos, sino también en sus metadatos funcionales, incorporando:

- **Interpretaciones específicas del sector público:** Contextualización de la información según su aplicabilidad en procesos administrativos y operativos.
- **Lógicas de validación avanzadas:** Definición y aplicación de reglas claras para verificar la calidad de los datos.
- **Umbrales de calidad:** Establecimiento de estándares mínimos que aseguren la confiabilidad de la información.

6.5. Informe final

El informe final muestra un análisis detallado que identifica iniciativas estratégicas para garantizar la calidad de los datos en todas las etapas del proceso. Esto incluye:

- **Limpieza preventiva de datos:** Recomendaciones específicas para optimizar la calidad de los datos en los sistemas fuente, evitando la propagación de errores.
- **Detección de inconsistencias:** Identificación de posibles inconsistencias en los repositorios existentes, asegurando la integridad y fiabilidad de la información.
- **Diagnóstico de fallas de calidad:** Registro detallado de los problemas detectados durante el proceso de perfilado de datos, con un enfoque en su resolución efectiva.

Estas acciones no solo buscan mitigar riesgos, sino también proporcionar insumos valiosos para ser aplicados en la fase de diseño de la migración de datos, garantizando que el proceso de transferencia de información sea eficiente, confiable y alineado con los estándares de calidad requeridos por el sector público.

6.6. Procesos de calidad

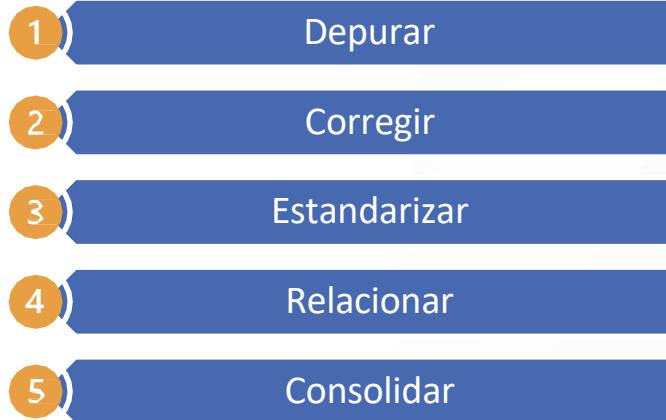
La calidad de los datos está intrínsecamente ligada a su limpieza y al cumplimiento de objetivos orientados a prevenir fallas en su captura. Estos procesos no solo garantizan la integridad de la información, sino que también permiten alinear los datos con las metas y necesidades específicas de la organización.

Los procesos de limpieza están determinados por las características y reglas de negocio de cada entidad pública, abordando preguntas clave como:

- ¿Cuál es el propósito de capturar estos datos?
- ¿Qué tipo de datos son necesarios para cumplir con los objetivos organizacionales?
- ¿Cómo debe estar estructurada la información para ser eficiente y procesable?

Dependiendo de las respuestas, se definirán procesos específicos de limpieza de datos o, en casos excepcionales, no será necesaria ninguna intervención.

Ilustración 30 Actividades limpieza de datos



Fuente. Elaboración propia

6.7. Planificación de la calidad

La planificación de la calidad es un proceso clave que permite a las entidades públicas establecer una estrategia robusta para gestionar eficazmente la calidad de los metadatos y datos. Este proceso define roles, responsabilidades, programas de calidad, secuencias de ejecución, interacción de procesos y el uso de herramientas especializadas.

- **Planeación:** En esta etapa se realiza un diagnóstico integral y se diseña el plan estratégico para garantizar la calidad de los metadatos y datos:
 - A. **Diagnóstico de calidad:** Identificación del estado actual de los metadatos y datos, detectando brechas y oportunidades de mejora.
 - B. **Definición del plan estratégico:** Desarrollo de un plan alineado con los objetivos institucionales, estableciendo metas claras y acciones específicas para mejorar la calidad.
- **Desarrollo:** Durante esta fase se materializan las acciones planificadas, ajustando y probando los procesos de gestión de calidad:
 - A. **Estimación de ajustes y desarrollos:** Evaluación de las necesidades de ajustes o nuevos desarrollos relacionados con la calidad de los datos.

- B. **Pruebas unitarias y funcionales:** Validación técnica y operativa de los procesos y herramientas implementadas para Data Quality, asegurando su correcto funcionamiento.
- **Ejecución:** La ejecución pone en marcha las acciones planificadas y garantiza su sostenibilidad a través de monitoreo y comunicación continua:
 - A. **Automatización de procesos:** Implementación de programas automáticos de control de calidad con frecuencia mensual o quincenal.
 - B. **Comunicación institucional:** Generación de reportes y actualización periódica de las entidades mediante correos electrónicos.
 - C. **Monitoreo permanente:** Seguimiento activo a través del canal oficial de datos abiertos: datosabiertos@mintic.gov.co.

6.8. Validación de la calidad

La validación de la calidad de datos en el marco de la apertura de información pública es una actividad clave para garantizar la confiabilidad, accesibilidad y utilidad de los datos abiertos. Este proceso, realizado de manera periódica, implica identificar manualmente características críticas de calidad en los conjuntos de datos publicados por las entidades.

- **Validaciones Semanales:** Las revisiones semanales permiten realizar ajustes rápidos y garantizar que los datos cumplan con los estándares básicos de calidad:
 - A. **Volumen insuficiente:** Conjuntos de datos con menos de 50 registros publicados.
 - B. **Enlaces no funcionales:** Direcciones que no permiten la descarga directa de los datos en formatos válidos como CSV, XLS, XLSX, JSON, KML, KMZ y ZIP (shapefiles de ESRI).
 - C. **Errores en la metadata:** Enlaces en los campos de URL de documentación y/o normativa que redirigen a páginas no disponibles.
 - D. **Metadata incompleta o ausente:** Campos clave vacíos o insuficientes que afectan la interpretación y el uso de los datos.
 - E. **Estructura deficiente:** Conjuntos de datos con menos de tres columnas o en formatos no admitidos como datos abiertos.

- **Validaciones Quincenales:** Las revisiones quincenales se enfocan en detectar y corregir problemas más complejos, alineando los datos con las mejores prácticas de calidad:
 - A. **Ajustes automáticos:** Corrección de nombres de municipios, departamentos, entidades y formatos de fechas para garantizar consistencia.
 - B. **Desactualización:** Conjuntos de datos que no están actualizados según la frecuencia definida en su metadata.
 - C. **Enlaces rotos:** URLs externas que redireccionan a páginas no disponibles.
 - D. **Identidad institucional:** Modificación del nombre del usuario en los datos abiertos por el nombre oficial de la entidad.
 - E. **Estructura de las columnas:** Conjuntos que contienen únicamente columnas de texto o que carecen de filas con información válida.

