
METODOLOGÍA DE GESTIÓN DE DATOS ABIERTOS

***Versión: 002
Alcaldía Municipal de Chía
Julio de 2025***

Carrera 7 N° 12-100
PBX: (601) 884 4444 Ext. 2300-2301
oficinatic@chia.gov.co
www.chia-cundinamarca.gov.co

Tabla de contenido

	PÁG.
INTRODUCCIÓN	3
NORMATIVA SOBRE DATOS ABIERTOS EN COLOMBIA	4
DEFINICIONES Y CONCEPTOS BÁSICOS SOBRE DATOS ABIERTOS	5
CRITERIOS DE CALIDAD DE DATOS ABIERTOS	8
ARCHIVOS PLANOS Y FORMAS DE PRESENTACIÓN	9
MANEJO DE DATOS ABIERTOS EN LA ALCALDÍA MUNICIPAL DE CHÍA.	11
ELEMENTOS DE CALIDAD PARA DATOS ABIERTOS	16
REQUISITOS DE PRESENTACIÓN DE REPORTES DE LAS DIFERENTES DEPENDENCIAS	19
GLOSARIO	20
REFERENCIAS	22

INTRODUCCIÓN

La implementación de **datos abiertos** en la Alcaldía Municipal de Chía representa un avance fundamental hacia un gobierno más **transparente, participativo y eficiente**. Estos datos, disponibles en formatos digitales accesibles y reutilizables, permiten a la ciudadanía, empresas y academia acceder a información clave sobre gestión pública, servicios municipales y proyectos estratégicos. En Colombia, este proceso está respaldado por un marco normativo robusto, que incluye la **Ley 1712 de 2014** (Ley de Transparencia), el **Decreto 1081 de 2015** y la **Política Nacional de Datos Abiertos**, los cuales exigen a las entidades públicas garantizar la divulgación proactiva de información bajo principios de apertura y calidad.

La importancia de los datos abiertos en Chía radica en su capacidad para **fortalecer la rendición de cuentas** y promover la **innovación social y económica**. Al publicar datos sobre cada una de las dependencias de la administración municipal, la Alcaldía no solo cumple con la normativa legal, sino que también fomenta la **participación ciudadana** y el control social. Esto se alinea con estándares internacionales como los de la **Open Government Partnership (OGP)**, a los que Colombia está adherida, y que buscan mejorar la gobernanza mediante la tecnología y la apertura de información.

Además, los datos abiertos son una **herramienta clave para el desarrollo local**, ya que permiten a emprendedores, investigadores y organizaciones crear soluciones basadas en evidencia. Por ejemplo, el análisis de datos de movilidad puede optimizar rutas de transporte, mientras que la transparencia en contratos públicos reduce riesgos de corrupción. La **Plataforma de Datos Abiertos de Colombia (datos.gov.co)** sirve como referente para que Chía consolide su propia estrategia, asegurando que los conjuntos de datos cumplan con criterios de **interoperabilidad, oportunidad y usabilidad**.

En este contexto, la Alcaldía de Chía tiene la oportunidad de posicionarse como un **referente en gobierno abierto** dentro de la región, aprovechando los datos como un activo público que impulsa la **competitividad y la confianza institucional**. El cumplimiento de la normativa no es solo un requisito legal, sino una apuesta por la modernización de la gestión pública, donde la tecnología y la transparencia se convierten en pilares para construir una **sociedad más informada y colaborativa**. La adopción de estas prácticas no solo beneficia a la administración municipal, sino que también empodera a los ciudadanos como actores clave en la construcción de un **territorio más inteligente y sostenible**.

NORMATIVA SOBRE DATOS ABIERTOS EN COLOMBIA

En Colombia, la constitución política consagra el acceso a la información como derecho fundamental en los artículos 20 y 74, reconociéndolo como un derecho habilitante de otros derechos. partiendo de esta premisa, Colombia adopta un marco legal que garantiza la transparencia y el derecho de acceso a la información pública a los ciudadanos en la ley 1712 de 2014¹, que establece procedimientos para garantizar este derecho y se obliga a las entidades, en el marco de la seguridad y privacidad de la información respetando los datos reservados y clasificados, a publicar o liberar datos abiertos.

Como instrumentos normativos adicionales tenemos:

Decreto 1078 de 2015²: "...sus objetivos y funciones se encuentran definidos en la ley 1341 de 2009, "por la cual se definen principios y conceptos sobre la sociedad de la información y la organización de las tecnologías de la información y las comunicaciones -TIC-, se crea la agencia nacional de espectro y se dictan otras disposiciones..."

Decreto 1081 de 2015³: en su título 1, "disposiciones generales en materia de transparencia y del derecho de acceso a la información pública nacional".

