Guía metodológica

Proyecto "Análisis, perfilado y depuración de los conjuntos de datos del portal de datos abiertos del Ayuntamiento de Madrid"

Subdirección General de Transparencia
C/ Montalbán, 1 – 6ª planta -28014 MADRID
Tlf. 915 132 390
sgtransparencia@madrid.es

INDICE

Control de versiones 5
1 CONTEXTO Y OBJETIVO DE LA GUÍA METODOLÓGICA 1
1.1 Qué son los datos abiertos 1
1.2 Principios de los datos abiertos 1
1.3 ¿Por qué es necesaria la calidad de los datos? 3
1.4 Estructura de un conjunto de datos abiertos del Ayuntamiento de Madrid 3
1.5 Beneficios para el propio Ayuntamiento de tener un Portal de Datos Abiertos. 6
2 ESTRUCTURA DE LA GUÍA METODOLÓGICA 7
3 MEJORES PRÁCTICAS PARA LA PUBLICACIÓN DE DATOS ABIERTOS 8
3.1 Metadatos 10
3.1.0 Criterio 1 - El nombre del conjunto de datos es correcto y claro 10
3.1.1 Criterio 2 - Existe descripción correcta para el dataset 10
3.1.2 Criterio 3 - El dataset tiene asignado un sector correctamente 12
3.1.3 Criterio 4 - Existe un conjunto de palabras clave correcto 13
3.1.4 Criterio 5 - Existe la fecha desde/hasta del dataset 14
3.1.5 Criterio 6 - Frecuencia de actualización correcta 15
3.1.6 Criterio 7 – Responsable del conjunto de datos 17
3.1.7 Criterio 8 - Existe un documento de estructura correctamente estructurado ... 18
3.1.8 Criterio 9 - Existe la licencia correcta del dataset 20
3.2 Formatos reutilizables de los ficheros 22
3.2.1 Criterio 1 – Fichero formalmente bien construido 22
3.2.2 Criterio 2 – No existen filas o columnas en blanco en el fichero 22
3.2.3 Criterio 3 - Formato correcto del encabezado y en minúsculas y guiones bajos 23
3.2.4 Criterio 4 - Evitar filas o columnas de totales o subtotales 25
3.2.5 Criterio 5 - Sólo un tipo de dato por columna 25
3.2.6 Criterio 6 - No incluir hojas vacías y preferentemente una hoja por fichero excel y preferentemente, una hoja por fichero excel 26
3.2.7 Criterio 7 - El fichero debe contener una única tabla de datos por hoja 27
3.2.8 Criterio 8 - Mismo orden y número de columnas en todas las filas, series temporales y formatos 28

3.2.9 Criterio 9 - El nombre del fichero de datos es consistente con la serie anterior.29

3.2.10 Criterio 10 - Uso de ";" como campo separador de caracteres en los ficheros CSV 30

3.2.11 Criterio 11 - El conjunto de datos está dividido en distintas distribuciones de manera que cada una de ellas sea suficientemente tratable con programas informáticos habituales 31

3.2.12 Criterio 12 – No existe demasiada anidación en los datos 32

3.2.13 Criterio 13 - Codificación correcta de caracteres 34

3.2.14 Criterio 14 – Organización vertical de la información, en vez de horizontal 35

3.2.15 Criterio 15 – Identificación en los datos, del año y mes a que hacen referencia 36

3.2.16 Criterio 16 – Mes mejor en formato numérico en vez de texto, para permitir una ordenación cronológica de los meses, en vez de alfabética 37