6.9. Control y aseguramiento de la calidad de los datos

La gestión de la calidad de los datos en las entidades públicas puede resumirse en seis pasos fundamentales, que integran acciones clave orientadas a garantizar información confiable y alineada con los objetivos institucionales:

1. **Descubrimiento de datos:** Este paso inicial consiste en la búsqueda, recopilación, organización y documentación de metadatos relevantes. Este proceso permite construir un mapa completo de los activos informacionales, facilitando su acceso y comprensión.
2. **Perfilado de datos:** Implica un análisis exhaustivo de los datos para compararlos con sus metadatos, calcular estadísticas clave y definir medidas de calidad aplicables en cada etapa. Este proceso asegura que los datos cumplan con los estándares establecidos y que sean adecuados para su propósito.
3. **Reglas de calidad de datos:** Estas reglas son el marco para optimizar la calidad de los activos informacionales. Se diseñan a partir de los requisitos de negocio, las reglas comerciales y las especificaciones técnicas aplicables, garantizando que los datos estén alineados con los objetivos operativos y estratégicos de la entidad.
4. **Monitorización de la calidad de los datos:** La mejora continua requiere un monitoreo constante. Este paso permite comparar los resultados alcanzados con los umbrales de calidad definidos, registrar excepciones y generar notificaciones que impulsan acciones correctivas. Es un componente esencial para consolidar una cultura

de calidad en la gestión de datos.

5. **Informes de calidad de datos:** El “reporting” es clave para mantener la transparencia en la gestión de calidad. Incluye la generación de informes detallados, el registro de excepciones y la actualización constante de las medidas correctivas implementadas.
6. **Limpieza de datos:** La limpieza es una actividad continua que aborda las excepciones y problemas de calidad detectados. Este paso asegura la corrección en tiempo real de errores, mejorando la precisión y confiabilidad de los datos.

7. Calidad de los metadatos

7.1. ¿Qué son los metadatos?

Los metadatos son elementos esenciales que describen y contextualizan otros datos, proporcionando información clave que facilita el acceso, la interpretación y la reutilización de los recursos informativos. Una adecuada gestión de metadatos no solo mejora la calidad de los datos, sino que también genera confianza al permitir identificar información deficiente, incorrecta o incompleta de manera oportuna.

La calidad de los datos está directamente relacionada con la calidad de sus metadatos. Contar con metadatos confiables permite:

- **Mejorar el acceso:** Facilitan la localización y comprensión de los conjuntos de datos.
- **Optimizar la interoperabilidad:** Promueven el intercambio eficiente de información entre diferentes plataformas y sectores.
- **Identificar problemas:** Ayudan a detectar errores y vacíos en la información publicada.
- **Estándar DCAT17:** Un Enfoque para la Interoperabilidad

El portal de Datos Abiertos de Colombia ha adoptado el estándar **DCAT17** como marco para la definición de metadatos en los conjuntos de datos publicados. Este estándar es fundamental para:

- **Asegurar la interoperabilidad:** Permitir que conjuntos de datos albergados en diferentes portales puedan ser integrados y utilizados de manera conjunta.
- **Homogeneizar la información:** Garantizar un enfoque estructurado y consistente en la publicación de datos abiertos.

Ilustración 31 Metadatos del conjunto de datos

Editar los metadatos

X

Título y descripción
Escriba su título y descripción lo más claro y simple como le sea posible.

* CAMPO OBLIGATORIO

Etiqueta de la fila
Describa lo que representa cada fila en el conjunto de datos (si procede).

Categorías y etiquetas
Clasifique el conjunto de datos para que le sea más fácil encontrarlo.

Licencia de Atribución

Email de contacto

Nombre del recurso

Anexos
Si desea dotar de contenido extra a este conjunto de datos, añada un anexo.

Cancelar
Guardar

Fuente. Tomado de Portal Datos abiertos. Herramientas.

Es fundamental garantizar que los datos publicados estén protegidos mediante licencias estándar que regulen su uso y promuevan el intercambio responsable de información. Estas licencias no solo salvaguardan la estructura y el contenido de los conjuntos de datos, sino que también aseguran el reconocimiento de los derechos de los creadores cuando los datos son reutilizados.

Se recomienda utilizar licencias abiertas estándar, como **Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)**, tal como se establece en los términos y condiciones de uso del portal de Datos Abiertos de Colombia (www.datos.gov.co). Esta licencia garantiza:

- **Protección de derechos:** Los creadores mantienen la titularidad de los derechos de los datos publicados.
- **Fomento del reuso responsable:** Los usuarios deben atribuir el origen de los datos y compartir cualquier modificación bajo la misma licencia, promoviendo la transparencia y la

interoperabilidad.

Para facilitar el proceso de publicación de datos abiertos, el portal de Datos Abiertos de Colombia ofrece recursos prácticos en la sección de herramientas, donde se incluyen:

- **Guías sobre diligenciamiento de metadatos:** Información detallada para garantizar que los datos cumplan con los estándares de calidad.
- **Videos tutoriales:** Material audiovisual que explica cómo publicar conjuntos de datos de manera adecuada y conforme a las mejores prácticas.
- **Capacitación en línea:** Cursos y talleres virtuales dirigidos a entidades públicas y usuarios interesados en aprender a gestionar y publicar datos abiertos de manera efectiva.
- **Acceso a APIs:** Documentación y ejemplos para facilitar la integración y el consumo automatizado de datos por parte de desarrolladores.

7.2. Como Diligenciar los Metadatos

7.2.1. Título y Descripción

- **Título:** Proporcione un título conciso y descriptivo que refleje el contenido principal del conjunto de datos. Evite incluir años o períodos de tiempo en el título para asegurar su relevancia a lo largo del tiempo. Ejemplo: "Estadísticas de Hospitales en Colombia".
- **Descripción:** Incluya un resumen claro y específico sobre qué trata el conjunto de datos. La descripción debe contener, en lo posible, más de 150 caracteres para garantizar un contexto completo. Ejemplo: "Incluye información sobre el número de hospitales, camas disponibles y ubicaciones geográficas".

7.2.2. Etiqueta de la Fila

Especifique el identificador principal para cada fila en el conjunto de datos. Este campo ayuda a identificar únicamente cada entrada del archivo. Ejemplo: "ID del hospital, código de institución".

7.2.3. Categorías y Etiquetas

- Seleccione una categoría de la lista desplegable que mejor represente el conjunto de datos.
- Agregue etiquetas o palabras clave que ayuden a los usuarios a localizar el conjunto de datos. Ejemplo: "Salud", "Hospitales", "Estadísticas".

7.2.4. Licencia de Atribución

Seleccione la licencia que regula el uso del conjunto de datos. Especifique si se requiere atribuir crédito al publicador o si el uso es completamente libre. Proporcione un enlace a la descripción de la licencia si es necesario.

7.2.5. Información de Contacto

Proporcione una dirección de correo electrónico institucional válida para que los usuarios puedan hacer consultas o reportar problemas con el conjunto de datos, asegurando mayor veracidad.

7.2.6. Nombre del Recurso

Escriba un nombre que identifique el archivo o recurso que será compartido dentro del conjunto de datos. Ejemplo: "Hospitales_Colombia.csv".

7.2.7. Anexos

Adjunte documentos adicionales que complementen el conjunto de datos, como manuales, guías o documentos de referencia. Utilice el botón "adjuntar" para cargar los archivos.

7.2.8. Información de la Entidad

- **Departamento:** Seleccione el departamento administrativo al que pertenece la entidad responsable del conjunto de datos.

- **Entidad Principal:** Ingrese el nombre de la institución o entidad responsable de la generación y publicación de los datos. Asegúrese de que el nombre sea preciso y no contenga errores tipográficos.
- **Área o Dependencia:** Especifique el área interna o dependencia dentro de la entidad principal que gestionó el conjunto de datos.