Resolución 3564 de 2015 de MinTIC⁴: "...tiene por objeto establecer los lineamientos respecto de los estándares para publicación y divulgación de la información..."

Desde la Alcaldía Municipal de Chía, buscamos enmarcar un plan de trabajo inmerso en la normatividad vigente estipulada previamente por el Ministerio de las TIC donde se garantizará la implementación de estrategias que permitan a la ciudadanía ejecutar procesos de participación en base a los proyectos de gobierno abierto y territorios inteligentes reduciendo así el riesgo de corrupción y manipulación de información sensible la cual puede ser consultada por cada uno de los ciudadanos.

¹ <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=56882>

² <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=77888>

³ <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=73593>

⁴ https://normograma.mintic.gov.co/mintic/compilacion/docs/resolucion_mintic_3564_2015.htm

DEFINICIONES Y CONCEPTOS BÁSICOS SOBRE DATOS ABIERTOS

Un **DATO ABIERTO** es información pública, generada o custodiada por entidades gubernamentales u organizaciones de interés público, que fundamenta su origen en la ley de transparencia de la información. Algunas de las características que deben tener los datos abiertos son:

Gratuitos	- Los datos abiertos deben estar disponibles sin costo alguno para los ciudadanos, empresas y organizaciones. Esto elimina barreras económicas y fomenta la igualdad en el acceso a la información pública. Ejemplo: La Alcaldía de Chía no puede cobrar por descargar datos sobre presupuesto municipal o contratación pública.
Accesibles	- Los datos deben publicarse en plataformas digitales de fácil acceso, preferiblemente en un portal único (como datos.gov.co), y sin restricciones técnicas (como requerir registros innecesarios). Ejemplo: Datos de transporte público en Chía deben estar en un enlace directo y descargable, no solo en informes impresos.
Reutilizables	- Los datos deben publicarse con licencias abiertas (como Creative Commons CC-BY 4.0) que permitan su uso, modificación y redistribución, incluso con fines comerciales. Ejemplo: Una universidad puede usar datos demográficos de Chía para un estudio académico, o una startup puede desarrollar una app de rutas de buses.
Oportunos	- La información debe actualizarse con una frecuencia definida (ej.: mensual, trimestral) para mantener su relevancia. Ejemplo: Los datos de contratación en Chía deben publicarse mensualmente, no solo al finalizar el año. esto se hace a través de vistas filtradas.
Estructurados	- Los datos deben estar en formatos procesables (como CSV, JSON, XML) y no en documentos estáticos (PDF, imágenes). Esto permite su análisis automatizado. Ejemplo: El padrón electoral de Chía en formato CSV puede ser leído por programas de estadística, mientras que un PDF escaneado no.
Interoperables	- Deben seguir estándares técnicos que permitan su conexión con otros sistemas o bases de datos (ej.: esquemas de metadatos, vocabularios comunes). Ejemplo: Datos de salud en Chía deben usar clasificaciones nacionales (como CUPS) para compararse con otros municipios.

Integrales

- Los conjuntos de datos deben ser **completos y exactos**, sin omisiones clave que limiten su utilidad.
- **Ejemplo:** Si la Alcaldía publica datos de accidentes viales, debe incluir ubicación, fecha, tipo de vehículo, etc., no solo cifras generales.

No discriminatorios

- El acceso debe estar abierto a **todos los ciudadanos**, sin requerir permisos especiales o justificación de uso.

Para el 2016, Colombia adoptó la **Carta Internacional de Datos Abiertos**; ésta como instrumento orientador y generados de datos en las estrategias de transparencia organizadas por el Gobierno Nacional. Al ser un adherente de esta Carta, Colombia acepta que los datos abiertos se rigen por los principios de:

PRINCIPIO	¿A QUÉ SE REFIERE?
Datos Abiertos por Defecto	Los datos públicos deben ser abiertos por defecto , salvo razones legales o de privacidad que lo impidan.
Oportunos y Exhaustivos	Los datos deben ser publicados de manera oportuna (actualizados) y exhaustiva (incluyendo toda la información relevante).
Accesibles y Utilizables	Los datos deben estar en formatos abiertos (CSV, JSON) y en plataformas accesibles (como datos.gov.co).
Comparables e Interoperables	Los datos deben seguir estándares comunes para integrarse con otros sistemas (nacionales o internacionales).
Para Mejorar la Gobernanza y la Participación Ciudadana	Los datos deben empoderar a la ciudadanía para tomar decisiones y ejercer control político y social .

*Para el Desarrollo Incluyente y la
Innovación*

Los datos deben usarse para reducir
desigualdades y fomentar **innovación social y
económica**.