3.2.17 Criterio 17 - Utilizar el mayor número de formatos posibles, para que los datos sean más accesibles. 38

3.2.18 Criterio 18 – No existen metadatos de autor 40

3.3 Datos 41

3.3.1 Criterio 1 – Orden lógico de las columnas 42

3.3.2 Criterio 2 - Los tipos de campos se ajustan a lo esperado 43

3.3.3 Criterio 3 - Asignación de un ID único 44

3.3.4 Criterio 4 - Los valores de datos de tipo fecha y fecha/hora deben describirse en formato ISO 8601 45

3.3.5 Criterio 5 - Cumplimiento de codificación para información de barrios y distritos 46

3.3.6 Criterio 6 – Formato de dirección válida. 47

3.3.7 Criterio 8 - Los valores nulos y no nulos se ajustan a lo esperado 48

3.3.8 Criterio 7 - Las coordenadas latitud y longitud, correctamente representadas . 49

3.3.9 Criterio 9 – Las coordenadas X e Y, correctamente representadas 51

3.3.10 Criterio 10 - Decimales representados con coma en números. 52

3.3.11 Criterio 11 - No se deben utilizar caracteres de formato de "miles". 53

3.3.12 Criterio 12 - No se deben incluir ceros a la izquierda. 53

3.3.13 Criterio 13 - Las unidades de medida y monedas deben indicarse por separado, o en el nombre de las columnas. 54

3.3.14 Criterio 14 - Valores de distribución de cada columna coherentes con la serie anterior 55

3.3.15 Criterio 14 - Confidencialidad y anonimización de los datos 57

3.3.16 Criterio 15 – Dato único. Consistencia entre datos del portal de Madrid y fuentes externas (Madrid.es y Banco de Datos de Estadística) 58
4 GUÍAS EXTERNAS 60
4.1 Guías de la iniciativa Ciudades Abiertas 60
4.2 Guías de la iniciativa Aporta (Datos.Gob) 63
4.3 Normas UNE de la Gestión del Dato y de la Calidad del Dato 68
Anexo I: Documento de estructura 72
Anexo II: Sectores según la NTI 75
Anexo III: Tabla de distritos (versión abreviada) 76
Anexo IV: Listado de criterios o checklists a revisar 78
Anexo V: Rangos orientativos de las coordenadas X-Y para la Ciudad de Madrid 80
5 Anexo VI: Rangos posibles para Latitud y Longitud en la Ciudad de Madrid (ETRS89) 81
Anexo VII: Como eliminar metadatos de autor en los ficheros excels 82

Control de versiones

Versión	Descripción	Autor	Fecha de creación
1.1	Desarrollo de la Guía Metodológica FASES I y II	Accenture	18/09/2023
1.2	Desarrollo de la Guía Metodológica FASE IV (prórroga)	Accenture	09/05/2024
1.3	Revisión	S.G. Transparencia	09/05/2024
1.4	Revisión	S.G. Transparencia	23/05/2024
1.5	Revisión	Accenture	28/05/2024

1 CONTEXTO Y OBJETIVO DE LA GUÍA METODOLÓGICA

De cara a seguir mejorando el proceso de publicación de datos abiertos del portal del Ayuntamiento de Madrid Accenture, dentro del marco del proyecto Análisis, perfilado y depuración de los conjuntos de datos del portal de datos abiertos del Ayuntamiento de Madrid, y en colaboración con la Subdirección General de Transparencia del Ayuntamiento de Madrid, ha preparado una **Guía Metodológica** con las mejores prácticas relativas a dicho proceso de publicación.

El objetivo de esta Guía Metodológica será ayudar al Ayuntamiento de Madrid y sus fuentes de datos a mejorar sus controles y validaciones de **calidad**, **utilidad**, **transparencia** y **reutilización de sus datos.** La Guía incluirá el mapeo de los **criterios** más relevantes, las **mejores prácticas** asociadas a tales criterios y **ejemplos de estas prácticas**.

Adicionalmente a este documento, para facilitar el entendimiento de la Guía y promover la **adopción efectiva** de las prácticas identificadas por parte de las fuentes de datos del portal de Madrid, Accenture impartirá en mayo de 2024 **dos sesiones formativas online de 4 horas**

1.1 Qué son los datos abiertos

Los datos abiertos se refieren a la idea de que ciertos datos deben estar disponibles para que cualquier persona los use, modifique y comparta libremente. Estos datos son accesibles sin restricciones de copyright, patentes u otros mecanismos de control. En resumen, son datos que están disponibles para el público en general de forma gratuita. Además, destacan por estar al **máximo nivel de detalle** y su capacidad para proporcionar información precisa y completa que pueda ser utilizada de manera efectiva por una amplia variedad de usuarios y aplicaciones. Es importante destacar que no estamos hablando de datos estadísticos, sino datos a máximo nivel de detalle.

Los datos abiertos son utilizados en una amplia gama de aplicaciones y contextos. Los desarrolladores de software aprovechan estos datos para crear aplicaciones que abordan necesidades diversas, desde el transporte público hasta la salud pública. Los investigadores académicos utilizan datos abiertos en sus estudios en disciplinas tan variadas como las ciencias sociales y las ciencias de la computación. Asimismo, los datos abiertos se emplean en el análisis de políticas públicas y son una herramienta para el activismo y el periodismo de investigación.