7.2.9. Información de Datos

- **Idioma:** Seleccione el idioma principal del conjunto de datos. Ejemplo: "español".
- **Cobertura Geográfica:** Indique la región o área geográfica cubierta por los datos. Ejemplo: "Internacional", "Nacional".
- **Frecuencia de Actualización:** Especifique con qué periodicidad se actualizará el conjunto de datos. Ejemplo: "Mensual", "Anual".

7.2.10. URL de Documentación y Normativa

- Proporcione enlaces a documentos que expliquen en detalle el contenido y el formato del conjunto de datos.
- Incluya también enlaces a normativas o políticas relacionadas con los datos publicados.

8. Clasificación y priorización de errores de calidad y errores de publicación

La siguiente tabla presenta información clave sobre los tipos de errores que se evalúan actualmente en los procesos de validación de conjuntos de datos. Este análisis permite a las entidades identificar las causas específicas detrás de rechazos, proporcionando información detallada que facilita la toma de decisiones estratégicas para optimizar los procesos de validación. Además, ofrece una guía clara para las entidades publicadoras al señalar qué aspectos deben corregirse, promoviendo la mejora continua y asegurando que las futuras cargas de datos cumplan con los estándares de calidad requeridos. Esta herramienta es fundamental para garantizar la precisión y confiabilidad de los datos, fortaleciendo la gestión pública y la transparencia en el uso de la información.

Tabla 1. 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 períodos y/o algún tipo de clasificación.
Compleitud	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.
Comprendibilidad	Comprendibilidad 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 comprendibilidad 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 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 que puede pertenecer a la plataforma, departamento, municipio, ciudad, región, área o dependencia.	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

8.1. Errores de publicación

Tabla 2. 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 campos faltantes.
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 vacias	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	El error "Conjunto de datos presenta columnas mal nombradas: Unnamed Column" ocurre cuando el conjunto de datos contiene columnas sin nombre, usualmente generadas por celdas vacías en los encabezados o problemas en la estructura del archivo
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), no es un dato abierto (está en un formato cerrado como PDF, .DOC, .PPT)
ERR011	Completitud del conjunto de datos	Conjunto de datos clasificado por periodos/por tipologías.	

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.
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.
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 datos 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 está 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, xml; y zip (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.	<p>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:</p> <ul style="list-style-type: none"> 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.
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 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.

Fuente. Elaboración propia

Como se observa en el siguiente ejemplo, el error ERR005_01 representa una combinación específica de los errores ERR005 y ERR018, permitiendo una evaluación más detallada y precisa de la calidad de los conjuntos de datos:

- **ERR005:** El conjunto de datos no cumple con las características de una base de datos o presenta una cantidad limitada de información, lo que restringe su potencial de reutilización.
- **ERR005_01:** Este error combina las características del ERR005 y añade que el conjunto de datos contiene agregaciones o totales, afectando aún más su utilidad para análisis específicos.
- **ERR018:** Indica que el conjunto de datos incluye exclusivamente agregaciones o totales, limitando su desagregación y profundidad analítica.

Gracias a la depuración y especificidad de los errores identificados, se asegura que los conjuntos de datos restantes cumplen con los criterios necesarios para una validación rigurosa de calidad, basada en los estándares definidos por la **norma ISO 25012**. Este enfoque sistemático refuerza la confiabilidad de los datos publicados, alinea las prácticas con estándares internacionales y fortalece la gestión de datos abiertos en las entidades públicas nacionales.

8.2. Clasificación de Datos

Tabla 3. Análisis de errores

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

Fuente. Elaboración propia

Como solución, se recomienda implementar un enfoque estructurado que utilice columnas específicas para clasificar los datos por períodos (como años, meses o días) y por temáticas o categorías relevantes. Además, se sugiere unificar los conjuntos de datos relacionados para mejorar su coherencia, accesibilidad y utilidad, garantizando que cumplan con los estándares de calidad y sean fácilmente reutilizables por las entidades públicas y otros usuarios.

8.3. Completitud

Tabla 4. Análisis errores completitud

Código del error	Categoría de error	Tipo de error	Descripción del 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 dé valor o aporte al contexto del conjunto de datos, se recomienda eliminarla.

Fuente. Elaboración propia

El cuadro de completitud refleja la importancia de garantizar que los conjuntos de datos publicados sean completos y consistentes. La falta de información en campos o columnas, como valores nulos o vacíos, afecta directamente su utilidad, interpretación y confianza, lo que limita su reutilización y su valor en contextos específicos. Para abordar este problema, se propone completar los campos faltantes con datos relevantes o eliminar aquellos que no aporten valor al conjunto de datos. Además, si la sumatoria de las columnas nulas o vacías representa más del 5 % del total de columnas, se recomienda rechazar el conjunto, ya que esto indica una deficiencia significativa en la calidad de los datos. Estas medidas no solo mejoran la calidad técnica de la información, sino que también fortalecen su utilidad para la toma de decisiones, incrementan su valor para análisis y políticas públicas, y aseguran el cumplimiento de estándares internacionales como la norma ISO 25012. Al implementar estas acciones, las entidades públicas refuerzan la transparencia, confianza y efectividad en la gestión de los datos abiertos.

8.4. Comprensibilidad

Tabla 5. Análisis errores comprensibilidad

Código del error	Categoría de error	Tipo de error	Descripción del 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

El cuadro de comprensibilidad resalta la relevancia de garantizar que los datos publicados sean claros, interpretables y accesibles para los usuarios. Un conjunto de datos que no puede ser interpretado adecuadamente limita su utilidad, dificulta su análisis y puede generar confusión o malentendidos en su uso. Para abordar este problema, se recomienda mejorar los metadatos, añadiendo información adicional que permita clasificar los registros de manera más precisa. Además, los datos deben publicarse en formatos que faciliten el procesamiento automático y con el más alto nivel de detalle posible, evitando agregaciones o modificaciones que dificulten su comprensión. También es importante proporcionar un contexto adecuado a los datos, de modo que tengan un sentido claro y estén alineados con un propósito específico. Estas acciones no solo incrementan la claridad y utilidad de los datos, sino que también fortalecen su capacidad para ser reutilizados en la toma de decisiones y el desarrollo de políticas públicas. La implementación de estas mejoras asegura que las entidades públicas cumplan con estándares internacionales de calidad y refuerzen la confianza en la publicación de datos abiertos.

8.5. Confidencialidad

Tabla 6. Análisis errores confidencialidad

Código del error	Categoría de error	Tipo de error	Descripción del 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.	Anonimización de la información. Se recomienda no publicar.

Fuente. Elaboración propia

El error de confidencialidad destaca la importancia de proteger los datos sensibles, asegurando que solo usuarios autorizados puedan acceder a ellos, en cumplimiento con estándares internacionales como la norma **ISO/IEC 13335-1:2004**. Este error, relacionado con el tratamiento de datos personales, implica riesgos legales, éticos y reputacionales si no se implementan medidas adecuadas. Para solucionarlo, se recomienda la anonimización de la información, eliminando cualquier dato que permita identificar personas o entidades, y evitar la publicación de información que pueda generar riesgos de reidentificación. Abordar este problema es esencial para cumplir con las normativas de privacidad, proteger la confianza ciudadana, reducir riesgos legales y fortalecer la seguridad en la gestión pública de datos abiertos.