Aunado a esto, es de resaltar que, al referirnos a datos abiertos, éstos **NO SON CONTENEDORES DE INFORMACIÓN SENSIBLE**, es decir, no incluyen datos que puedan ser de carácter privado y/o reservado. En este sentido, y para dar cumplimiento a las políticas de acceso a la información, surge el concepto de **ANONIMIZACIÓN**, este concepto se desarrolla sobre **MICRODATOS**, es decir,

“... cada uno de los datos sobre las características de las unidades de estudio de una población (individuos, hogares, establecimientos, entre otras) que se encuentran consolidados en una base de datos...” (Decreto 1743 de 2016: Art. 2.2.3.1.1).

Y se centra en

“... transformar los datos individuales de las unidades de observación, de tal modo que no sea posible identificar sujetos o características individuales de la fuente de información, preservando así las propiedades estadísticas en los resultados...” (Decreto 1743 de 2016: Art. 2.2.3.1.1).

El proceso de anonimización de una base de datos se encuentra compuesto por seis etapas que son: **i)** Revisiones previas; **ii)** Análisis de riesgos de identificación de las fuentes de información; **iii)** Identificación y selección de técnicas de anonimización; **iv)** Análisis de viabilidad del proceso, **v)** Aplicación de técnicas de anonimización, y **vi)** Evaluación de resultados del proceso.

Basados en la normativa legal, desde la Alcaldía Municipal de Chía, entendemos que resultan importantes los conceptos de:

Metadato

Son "datos sobre los datos" que describen su contenido, origen, formato, frecuencia de actualización y otros atributos técnicos. Facilitan la **búsqueda y comprensión** de los conjuntos de datos.

Estándar

Un **estándar** en datos abiertos es una especificación técnica acordada (nacional o internacionalmente) que define cómo deben estructurarse, describirse o publicarse los datos para garantizar su **interoperabilidad y calidad**.

Formato abierto

Formatos digitales **no propietarios**, que pueden ser procesados por software libre o estándar, sin restricciones legales o técnicas.

Reutilización de datos

Es el uso de datos abiertos por terceros (ciudadanos, empresas, academia) para crear **servicios derivados**, análisis o aplicaciones, sin restricciones legales.

CRITERIOS DE CALIDAD DE DATOS ABIERTOS

Con criterios de calidad para la evaluación de datos abiertos:

Criterio de confidencialidad	Criterio de relevancia	Criterio de actualidad	Criterio de trazabilidad
Criterio de conformidad	Criterio de exactitud	Criterio de completitud	Criterio de consistencia
Criterio de precisión	Criterio de portabilidad	Criterio de credibilidad	Criterio de comprensibilidad
Criterio de accesibilidad	Criterio de eficiencia	Criterio de recuperabilidad	Criterio de disponibilidad

Estos pueden ser consultados ampliamente en:

<https://herramientas.datos.gov.co/sites/default/files/2020->

CICLO DE DATOS



Establecer un ciclo para el proceso de apertura, mejora y uso de datos abiertos permite que los actores del ecosistema definan acciones a tomar en cada una de las fases estratégicas de este ciclo: construcción de plan de apertura, estructuración, comunicación y monitoreo. La buena implementación y seguimiento a estas actividades redundará en calidad de datos y, por tanto, en usos de impacto. Adicionalmente es un proceso recurrente, que debe presentarse con una vigencia definida, como recomendación este ciclo se planea y ejecuta cada dos años.

1<https://herramientas.datos.gov.co/sites/default/files/>

ARCHIVOS PLANOS Y FORMAS DE PRESENTACIÓN

Los **archivos planos** (o *flat files*) son un tipo de estructura de datos que almacena información en formato de texto sin relaciones jerárquicas o tablas vinculadas (a diferencia de las bases de datos relacionales). Son ampliamente usados en **datos abiertos** por su simplicidad, portabilidad y facilidad de procesamiento.

CARACTERÍSTICAS PRINCIPALES:

1. Simplicidad Estructural

- Son como bloc de notas con datos
- Sin fórmulas, gráficos ni relaciones complejas
- Ejemplo real: Un CSV de contratos públicos con solo 5 columnas: ID, Proveedor, Valor, Fecha

2. Autocontenidos

- Todo está en un solo archivo

- No dependen de otros archivos para funcionar
- Caso típico: Un TXT con registros de asistencia a eventos municipales

3. Legibilidad Universal

- Los puede abrir hasta el Notepad más básico
- No requieren software especializado