1.2 Principios de los datos abiertos

Para poder hablar de Datos Abiertos y no únicamente de “publicar información” o incluso de información estadística, es necesario cumplir una serie de características. Para ello nos quedaremos con los 8 principios de los datos abiertos, desarrollados por treinta defensores del gobierno abierto en 2007, que establecen qué características se necesitan para considerarse Datos Abiertos:

1. **Completos:**
   Todos los datos son públicos. Es una tendencia en todas las administraciones de ir publicando todos los datos que no estén protegidos o limitados por temas de privacidad, seguridad o cualquier otra restricción legal. Al ser completo, implica también el publicar los datos de toda la Ciudad y no por ejemplo publicar de unas zonas si y otras zonas no. Con respecto a este tema, hay que tener en cuenta el tema competencial de la Ciudad y que, si hablamos de inventario de elementos, dependiendo de si está en vía pública, parques y jardines normales o parques y jardines históricos, la gestión es realizada por distintas unidades.

2. **Primarios:**
   Los datos se publican desde la fuente, con el más alto nivel posible de granularidad, no en formas agregadas o modificadas. En el momento que se agregan ya podría tratarse de estadísticas.
   Algunas veces por temas de protección de datos, seguridad o por otro motivo, no queda más remedio de proceder a la agregación.

3. **Oportunos**
   Los datos se actualizarán con la frecuencia necesaria para que tengan valor.

4. **Accesibles**
   Los datos están disponibles para la gama más amplia de usuarios y para la más amplia gama de propósitos. Por ello se tiene siempre que intentar ofrecer los datos en diferentes formatos, algunos más técnicos y otros más fácilmente a utilizar por cualquier ciudadano.

5. **Procesables por máquinas**
   Los datos se estructuran de forma razonable para permitir el procesamiento automatizado.

6. **No discriminatorios**
   Los datos están disponibles para cualquier persona, sin necesidad de registro.

7. **No propietarios**
   Los datos están disponibles en un formato sobre el cual ninguna entidad tiene el control exclusivo o hay que pagar un software específico para su empleo.

8. **Licencia libre**
   La licencia establecida mayoritariamente es aquella que habilita a utilizar los datos para cualquier propósito, con el único requisito de atribuir la fuente de los datos.

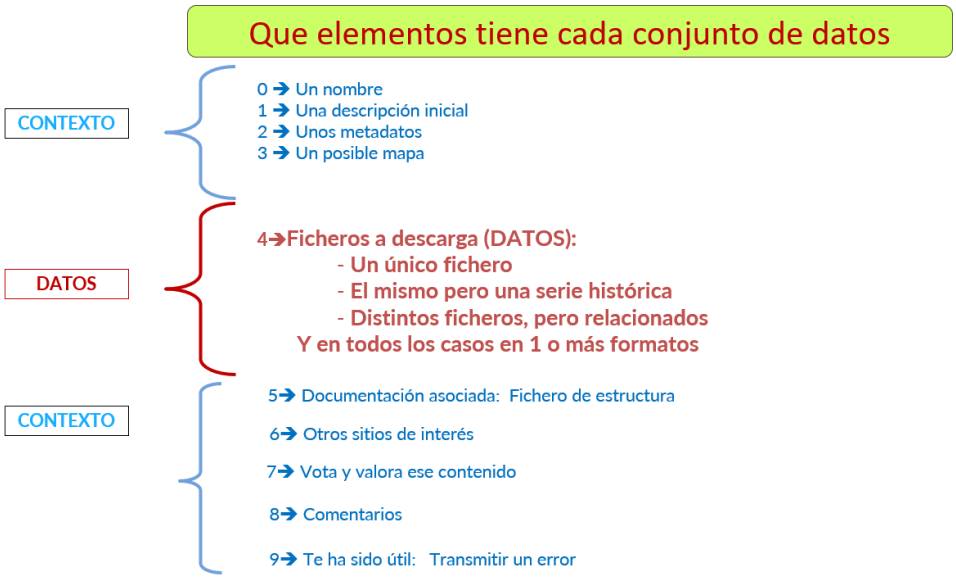
1.3 ¿Por qué es necesaria la calidad de los datos?