8.6. Conformidad

Tabla 7. Análisis errores conformidad

Código del error	Categoría de error	Tipo de error	Descripción del 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

El error de conformidad se enfoca en garantizar que los valores de los datos cumplan con formatos estándares y legibles para facilitar su correcta interpretación y utilización. Este problema ocurre cuando los datos presentes en los campos de una tabla no tienen un formato adecuado, lo que limita su utilidad y dificulta su integración en procesos de análisis o sistemas automatizados. Para solucionar este error, se recomienda ajustar y normalizar los formatos de las columnas del dataset, asegurando que los valores sean coherentes, estandarizados y alineados con las necesidades del usuario final. Corregir este tipo de error es fundamental para incrementar la calidad y la interoperabilidad de los datos, fortaleciendo su utilidad en la toma de decisiones y promoviendo una gestión pública más eficiente y profesional.

8.7. Duplicidad

Tabla 8. Análisis errores duplicidad

Código del error	Categoría de error	Tipo de error	Descripción del 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 al 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

El error de duplicidad destaca la necesidad de evitar la publicación repetida de información en el portal de Datos Abiertos por diferentes entidades, especialmente cuando los datos corresponden al mismo departamento, municipio, ciudad, región, área o dependencia. Este problema genera redundancia, confusión y puede afectar la confianza en la calidad de los datos publicados. La solución propuesta consiste en contactar a la entidad que ya ha publicado la misma información, consolidar los datos en una única fuente oficial y establecer claramente el rol y la pertinencia de la publicación. Abordar este tipo de error es esencial para optimizar la organización de los datos abiertos, mejorar la experiencia del usuario y garantizar una gestión pública eficiente, coherente y alineada con los estándares de calidad.

8.8. Metadata errada, incompleta y/o vacía

Tabla 9. Metadata errada, incompleta y / o vacía

Código del error	Categoría de error	Tipo de error	Descripción del 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

El cuadro de errores de metadata errada, incompleta y/o vacía resalta problemas críticos que afectan la calidad y la claridad de los conjuntos de datos publicados. Estos errores incluyen títulos y descripciones mal nombrados, metadatos incompletos o vacíos, y campos de usuario que no reflejan la entidad publicadora. Estos problemas generan confusión entre los usuarios y disminuyen la confianza en los datos publicados. Para solucionarlos, se recomienda actualizar los títulos y descripciones siguiendo estándares de calidad, completar los metadatos con la información requerida y vincular los campos de usuario al nombre de la entidad publicadora, asegurando un reconocimiento adecuado. Abordar estos errores mejora la claridad, consistencia y profesionalismo de los datos abiertos, fortaleciendo su utilidad y la confianza de los usuarios.

8.9. Error sin filas

Tabla 10. Error sin filas

Código del error	Categoría de error	Tipo de error	Descripción del 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. 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

El error ERR004 se centra en conjuntos de datos que no contienen filas con información, lo que limita significativamente su utilidad y el aprovechamiento de la información por parte de los ciudadanos. Este problema afecta la transparencia y el valor de los datos abiertos, ya que un conjunto de datos vacío no genera beneficios ni cumple con los objetivos de accesibilidad y reutilización. La solución propuesta es eliminar estos conjuntos vacíos y crear nuevos que incluyan columnas y filas relevantes, complementando lo publicado o disponiendo la información directamente en el sitio web de la entidad, especialmente si es de interés limitado. Abordar este error es fundamental para garantizar que los datos abiertos sean valiosos, accesibles y alineados con las necesidades de los usuarios, fortaleciendo la confianza y la efectividad de la gestión pública.

8.10. Error pocas filas

Tabla 11. Error pocas filas

Código del error	Categoría de error	Tipo de error	Descripción del 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 permitirá la reutilización por parte de los ciudadanos para realizar un producto o servicio.	Se recomienda para que se disponga una base de datos consolidada integrada que relacione los datos relacionados con turismo en un solo, denominado 'Oferta Turística del Municipio de xxx', de esta manera se consolidarían hoteles, comercios, 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, adicionando información más amplia y brindando mayor información al ciudadano como: datos de geolocalización, tipo de establecimiento, ciudad, departamento, sitio web, correo, horario de atención.
ERR005_02	Error pocas filas y clasificado por periodos	Conjunto de datos, no es una base de datos y presenta poca información para reutilización y adicionalmente está publicada por periodos (año, mes).	El conjunto de datos tiene muy pocos registros (menos de 50), lo que no permitirá 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 para que sea de utilidad para los ciudadanos consolidando la información en una sola publicación y no dividida por años o períodos (año, mes) con el fin de facilitar el análisis y generar una visualización más adecuada en el portal y en dispositivos móviles.
ERR007	Error vacías filas	Conjunto de datos con campos vacíos y/o basura.	El conjunto de datos presenta campos vacíos en las columnas XX, el campo XX no presenta formato tipo fecha en todos sus campos.	Se recomienda hacer uso de herramientas para la estandarización de los campos XX, el campo XX en todas las filas debería estar completo y en el formato que corresponde para presentarlas.

				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	Error Filas	Poca	Conjunto de datos, no es una base de datos, o presenta información reutilización.	El conjunto de datos tiene muy pocos registros (menos de 50 registros), lo que no permitirá la reutilización por parte de los ciudadanos para realizar un producto o servicio.

				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 genere el reporte y hacer uso de la funcionalidad de vistas filtradas del portal para generar reportes. En caso de que la información solo sea de interés particular de la entidad, se recomienda disponer esta información en la página web.
ERR005_01	Error Filas y Agregado	Poca y	Conjunto de datos no es una base de datos o presenta 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 permitirá la reutilización por parte de los ciudadanos para realizar un producto o servicio. Adicionalmente el conjunto de datos tiene registros agregados y totales.

Fuente. Elaboración propia

Los errores ERR005_02 y ERR007 evidencian desafíos relacionados con la calidad y completitud de los conjuntos de datos. El error ERR005_02, asociado a conjuntos con pocas filas, pone de manifiesto la necesidad de consolidar y estructurar la información de manera integral, especialmente en temáticas como el turismo o datos clasificados por períodos. Para mejorar estos conjuntos, se recomienda crear bases de datos unificadas que permitan a los ciudadanos acceder a información útil, organizada y enriquecida con datos adicionales como geolocalización o características específicas que mejoren la experiencia de consulta.

Por otro lado, el error ERR007 se centra en la presencia de campos vacíos o con datos basura, lo que compromete la utilidad y la interpretación de los conjuntos de datos. La solución consiste en estandarizar los campos y garantizar que estén completos y en formatos adecuados, asegurando su coherencia y su valor para el usuario final.

Aunque el error de pocas filas (ERR005_02) es uno de los más comunes, hay excepciones que justifican la aprobación de conjuntos con menos de 50 registros, siempre y cuando representen el universo completo de los datos. Ejemplos incluyen departamentos de Colombia, municipios específicos como los del Cauca, u otras categorías con un alcance limitado pero completo. En estos casos, la validez del conjunto no está en el número de registros, sino en su representatividad y utilidad.

8.11. Error pocas columnas

Tabla 12. Error pocas columnas

Código del error	Categoría de error	Tipo de error	Descripción del 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. 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	El 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

Los errores ERR008 y ERR008_1 resaltan problemas estructurales en los conjuntos de datos relacionados con la cantidad y calidad de las columnas. El error ERR008 señala conjuntos que tienen una única columna o campo de datos, lo cual limita significativamente su utilidad y aprovechamiento en el portal de datos abiertos. Por su parte, el error ERR008_1 identifica conjuntos con menos de tres columnas, lo que dificulta su reutilización por parte de los ciudadanos para realizar productos o servicios útiles. Para abordar estos problemas, se recomienda eliminar los conjuntos de datos con estructuras deficientes y crear nuevos con más columnas y contenido relevante que complemente lo publicado. Además, cuando la información sea de interés limitado, se sugiere disponerla directamente en el sitio web de la entidad. Estas medidas garantizan que los datos sean más completos, útiles y alineados con las necesidades de los usuarios, fortaleciendo la calidad de los datos abiertos y su impacto en la gestión pública.