- Ejemplo práctico: Un TSV con resultados de pruebas educativas que cualquier padre puede revisar
- 4. Delimitación Clara**
Usan separadores consistentes:
- CSV: comas (val1,val2,val3)
 - TSV: tabulaciones (val1 val2 val3)
 - Pipe: | (val1|val2|val3)
- 5. Metadatos Externos**
- La explicación de campos va aparte

- Buen ejemplo: Un archivo CSV de presupuesto municipal acompañado de un PDF explicativo

6. Problemas Comunes

- Falta de estandarización
- Errores con caracteres especiales (tildes, eñes)
- Confusión en delimitadores cuando hay textos largos

Los archivos planos se clasifican, básicamente, de dos formas:

1. Por Formato de Almacenamiento

Esta clasificación depende de cómo se organizan los datos dentro del archivo:

Tipo	Descripción	Ejemplo en Datos Abiertos
Delimitados	Campos separados por un carácter (coma, tabulación, pipe). Ejemplos: CSV, TSV.	<code>contratos_chia_2023.csv</code> (columnas separadas por comas).
Ancho Fijo	Cada campo ocupa un número fijo de caracteres (sin delimitadores). Ejemplo: Formatos bancarios antiguos.	<code>paz_y_salvo.txt</code> (campo "valor" siempre ocupa 10 caracteres).
JSON/XML	Aunque son estructurados, se consideran "planos" si no hay relaciones externas. Ejemplo: Datos de sensores en JSON.	<code>calidad_aire_chia.json</code> (una línea por registro).

2. Por Tipo de Contenido

Esta clasificación depende del uso o semántica de los datos:

Tipo	Descripción	Ejemplo
Tabulares	Datos en filas/columnas (ej.: hojas de cálculo).	<code>matriculas_educativas.csv</code>
Registros	Datos transaccionales (ej.: logs, facturas).	<code>historico_pagos.txt</code> (un pago por línea).
No estructurados	Texto libre, pero almacenado en archivo plano.	<code>actas_consejo_2023.txt</code>

Dentro de la Alcaldía Municipal de Chía y teniendo en cuenta que TODAS⁵ las dependencias de la misma están obligadas por Ley a reportar los datos abiertos y a publicarlos. Los lineamiento básicos de presentación para datos abiertos y validación en el portal datos.gov.co son:

Tipo de archivo	PLANO de preferencia formato Excel (.xlsx)
Extensión	Filas: 61 (mínimo) Columnas: 05 (mínimo)

⁵ La Dirección de Control Interno Disciplinario, en la administración municipal de Chía, no realiza un reporte para datos abiertos pues la información que ésta Dirección procesa y su misionalidad es de carácter confidencial y/o reservado. **NO HAY** ninguna otra dependencia que se encuentre excluida de la presentación de los reportes correspondiente.

Completez	Todas las celdas del reporte deben contener información, es decir, NO DEBE haber ninguna celda en blanco. De presentarse el caso en el que las celdas no coincidan por la característica de la columna deben ser completadas así: - Si el valor es numérico: con el 0 (cero) - Si el valor es de texto: con las siglas ND cuando el valor no está definido o NA cuando no aplica. NO deden ser escritas variaciones de éstas, como: N/A, o, N.A., ni ninguna que no sea especificada aquí.
Nombre	El archivo debe ser nombrado como será reportado en el portal de datos abiertos el conjunto que explica el mismo. Es decir, si el conjunto, por ejemplo, responde a procesos de acompañamiento de la secretaria de medio ambiente, el nombre del archivo debería ser: “Procesos de acompañamiento de la Secretaría de Medio Ambiente del Municipio de Chía”
Formato	El archivo reportado NO debe contener: • Celdas combinadas • Títulos incluidos dentro del formato • Colores • Fórmulas • Celdas vacías • Imágenes • Filtros • Filas o columnas fijas

MANEJO DE DATOS ABIERTOS EN LA ALCALDÍA MUNICIPAL DE CHÍA.

ALGUNOS EJEMPLOS DE CONJUNTO DE DATOS DEL MUNICIPIO DE CHÍA Y CARACTERÍSTICAS

Todos los elementos aquí mostrados se encuentran publicados en el portal de datos abiertos <https://www.datos.gov.co/>. Explicaremos las características primordiales de los conjuntos que reposan en el sitio.

A continuación presentaremos un ejemplo de conjunto de datos en estado **PÚBLICO** dentro del portal <https://www.datos.gov.co/> y sus características.

ACTIVIDADES DE LA BIBLIOTECA HOQABIGA DEL MUNICIPIO DE CHÍA	
Filas	15,3k
Columnas	9

Dentro del sitio es posible observar las siguientes opciones:

ACTIVIDADES DE LA BIBLIOTECA HOQABI...
 **Público**

Información sobre esta Categoría
Datos
Contenido Relacionado

En esta sección encontrará las características generales que solicita el portal para la publicación de un conjunto de datos.

A continuación, mostramos los elementos que allí se esbozan:

ACTIVIDADES DE LA BIBLIOTECA HOQABIGA DEL MUNICIPIO DE CHÍA
Cultura

La Biblioteca Hoqabiga de Chía ofrece talleres, charlas, lectura y eventos culturales, promoviendo el acceso a la información y la cultura

Última Actualización
10 de febrero de 2025

Datos suministrados por
ALCALDIA MUNICIPAL DE CHIA

Nombre del conjunto de datos, asignado en el archivo reportado por cada dependencia, la fecha de actualización del conjunto y el nombre de la Alcaldía Municipal de Chía como responsable de la misma.

Actualizado
10 de febrero de 2025

Última actualización de los datos
19 de diciembre de 2024

Última actualización de metadatos
10 de febrero de 2025

Fecha de creación
14 de noviembre de 2024

Vistas
143

Descargas
81

Suministró los datos
ALCALDÍA MUNICIPAL DE CHÍA

Propietario de conjunto de datos
Alcaldía Municipal de Chía

Información de la Entidad

Departamento	Cundinamarca
Municipio	Chía
Nombre de la Entidad	Alcaldía de Chía
Orden	Territorial
Sector	No Aplica
Área o dependencia	DIRECCIÓN DE CULTURA

Información de Datos

Idioma	Español
Cobertura Geográfica	Municipal
Frecuencia de Actualización	Anual

Temas

Categoría	Cultura
Etiquetas	Este conjunto de datos no tiene ninguna etiqueta

Licencia y atribución

Licencia	
----------	--

En esta sección se deja evidencia de la **METADATA** del conjunto que ha sido reportado, es decir, las características generales del conjunto y su forma de identificación. Además, en esta sección se observa el licenciamiento del que hace parte el conjunto y con el que se reportan todos los conjuntos de datos.

Filas
15,3K

Columnas
9

Cada fila es un
BibliotecaHoqabiga, Actividades Culturales, Cultura Chía, Eventos Culturales

En esta sección, las características elementales, en cuanto al tamaño del archivo que se ha reportado.

Columnas (9)

Nombre de la columna	Descripción	Nombre del campo API	Tipo de Dato
 Fecha	Fecha de realización	fecha	Texto
 Actividad	Nombre de la actividad	actividad	Texto
 Línea de Servicio	Área temática a la que pertenece la actividad	l_nea_de_servicio	Texto
 Espacio donde se desarrolla la actividad	Lugar de la biblioteca donde se desarrolló	espacio_donde_se_desarrolla	Texto
 Asistencias	Número total de personas que asistieron	asistencias	Número
 Rangos de Edad	Rango de edad de los asistentes	rangos_de_edad	Texto
 Femenino	Número de mujeres asistentes	femenino	Número
 Masculino	Número de hombres asistentes	masculino	Número
 LUGAR	Sede específica de la biblioteca	lugar	Texto

Rows per page: 1-9 of 9 < >

En esta sección, la caracterización de cada una de las columnas del conjunto de datos reportado, así como el tipo de información contenida en cada una de las celdas. En la parte inferior, se observa el lugar (cantidad de páginas) que el conjunto de datos ocupa en el portal.

ACTIVIDADES DE LA BIBLIOTECA HOQABI...		 Público
Información sobre esta Categoría	Datos	Contenido Relacionado

En ésta, el conjunto de datos reportado con todas las características que se mencionaron anteriormente en este documento.



APARIENCIA GENERAL DE UN
CONJUNTO DE DATOS EN EL

ACTIVIDADES DE LA BIBLIOTECA HOQABIGA DEL MUNICIPIO DE CHÍA

Q Buscar

Fecha	Actividad	Línea de Servicio	Espacio donde se desarrolla ...	Asistencias	Rangos de Edad	Femenino	Masculino	LUGAR
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	1	0	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	0	1	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	0	1	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	1	0	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	0	1	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	0	1	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	0	1	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	0	1	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADULTOS (29 A 59 AÑOS)	0	1	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	NIÑOS Y NIÑAS (6 A 12 AÑOS)	1	0	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	NIÑOS Y NIÑAS (6 A 12 AÑOS)	0	1	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	1	ADOLESCENTES (13 A 17 AÑOS)	1	0	CENTRO
2024	CONSULTA EN SALA (LIBROS, PE	CONSULTA EN SALA	SALA GENERAL	2	ADOLESCENTES (13 A 17 AÑOS)	2	0	CENTRO

Lo expuesto en esta sección corresponde a los elementos generales de los conjuntos de datos abiertos publicados en el portal de la Alcaldía Municipal de Chía.