8.12. Error pocas columnas mal nombradas

Tabla 13. Error pocas columnas mal nombradas

Código del error	Categoría de error	Tipo de error	Descripción del error	Solución
ERR008_2	Error columnas mal nombradas	Conjunto de datos presenta columnas mal nombradas (Unnamed Column)	El conjunto de datos contiene columnas sin un nombre definido o con nombres genéricos como 'Unnamed Column', lo que dificulta su interpretación y uso. Este problema afecta la claridad de los datos y puede generar confusión en los usuarios.	Se recomienda hacer uso de herramientas de calidad de datos para estandarizar los campos e identificar los errores presentados.

Fuente. Elaboración propia

El error ERR008_2 resalta la problemática de conjuntos de datos con columnas mal nombradas, lo que afecta directamente su claridad y utilidad. Este tipo de error se presenta cuando las columnas tienen títulos genéricos o indeterminados como "Unnamed Column", "column1", "dato1", entre otros, dificultando la identificación de los atributos que contienen. Este problema genera confusión y limita el aprovechamiento de los datos tanto para análisis como para la reutilización por parte de los usuarios.

Para abordar este error, se recomienda ajustar la descripción del problema para que sea más general y comprensible, evitando especificaciones técnicas innecesarias. Por lo tanto, el tipo de error debería enfocarse en señalar que el conjunto de datos presenta columnas mal nombradas. La descripción del error debería enfatizar que las columnas poseen títulos genéricos que no permiten identificar claramente los atributos del conjunto de datos. La solución propuesta es utilizar herramientas de calidad de datos para estandarizar los nombres de las columnas y corregir los errores, asegurando que los títulos sean claros, descriptivos y representen adecuadamente el contenido de cada columna.

8.13. Error falta campo de geolocalización del conjunto de datos

Tabla 14. Error falta de geolocalización del conjunto de datos

Código del error	Categoría de error	Tipo de error	Descripción del 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

El error ERR009 destaca la ausencia de campos de geolocalización (latitud y longitud) en conjuntos de datos que contienen información de dirección sin estandarización. Este problema limita la reutilización del conjunto de datos para generar mapas de ubicación o realizar análisis espaciales, reduciendo su valor para los usuarios. La solución propuesta es incluir campos de geolocalización con coordenadas estándar, garantizando que la información de ubicación sea clara, precisa y utilizable en herramientas que permitan el análisis geográfico. Abordar este error no solo mejora la calidad técnica de los datos, sino que también amplía su potencial de reutilización, optimizando su impacto en la toma de decisiones y en aplicaciones prácticas, como la planificación urbana o la gestión de recursos.



TIC



8.14. Enlace invalido

Tabla 15. Enlace inválido

Código del error	Categoría de error	Tipo de error	Descripción del 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). NO es un dato abierto si está 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. Los datos abiertos deben cumplir con los principios de completitud, ser de fuente primaria y presentar un alto nivel de desagregación.

Fuente. Elaboración propia

El error ERR010 resalta un problema crítico relacionado con enlaces externos que no permiten la descarga directa de datos en formatos abiertos y estándar, como CSV, XLS, XLSX, JSON, KML, KMZ o ZIP (shapefile de ESRI). En lugar de ello, los datos están en formatos cerrados como PDF, DOC o PPT, lo que impide su reutilización y dificulta el cumplimiento de los principios de datos abiertos. La solución propuesta consiste en corregir los enlaces para garantizar la disponibilidad de los datos en formatos estructurados y abiertos, que permitan su descarga y uso directo. Además, se recomienda que estos conjuntos cumplan con los principios de completitud, sean provenientes de fuentes primarias y estén desagregados para maximizar su valor. Abordar este error es esencial para garantizar la interoperabilidad, accesibilidad y utilidad de los datos publicados, fortaleciendo la confianza en la gestión pública y el cumplimiento de los estándares internacionales de datos.

8.15. Conjunto o subconjunto de errores

Tabla 16. Conjunto o subconjunto de errores

Código del error	Categoría de error	Tipo de error	Descripción del error	Solución
ERR011	Compleitud del conjunto de datos	Conjunto de datos clasificado por periodos/tipologías	Actualmente los conjuntos de datos se encuentran divididos por periodos y/o tipologías, lo cual se recomienda unificarlos para ofrecer un conjunto con alto grado de completitud.	Se recomienda crear un solo conjunto de datos consolidado que ofrezca información completa y organizada, eliminando duplicidad por periodos o tipologías.
ERR012_01	Subconjunto de dato maestro Contratación	Dato es un subconjunto de un dato maestro a nivel nacional	Este conjunto de datos es un subconjunto de un dato maestro relacionado con la plataforma SECOP del Sistema Electrónico de Contratación Pública.	Se recomienda centralizar estos datos en el Portal Nacional de Datos Abiertos y filtrarlos por departamento, municipio o región, reduciendo duplicidad.
ERR012_02	Subconjunto de dato maestro - ICFES	Dato es un subconjunto de un dato maestro a nivel nacional	Este conjunto de datos es un subconjunto de un dato maestro relacionado con el Instituto Colombiano para la Evaluación de la Educación (ICFES).	Se recomienda centralizar estos datos en el Portal Nacional de Datos Abiertos y filtrarlos por departamento, municipio o región, reduciendo duplicidad.
ERR012_03	Subconjunto de dato maestro - Zonas WIFI	Dato es un subconjunto de un dato maestro a nivel nacional	Este conjunto de datos es un subconjunto de un dato maestro de una entidad nacional.	Se recomienda centralizar estos datos en el Portal Nacional de Datos Abiertos y filtrarlos por departamento, municipio o región, reduciendo duplicidad.
ERR012_04	Subconjunto de dato maestro - COVID-19	Dato es un subconjunto de un dato maestro a nivel nacional	Este conjunto de datos es un subconjunto de un dato maestro relacionado con el Instituto Nacional de Salud (INS).	Se recomienda centralizar estos datos en el Portal Nacional de Datos Abiertos y filtrarlos por departamento, municipio o región, reduciendo duplicidad.

ERR012_05	Subconjunto de dato maestro - Trámites y Servicios	Dato es un subconjunto de un dato maestro a nivel nacional	Este conjunto de datos es un subconjunto de un dato maestro relacionado con el Departamento Nacional de Planeación (DNP).	Se recomienda centralizar estos datos en el Portal Nacional de Datos Abiertos y filtrarlos por departamento, municipio o región, reduciendo duplicidad.
ERR012_06	Subconjunto de dato maestro - DNP	Dato es un subconjunto de un dato maestro a nivel nacional	Este conjunto de datos es un subconjunto de un dato maestro relacionado con el Departamento Nacional de Planeación (DNP).	Se recomienda centralizar estos datos en el Portal Nacional de Datos Abiertos y filtrarlos por departamento, municipio o región, reduciendo duplicidad.