ELEMENTOS DE CALIDAD PARA DATOS ABIERTOS

En el contexto de la Alcaldía de Chía, Cundinamarca, la calidad de los datos abiertos es un pilar fundamental para la transparencia, la toma de decisiones informada y la participación ciudadana. Para garantizar que los datos publicados sean precisos, consistentes, completos, válidos, únicos y oportunos, la Alcaldía ha adoptado **OpenRefine** como la herramienta principal para la mejora de la calidad de sus conjuntos de datos.

OpenRefine (anteriormente Google Refine) es una aplicación de escritorio de código abierto y auto-alojada para la manipulación y transformación de datos semi-estructurados. Su arquitectura se basa en un servidor web local (servidor Jetty incrustado) que se ejecuta en la máquina del usuario, y una interfaz de usuario accesible a través de cualquier navegador web estándar (generalmente en <http://127.0.0.1:3333/>). No es una herramienta de base de datos en sí misma, sino un entorno interactivo y visual para la preparación de datos (data wrangling o data munging), con un enfoque particular en la mejora de la calidad de los datos.

Desde el concepto de calidad de los datos, OpenRefine aborda las siguientes dimensiones y características clave:

PERFILADO DE DATOS (DATA PROFILING) Y DESCUBRIMIENTO INTERACTIVO

- **Facetas (Faceting):** Es la característica central para el perfilado. Permite la exploración multidimensional de los datos.
- **Facetas de Texto:** Agrupan valores de texto únicos y muestran sus recuentos de ocurrencias. Son cruciales para identificar inconsistencias ortográficas (errores de digitación), variaciones en la capitalización, espaciado inconsistente o sinónimos no normalizados. La visualización de la distribución de valores revela directamente anomalías.
- **Facetas Numéricas/Personalizadas:** Permiten el análisis de la distribución de valores numéricos a través de histogramas interactivos, facilitando la detección de valores atípicos (outliers), rangos inválidos o datos faltantes implícitos.
- **Facetas por Dispersión (Scatterplot Facet):** Para datos numéricos de dos dimensiones, ayuda a visualizar correlaciones y aglomeraciones, identificando patrones o anomalías que sugieren errores en la medición o registro.
- **Clústeres (Clustering):** Emplea algoritmos de coincidencia aproximada (approximate matching algorithms) como n-gram fingerprint, metaphone, y key collision para identificar automáticamente entradas que son "casi" idénticas pero tienen ligeras variaciones. Esto es fundamental para la unificación de datos y la reducción de duplicidad, abordando la dimensión de la consistencia y la unicidad.
- **Filtros (Filtering):** Permiten la segmentación y el enfoque en subconjuntos específicos de datos para inspección detallada, usando expresiones regulares, condiciones numéricas, o condiciones booleanas. Esto facilita la localización de registros que incumplen reglas de negocio o patrones esperados.

ESTANDARIZACIÓN Y NORMALIZACIÓN DE DATOS (STANDARDIZATION & NORMALIZATION)

Transformaciones GREL (General Refine Expression Language): Un lenguaje de expresión basado en JavaScript pero con una sintaxis más concisa, diseñado para la manipulación de cadenas, fechas, números y estructuras de datos. Permite la aplicación de reglas de limpieza complejas de forma programática. Ejemplos incluyen:

- `value.trim()`: Elimina espacios en blanco iniciales/finales (aborda la consistencia del formato).
- `value.toUpperCase()` / `value.toLowerCase()`: Normaliza la capitalización (aborda la consistencia de la representación).
- `value.replace("viejo", "nuevo")`: Estandariza términos.
- **Manipulación de fechas/horas:** Conversión a formatos estándar (AAAA-MM-DD), vital para la precisión temporal.
- **Operaciones de Celda y Columna:**

- Dividir columnas: Basado en delimitadores o expresiones regulares, útil para granularidad adecuada y separación de conceptos mezclados en una sola columna.
- Unir columnas: Para consolidar información relevante que estaba dispersa.
- Reordenar y renombrar columnas: Mejora la legibilidad y la interpretabilidad del esquema de datos.
- Rellenar hacia abajo/arriba (Fill down/up): Propagar valores en celdas vacías, crucial para datos jerárquicos o anidados donde la información solo se registra una vez.

DETECCIÓN Y RESOLUCIÓN DE ERRORES (ERROR DETECTION & RESOLUTION)

- Visualización Guiada: La interfaz de usuario intuitiva permite a los usuarios identificar visualmente los errores y aplicar transformaciones de forma interactiva, lo que reduce la curva de aprendizaje y acelera el proceso de limpieza.
- Aplicación de Transformaciones en Lote (Batch Transformations): Las operaciones aplicadas a una celda pueden ser propagadas a todas las celdas coincidentes dentro de una faceta o a toda la columna, asegurando la consistencia global tras la corrección de errores.
- Soporte para Expresiones Regulares (Regex): Permite la identificación y extracción de patrones complejos, así como la validación de formatos, contribuyendo a la exactitud y la conformidad de los datos.

ENRIQUECIMIENTO Y RECONCILIACIÓN DE DATOS (DATA ENRICHMENT & RECONCILIATION)

- Reconciliación con Servicios Externos (Reconciliation with External Services): Permite comparar y mapear valores de datos locales con entidades en bases de conocimiento externas (ej., Wikidata, VIAF, GeoNames) utilizando APIs estándar como el protocolo de reconciliación de Google. Esto es fundamental para:
 - Mejorar la completitud: Añadir atributos faltantes.
 - Aumentar la exactitud: Validar la existencia y unicidad de entidades.
 - Desambiguar: Resolver ambigüedades (ej., "Washington" ¿estado, ciudad, persona?).
- Normalizar a identificadores persistentes: Reemplazar cadenas de texto con URIs o IDs estables, lo que mejora la interoperabilidad y la vinculación de datos (Linked Data).
- Extracción de Datos (Extract Data): Permite aplicar transformaciones para generar nuevas columnas basadas en la extracción de información de otras columnas, enriqueciendo el conjunto de datos con información derivada.

TRAZABILIDAD Y REPLICABILIDAD (TRACEABILITY & REPLICABILITY)

- Historial de Operaciones (Undo/Redo): Cada operación realizada en OpenRefine se registra en un historial, permitiendo deshacer y rehacer cambios. Esto es crítico para la auditoría y la recuperación de errores.
- Exportar/Aplicar Operaciones (Export/Apply Operations): El conjunto de todas las transformaciones aplicadas puede ser exportado como un script JSON. Este script puede ser

reproducido en otros conjuntos de datos con una estructura similar, garantizando la consistencia en los procesos de limpieza y facilitando la automatización y la reutilización de flujos de trabajo. Esto es vital para la fiabilidad y la replicabilidad de los resultados de limpieza.

ESCALABILIDAD Y FORMATOS DE SALIDA

- Manejo de Conjuntos de Datos Grandes (sin sobrecargar la RAM): Aunque es una aplicación de escritorio, OpenRefine está diseñado para manejar eficientemente conjuntos de datos de millones de filas, gestionando la memoria de forma inteligente, lo que lo hace adecuado para la limpieza de big data a pequeña escala.
- Múltiples Formatos de Exportación: Permite exportar los datos limpios en diversos formatos (CSV, TSV, Excel, JSON, XML, bases de datos SQL, RDF Turtle/N-Triples), facilitando la integración de los datos procesados con otros sistemas o herramientas analíticas.

En resumen, OpenRefine es una herramienta excepcionalmente potente para la curación de datos, proporcionando un conjunto rico de funcionalidades que abordan directamente las dimensiones críticas de la calidad de los datos: exactitud, completitud, consistencia, unicidad, validez, actualidad y oportunidad, a través de su enfoque interactivo, visual y programático. Su capacidad para combinar la exploración de datos con transformaciones potentes lo convierte en un componente esencial en cualquier proceso de pipeline de datos que requiera un alto nivel de fiabilidad y limpieza.

REQUISITOS DE PRESENTACIÓN DE REPORTES DE LAS DIFERENTES DEPENDENCIAS

En cada una de las entregas realizadas, las diferentes dependencias deberán entregar:

- Archivo plano exportado desde OpenRefine según especificaciones dadas en ciclos de capacitación y en la presente metodología.
- Pantallazo de OpenRefine
- Formato **“REPORTE DE CONJUNTOS DE DATOS ABIERTOS POR DEPENDENCIA”** completamente diligenciado