Fuente. Elaboración propia

El cuadro analiza diversos errores relacionados con la completitud y la duplicidad de datos, específicamente en subconjuntos de datos maestros a nivel nacional. Estos problemas se presentan cuando los conjuntos de datos están divididos por períodos, tipologías o corresponden a subconjuntos redundantes de datos maestros, como los relacionados con SECOP, ICFES, Zonas WIFI, COVID-19, trámites y servicios, o datos del DNP. Estas divisiones fragmentan la información, dificultan su reutilización y generan duplicidad en los portales de datos abiertos. La solución propuesta es centralizar estos datos en un solo conjunto consolidado, filtrando por categorías como región, municipio o departamento, para garantizar completitud, evitar redundancias y ofrecer datos más útiles y accesibles a los usuarios. Implementar estas recomendaciones no solo mejora la calidad y eficiencia de los datos publicados, sino que también fortalece la transparencia y optimiza su uso para análisis y toma de decisiones.

8.16. El conjunto de datos está mal cargado

Tabla 17. Conjunto de datos mal cargado

Código del error	Categoría de error	Tipo de error	Descripción del 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

El error ERR013 destaca un problema relacionado con la calidad de la publicación de un conjunto de datos, específicamente cuando este está mal cargado. Esto puede incluir errores técnicos o estructurales que afectan la utilidad y confiabilidad de los datos publicados. Estos problemas generan una experiencia negativa para los usuarios y afectan la percepción de calidad de la información proporcionada por la entidad. La solución propuesta es eliminar el conjunto de datos mal cargado y crearlo nuevamente de manera correcta, asegurando que cumpla con los estándares de calidad establecidos. Este enfoque garantiza que los datos publicados sean precisos, completos y útiles para su reutilización, fortaleciendo la transparencia y la confianza en la gestión pública de datos abiertos.

8.17. Desactualizado

Tabla 18. Desactualizado

Código del error	Categoría de error	Tipo de error	Descripción del error	Solución
ERR015	Desactualizado	El conjunto de datos está 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 datos y mejorar la completitud, 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

El error ERR015 aborda un problema de desactualización en los conjuntos de datos, lo cual reduce significativamente su utilidad y relevancia para los usuarios del portal. Este error ocurre cuando los conjuntos de datos no se actualizan conforme a la periodicidad establecida, lo que los convierte en recursos no reutilizables para análisis o toma de decisiones. Esto afecta la confianza en la calidad de los datos abiertos y su capacidad de generar valor público.

La solución propuesta es actualizar el conjunto de datos de manera oportuna y garantizar que cumpla con los estándares de completitud, para que sea de utilidad para los usuarios. En caso de que la información sea de interés particular y no se pueda mantener actualizada en el portal, se recomienda eliminar el conjunto de datos y disponerlo en el sitio web de la entidad. Este enfoque asegura que solo se publiquen datos relevantes, actuales y útiles, fortaleciendo la confianza en la gestión pública de datos abiertos y promoviendo una cultura de transparencia y calidad en la información publicada.

8.18. ERR017 Enlace roto

Tabla 19. Enlace roto

Código del error	Categoría de error	Tipo de error	Descripción del 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). Los datos deben cumplir con los principios de completitud, ser de fuente primaria y presentar un alto nivel de desagregación.	Se recomienda corregir el enlace del conjunto de datos externo, que permita la descarga directa de alguno de los formatos estándares abiertos soportados: csv, xls, xlsx, json, kml, kmz y zip (shapefile de ESRI). Los datos deben cumplir con los principios de completitud, ser de fuente primaria y presentar un alto nivel de desagregación.

Fuente. Elaboración propia

El error ERR017 se relaciona con enlaces rotos que dirigen a direcciones inválidas, lo que impide la descarga de los conjuntos de datos en formatos abiertos y válidos como CSV, XLS, XLSX, JSON, KML, KMZ y ZIP (shapefile de ESRI). Este problema no solo afecta la accesibilidad de los datos, sino también su reutilización y el cumplimiento de los principios de datos abiertos.

La solución propuesta es corregir los enlaces para garantizar que los usuarios puedan acceder directamente a los datos en formatos estándar y abiertos. Además, se resalta la importancia de que estos datos cumplan con los principios de completitud, sean provenientes de fuentes primarias y presenten un alto nivel de desagregación. Resolver este tipo de errores es esencial para mejorar la experiencia del usuario, fortalecer la confianza en los datos abiertos y asegurar que estos sean útiles para análisis y toma de decisiones.

8.19. El conjunto de datos presenta agregaciones o totales

Tabla 20. El conjunto de datos presenta agregaciones o totales

Código del error	Categoría de error	Tipo de error	Descripción del 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

El error ERR018 se refiere a conjuntos de datos que contienen registros agregados o totales, lo que limita su utilidad para análisis detallados o específicos. Este tipo de error contradice los principios fundamentales de los datos abiertos, que establecen que la información debe publicarse en su nivel máximo de desagregación y completitud para maximizar su valor y reutilización por parte de los ciudadanos.

La solución recomendada es garantizar la disponibilidad de la fuente de datos detallada que fue utilizada para generar los reportes. Esto permitirá a los usuarios acceder a la información base y realizar análisis más profundos y personalizados. Además, se sugiere utilizar la funcionalidad de vistas filtradas en el portal de datos abiertos para generar reportes específicos, manteniendo siempre la integridad y desagregación de los datos.

8.20. ITA – Ley de Transparencia y Derecho de Acceso a la Información Pública

Tabla 21. El conjunto de datos presenta agregaciones o totales

Código del error	Categoría de error	Tipo de error	Descripción del 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 en portales de Datos Abiertos como instrumentos de gestión de información pública.	Los datos no hacen parte de los conjuntos publicados en el portal de Datos Abiertos como instrumentos de gestión de información pública.	Como mínimo, el índice de información pública reservada y clasificada, y los registros, deben estar disponibles en los portales. Esto incluye un catálogo de datos, y la publicación de activos en formatos abiertos (CSV).
ITA_2	Errores en activos de información	Error en conjunto de datos de registro de activos de información.	El conjunto de datos de los activos de información no cumple con la estructura establecida o está incompleta.	El portal cuenta con una herramienta activa de diligenciamiento correcto que facilite el cumplimiento normativo y técnico, tal como lo establece la Ley 1712 de 2014 .



TIC



103

ITA_3	Ley de Transparencia y Derecho de Acceso a la Información Pública	El conjunto de datos no está publicado en el portal propio de la entidad.	De acuerdo con la Ley 1712 de 2014²⁶ y la Resolución 1519 de 2020²⁷ , el conjunto de datos no está publicado en el portal propio de la entidad, como mínimo debe incluir: <ol style="list-style-type: none"> 1. Normatividad aplicable. 2. Estructura orgánica. 3. Directorio de servidores públicos. 4. Tarifas. 5. Presupuestos. 6. Información de interés. 7. Otros grupos de interés. 	Se recomienda publicar en el sitio propio de la entidad toda la información correspondiente a la Ley 1712 de 2014 , incluyendo los puntos mencionados: <ol style="list-style-type: none"> 1. Normatividad. 2. Estructura orgánica. 3. Directorio. 4. Tarifas. 5. Presupuestos. 6. Información de interés. 7. Grupos de interés.
-------	---	---	--	---

Fuente. Elaboración propia

El cuadro relacionado con los errores ITA_1, ITA_2 y ITA_3 aborda problemáticas específicas en el cumplimiento de la Ley de Transparencia y Derecho de Acceso a la Información Pública y en la gestión de activos de información.