GLOSARIO

- **Pipeline (de datos):** Se refiere a un sistema utilizado para mover datos de un origen a un destino, como un lago de datos o un almacén de datos. Pueden ser en tiempo real, basados en la nube, de procesamiento por lotes o de transmisión.
- **API (Application Programming Interface):** Interfaz de programación de aplicaciones. Es un conjunto de reglas y protocolos para crear e interactuar con aplicaciones de software. Define cómo dos aplicaciones se comunican entre sí mediante solicitudes y respuestas, facilitando la integración, innovación, ampliación y mantenimiento de sistemas.
- **JSON (JavaScript Object Notation):** Un formato de texto sencillo para el intercambio de datos. Es un subconjunto de la notación literal de objetos de JavaScript, considerado un formato independiente del lenguaje.
- **XML (Extensible Markup Language):** Un estándar para el intercambio de datos entre dos sistemas, legible tanto por humanos como por ordenadores. Es un formato de datos basado en texto que utiliza etiquetas que se pueden definir libremente.
- **CSV (Comma Separated Values):** Un archivo de texto que tiene un formato específico para guardar los datos en una tabla estructurada, donde los valores están separados por comas.
- **TSV (Tab Separated Values):** Similar a CSV, pero los valores dentro del archivo de texto están separados por tabulaciones en lugar de comas.
- **GREL (General Refine Expression Language) - OpenRefine:** Un lenguaje de programación utilizado en la herramienta OpenRefine que permite perfilar y limpiar conjuntos de datos. Facilita tareas avanzadas de depuración y manipulación de datos.
- **N-gram fingerprint:** Un algoritmo utilizado para indexar palabras o frases basándose en su pronunciación. Se usa para encontrar similitudes y duplicados en datos textuales, involucrando la normalización de caracteres, la obtención de n-gramas, la clasificación y la unión de los mismos.
- **Algoritmo Metaphone:** Un algoritmo fonético diseñado para indexar palabras por su pronunciación. Se utiliza comúnmente para nombres propios y otras palabras, asignando un código único a cada palabra basado en cómo suena.
- **Algoritmo de colisión de claves (Key Collision Algorithm):** En criptografía de bloques, se refiere a dos claves distintas que producen subclaves idénticas a través de la función de programación de claves, lo que significa que cualquier texto sin formato puede cifrarse en el mismo texto cifrado.

- **Wikidata:** Un proyecto alojado y mantenido por Wikimedia que tiene como objetivo crear una base de conocimiento libre sobre el mundo que pueda ser leída y editada tanto por humanos como por máquinas.
- **VIAF (Virtual International Authority File):** Un proyecto de bibliotecas nacionales operado por el Online Computer Library Center para unir los ficheros de autoridad virtual de esas bibliotecas.
- **GeoNames:** Una base de datos geográfica gratuita accesible a través de Internet, que contiene más de 25 millones de nombres geográficos y más de 11.8 millones de características únicas, incluyendo coordenadas, elevación, población y subdivisiones administrativas.
- **URI (Uniform Resource Identifier):** Una cadena de caracteres que identifica de forma general cualquier recurso web mediante un nombre, una ubicación o ambos. Los Localizadores Uniformes de Recursos (URL) y los Nombres Uniformes de Recursos (URN) son dos tipos de URI.
- **RDF Turtle:** Un formato para serializar RDF (Resource Description Framework). RDF representa la información mediante "tripletas" (sujeto, predicado, objeto), donde cada elemento puede expresarse como un URI.
- **N-Triples:** Un formato para almacenar y transmitir datos, que es una serialización en texto plano y en línea para grafos RDF, y un subconjunto de Turtle.

REFERENCIAS

1. Gobierno de Colombia. (2020). *Guía de estándares para datos abiertos*. Herramientas de Datos Abiertos. https://herramientas.datos.gov.co/sites/default/files/2020-11/A_guia_de_estandares_final_0.pdf
2. Gobierno de Colombia. (s.f.). *Guía de Datos Abiertos de Colombia*. Herramientas de Datos Abiertos. <https://herramientas.datos.gov.co/sites/default/files/Guia%20de%20Datos%20Abiertos%20de%20Colombia.pdf>
3. Función Pública. (s.f.). *ABC de datos abiertos*. Sistema de Información Empresarial. <https://www1.funcionpublica.gov.co/web/sie/abc-de-datos-abiertos>
4. Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2012). *Política de Datos Abiertos del Gobierno Nacional*. https://www.mintic.gov.co/portal/604/articles-3707_documento.pdf
5. Departamento Nacional de Planeación (DNP). (2021). *Conpes 3920: Estrategia Nacional de Datos Abiertos*. <https://colaboracion.dnp.gov.co/CDT/Conpes/Economicos/3920.pdf>
6. Congreso de la República de Colombia. (2014). *Ley 1712 de 2014 (Ley de Transparencia y Acceso a la Información Pública)*. Diario Oficial No. 49.169. <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=57581>

Elaboró	Revisó	Aprobó
Pierre Ángel Gómez Cortés Contratista – Oficina TIC	Ing. Eliany R. Montejo Carrascal Profesional especializado – Oficina TIC	Ing. Gustavo Carvajal Millán Jefe Oficina TIC
Julio 14 de 2025	Julio 14 de 2025	Julio 14 de 2025