El error ITA_1 se refiere a la ausencia de datos clave en los portales de Datos Abiertos, lo que impide que estos actúen como instrumentos efectivos de gestión de la información pública. Esto afecta la transparencia y dificulta el acceso de los ciudadanos a información relevante. La solución propone garantizar la publicación mínima de índices de información reservada y clasificada, además de registrar los datos en formatos abiertos y accesibles.

El error ITA_2 identifica inconsistencias en los conjuntos de datos relacionados con los activos de información. Estos conjuntos no cumplen con las estructuras establecidas o están incompletos, lo que compromete su calidad y utilidad. Se recomienda usar herramientas para garantizar el diligenciamiento correcto y cumplir con los requisitos normativos, como los establecidos en la Ley 1712 de 2014.

Finalmente, el error ITA_3 señala que ciertas entidades no publican en sus portales toda la

²⁶ <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=56882>

²⁷ <https://gobiernodigital.mintic.gov.co/portal/Noticias/160997:MinTIC-publica-la-Resolucion-1519-del-2020-sobre-transparencia-en-el-acceso-a-la-informacion-accesibilidad-web-seguridad-digital-web-y-datos-abiertos>

información requerida por la normativa, como normatividad aplicable, estructura orgánica, directorio de servidores públicos, presupuestos, tarifas, y otros datos de interés. La solución consiste en asegurar que toda esta información esté disponible en el portal de la entidad, cumpliendo con la normativa vigente y mejorando la accesibilidad para los ciudadanos

8.21. Subconjunto de datos maestros

Tabla 22. Subconjunto de datos maestros

Código del error	Categoría de error	Tipo de error	Descripción del 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 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

El error relacionado con los subconjuntos de datos maestros señala un problema frecuente en los portales de datos abiertos, donde ciertos conjuntos de datos se publican como subconjuntos de información que ya está consolidada en un dato maestro a nivel nacional. Esto genera redundancia, desactualización y confusión para los usuarios que acceden a la plataforma, dificultando la reutilización y análisis de los datos.

La solución propuesta es eliminar estos subconjuntos de datos duplicados y desactualizados y, en su lugar, crear vistas filtradas directamente desde el conjunto de datos maestro publicado por la entidad productora. Esto garantiza que la información sea única, actualizada y representativa del universo completo de datos, facilitando su uso y evitando inconsistencias.

8.22. Unicidad

Tabla 23. Unicidad

Código del error	Categoría de error	Tipo de error	Descripción del 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

El error relacionado con la unicidad se refiere a la presencia de registros duplicados dentro de un conjunto de datos. Este problema afecta la calidad y confiabilidad de la información, ya que genera redundancia y puede llevar a interpretaciones erróneas o análisis incorrectos. La duplicación de datos también reduce la eficiencia en el manejo de información y afecta la percepción de profesionalismo en la publicación de datos.

La solución propuesta es ajustar el conjunto de datos para garantizar que todos los valores sean únicos, eliminando registros duplicados. Esto no solo mejora la claridad y la calidad del conjunto de datos, sino que también facilita su reutilización y análisis por parte de los usuarios, incrementando su valor para la toma de decisiones.

8.23. Poca reutilización

Tabla 24. Poca reutilización

Código del error	Categoría de error	Tipo de error	Descripción del error	Solución
Uso Aprovechamiento y	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 datos 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

El error relacionado con Uso y Aprovechamiento señala un problema de baja reutilización de los datos abiertos por parte de los ciudadanos. Esto ocurre cuando la información contenida en un conjunto de datos no está estructurada ni presentada de manera que sea considerada útil o accesible como dato abierto. Este problema limita el potencial de los datos para generar productos, servicios o análisis en contextos prácticos, afectando su valor para los usuarios.

La solución propuesta enfatiza la necesidad de actualizar el conjunto de datos para mejorar su pertinencia, asegurando que la información publicada sea relevante y reutilizable. En caso de que la información no sea de interés general o aplicable fuera de la entidad que la produce, se recomienda eliminar el conjunto de datos del portal de datos abiertos y disponerlo únicamente en el sitio web de la entidad. Esto permite optimizar la gestión de datos abiertos, enfocándose en información verdaderamente útil para los ciudadanos.

8.24. Revisión de errores

Tabla 25. 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.
Complejidad	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 cuáles no.

Fuente. Elaboración propia

Se identifican los aspectos clave relacionados con la calidad de los datos y los errores frecuentes asociados a cada dimensión, destacando los desafíos en la gestión de información abierta.

- **Exactitud:** Se identifican problemas relacionados con la precisión de los datos, tanto sintáctica como semántica, debido a la falta de fórmulas específicas que permitan evaluar su exactitud. Esto subraya la necesidad de establecer métodos claros y validados para medir la calidad de los datos en este aspecto.
- **Complejidad:** Aunque existen reglas suficientes para analizar la completitud en términos de columnas, filas y volumen de información, el reto radica en definir fórmulas y niveles de aprobación que garanticen la integridad de los conjuntos de datos, asegurando que sean completos y útiles para su reutilización.
- **Consistencia:** Si bien se cuenta con reglas para subconjuntos y conjuntos globales, se sugiere explorar el uso de inteligencia artificial (IA) y aprendizaje automático (Machine Learning) para analizar de manera más robusta la consistencia de los datos, aplicando modelos propios desarrollados a partir de datos abiertos.
- **Credibilidad:** Este ítem requiere trabajar en técnicas avanzadas de IA y Machine Learning para validar la veracidad y confiabilidad de los datos. Esto es esencial para generar

confianza en la información publicada y evitar posibles malinterpretaciones.

- **Actualidad:** Aunque se cuenta con reglas como "Desactualizado", es importante definir criterios claros sobre hasta qué año o rango temporal se aceptarán datos y cuáles deben ser publicados, asegurando que la información disponible sea relevante y útil para los usuarios.

9. Cambios en los criterios de evaluación

En esta sección se presenta un resumen de los cambios realizados a los criterios de calidad frente a la versión anterior, y que se podrán evidenciar en las herramientas de evaluación disponibles en la sección de calidad del portal de datos abiertos. La tabla siguiente describe el criterio, enuncia sus cambios realizados y la justificación del cambio. Cabe mencionar que dichos cambios son realizados a partir de la información de la guía de calidad e interoperabilidad en su versión 1 del 2022 y los documentos de las pruebas de concepto que describieron la evaluación para la definición, creación y cálculo de la medida justificación, ecuación y algoritmo de los criterios de calidad.

Tabla 26. Tabla de criterios de evaluación de confidencialidad

Criterio	Descripción	Cambios	Justificación
Confidencialidad	Evalúa el grado de protección de los datos personales y sensibles.	Se incorporan los datos anteriores en una categoría llamada seguridad. Se agregan dos nuevas categorías: salud y financieros. Se crea un factor de riesgo para ponderar cada dato dentro de las categorías.	Permite una evaluación que pondera los datos por sus distintos riesgos de identificación. Es adaptativa ante un cambio en las definiciones de riesgo. Es detallada por el cálculo de riesgo en cada categoría.
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.	Se añaden técnicas de procesamiento de lenguaje natural (PLN) para considerar la coherencia semántica de los metadatos en la medida de categoría Se añaden dos factores de columnas y completitud a la medida de número mínimo de filas	Implementar los últimos avances de PLN al criterio y añadir nuevos factores para completar la descripción del criterio
Trazabilidad	Evalúa la proporción de atributos que han sido diligenciados en comparación con aquellos que están incompletos	Añade una penalización por cada metadato faltante a la medida de metadatos diligenciados Asigna diferentes ponderaciones a los metadatos para asignar importancia en datos significativos para la trazabilidad. Además que añade una penalización si el metadato no fue diligenciado en la medida de acceso auditado. Cambia el método de identificación de fechas en el título por un indicador binario en lugar de recorrer una lista de los años 1900 a 2100 para la medida de título sin fecha.	Añadir penalizaciones para darle mayor importancia a los datos que permiten la trazabilidad del conjunto de datos
Conformidad	Relación entre los	Se optimiza la medida de	Identificar con mayor

	valores correctos y los incorrectos en un conjunto de datos	evaluación de columnas para dar eficiencia en la validación de los valores correctos que se usan para comparar. Añade una penalización exponencial para disminuir en mayor magnitud en criterio cada vez que la proporción de los errores en los datos aumente.	sensibilidad el aumento de la proporción de errores en el conjunto de datos
Exactitud sintáctica	Cercanía de los valores de los datos, a un conjunto de valores definidos en un dominio considerado sintácticamente correcto.	<p>Se cambia el método de normalización con librerías reciente de PLN en español.</p> <p>Se añade un método de lematización para mejorar la detección de similitudes entre palabras.</p> <p>Se cambia la librería de comparación de texto debido a la disposición de una más nueva.</p> <p>Se añade un estado de no modificación a la lista original que contiene las palabras únicas.</p> <p>Se añade una propiedad cuadrática al criterio para dar mayor sensibilidad o mayor impacto de los errores en el total de columnas de texto.</p>	Implementar librerías más recientes de PLN al proceso base de similitud y agregar una penalización de mayor impacto que otorgue mayor relevancia a los errores en el total de columnas con textos
Exactitud semántica	Cercanía de los valores de los datos, a un conjunto de valores definidos en un dominio considerado semánticamente correcto.	<p>Se agrega una penalización por función cuadrática a los errores en campos de textos</p> <p>Se modifica la fórmula del criterio para agregar un parámetro de suavizado a la caída por la nueva penalización.</p> <p>Se agrega un límite para que el criterio con la nueva fórmula no sea negativa.</p>	Se añade una propiedad cuadrática al criterio para dar mayor sensibilidad o mayor impacto y adaptabilidad a la cantidad de columnas y la proporción de errores encontrados.
Completitud	Datos completamente diligenciados.	<p>Añade una penalización exponencial para pronunciar más la caída de las medidas de completitud datos cuando hay datos en celdas nulas.</p> <p>Añade una penalización exponencial para pronunciar más la caída de las medidas de completitud columnas cuando</p>	Se añade una propiedad de penalización exponencial para identificar con mayor claridad los cambios en proporción a los errores de información completa.

		<p>hay columnas incompletas.</p> <p>Añade una proporción por columnas vacías con respecto al total de columnas en lugar de colocar 0 en la medida de columnas no vacías</p>	
Consistencia	Los datos son consistentes cuando están libres de contradicción y son coherentes respecto a otros datos en el mismo contexto de uso.	<p>Se agrega un <i>docstring</i> en las dos medidas de: atributo de columnas de texto con valores cortos y consistencia por número de caracteres para documentar mejor los parámetros de entrada y salida</p> <p>Reduce el tiempo de ejecución de preprocessamiento de texto con un método más sencillo que permite el procesamiento rápido de grandes volúmenes de datos textuales en la medida de atributo de columnas de texto con valores cortos</p> <p>Se añade un factor de penalización a la medida de consistencia por número de caracteres para dar mayor impacto en la inconsistencia de esta medida</p>	<p>Se añade factor de penalización y simplificación de métodos de preprocessamiento para mejorar los tiempos de cálculo del criterio. Además, se mejora la documentación para explicar mejor los resultados del criterio</p>
Precisión	Los conjuntos de datos se deben publicar con el más alto nivel de desagregación	<p>Se incluye una varianza para evaluar la dispersión de los valores en el objetivo de detectar columnas con información limitada</p> <p>El criterio de valores únicos asegura que las columnas tengan suficiente diversidad en sus datos.</p>	<p>Los cambios en la función ofrecen una evaluación más completa de la calidad de los datos al integrar múltiples criterios</p>
Portabilidad	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.	<p>Se crea una nueva función para el cálculo del criterio que incluye aspectos de presencia de caracteres especiales, ausencia de valores nulos y el tamaño de conjuntos de datos</p>	<p>Al considerar múltiples factores que afectan la capacidad de los datos para ser transferidos y reutilizados, se logra una puntuación que refleja de manera más fiel la realidad del conjunto de datos.</p>
Comprendibilidad	Los datos deben	<p>Se introduce una función</p>	Estas modificaciones introducen

	poseer atributos que permitan ser leídos e interpretados por los usuarios.	exponencial para la medida de descripción con tamaño mínimo requerido aumentando a medida que se acerca al mínimo requerido Se añade una función logarítmica para calcular el puntaje de forma no lineal en función de la longitud de la etiqueta	cálculos no lineales, lo que permite que el puntaje refleje mejor la calidad de los datos de entrada sin depender de límites estrictos.
Accesibilidad	Evalúa si los datos incluyen metadatos que permiten a los usuarios encontrar y acceder fácilmente a la información	Cuenta la cantidad de parámetros de etiquetas o enlaces encontrados. Si existe el parámetro, asigna 5 puntos.	Incluir un conteo de parámetros que evidencie la cantidad de insumos existentes para dar accesibilidad real al conjunto de datos.
Unicidad	Grado en que los datos no contienen duplicados, tanto a nivel de filas como en columnas clave	Agregar los tres niveles de riesgo de forma que un riesgo mayor hace una penalización de mayor magnitud en el valor del criterio.	Agregar los niveles de riesgo para darle importancia a la información que representa.
Eficiencia	Mide la completitud y ausencia de duplicados en los datos.	Se agrega un cálculo con las medidas de: Medida completitud en datos, Medida columnas no duplicadas y Medida filas no duplicadas, en lugar de un valor fijo 10.	Reflejar las realidades de los conjuntos para la descripción del criterio más allá de un valor fijo colocado arbitrariamente.
Recuperabilidad	Evalúa la facilidad de recuperación y comprensión de los datos, con énfasis en los metadatos y accesibilidad.	Se agrega un cálculo con las medidas de: Medida metadatos completos, Medida accesibilidad y Medida metadatos acceso auditado.	Reflejar las realidades de los conjuntos para la descripción del criterio más allá de un valor fijo colocado arbitrariamente.
Disponibilidad	Verifica que los datos estén actualizados y sean accesibles.	Se agrega un cálculo con las medidas de: Medida actualidad y Medida accesibilidad.	Reflejar la realidad de los conjuntos para la descripción del criterio más allá de un valor fijo colocado arbitrariamente.



10. 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.

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.

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.

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.

11. Referencias

- Estado Peruano. (s.f.). Infraestructura 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
- Presidencia de la Republica de Colombia. (2019, Septiembre). Presidencia de la Republica de Colombia. Retrieved from <https://dapre.presidencia.gov.co/dapre/DocumentosSIGEPRE/G-GD-02-calificacion-informacion.pdf>