La utilidad y credibilidad de estos datos dependen en gran medida de su calidad. En primer lugar, los datos precisos y confiables son indispensables para la toma de decisiones informadas y la generación de análisis confiables. Además, los datos de alta calidad son relevantes y útiles para una amplia gama de usuarios y aplicaciones, ya que proporcionan información precisa y detallada que puede ser utilizada de manera efectiva para resolver problemas y generar conocimiento en diversos campos.

Para garantizar la calidad de los datos, es necesario considerar varios aspectos. La exactitud es fundamental; los datos deben reflejar con precisión la realidad que intentan representar. La completitud es otro aspecto importante; todos los datos relevantes deben estar presentes y disponibles, sin omisiones significativas. Además, la consistencia y la actualidad son aspectos cruciales; los datos deben ser coherentes en su estructura y contenido, y estar actualizados para reflejar la información más reciente disponible.

1.4 Estructura de un conjunto de datos abiertos del Ayuntamiento de Madrid

La estructura de los conjuntos de datos abiertos del Ayuntamiento de Madrid se caracteriza por su organización categorizada en diversas áreas temáticas, abarcando una amplia gama de aspectos relacionados con la ciudad y su funcionamiento. Cada conjunto de datos se presenta de manera clara y coherente, siguiendo estándares de codificación y etiquetado. Los conjuntos de datos tienen tres principales elementos:



- **Contexto:** Cada conjunto de datos tiene un nombre propio con una descripción, metadatos, y dependiendo del conjunto un posible mapa de geolocalización.

El nombre y la descripción aparecen primero en la estructura, y la descripción puede ser corta o extensa, y a continuación los otros descriptores. El siguiente ejemplo del fichero "Aforos de tráfico en la ciudad de Madrid permanentes" expone lo anterior.

Aforos de tráfico en la ciudad de Madrid permanentes

← Volver

Escuchar

Este conjunto de datos ofrece los datos de aforos de tráfico registrado en las 60 estaciones permanentes disponibles en la ciudad de Madrid.

En el apartado '**Descarga**s' están disponibles los recursos siguientes:

- Datos de aforos desde enero de 2018, con el número de vehículos que pasan por las estaciones permanentes de medición de tráfico
- Ubicación de las estaciones permanentes
- Ubicación de las estaciones permanentes con el sentido y orientación de la medición

En este portal también están disponibles otros conjuntos de datos relacionados con aforos de tráfico:

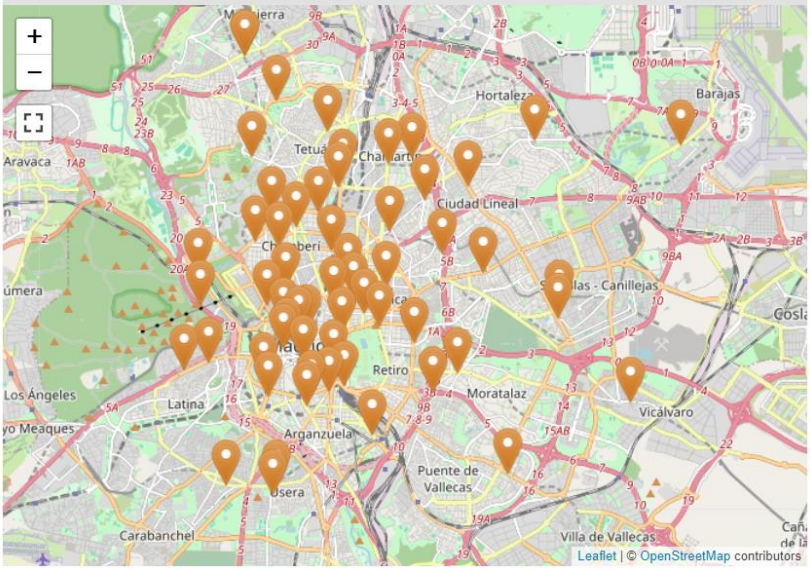
- Aforos de tráfico de la Ciudad de Madrid no permanentes
- Aforos de peatones y bicicletas
- Campañas de aforos de bici, motos y peatonales

IMPORTANTE: A partir del 14 de marzo de 2020 se decreta el estado de alarma por el COVID-19.

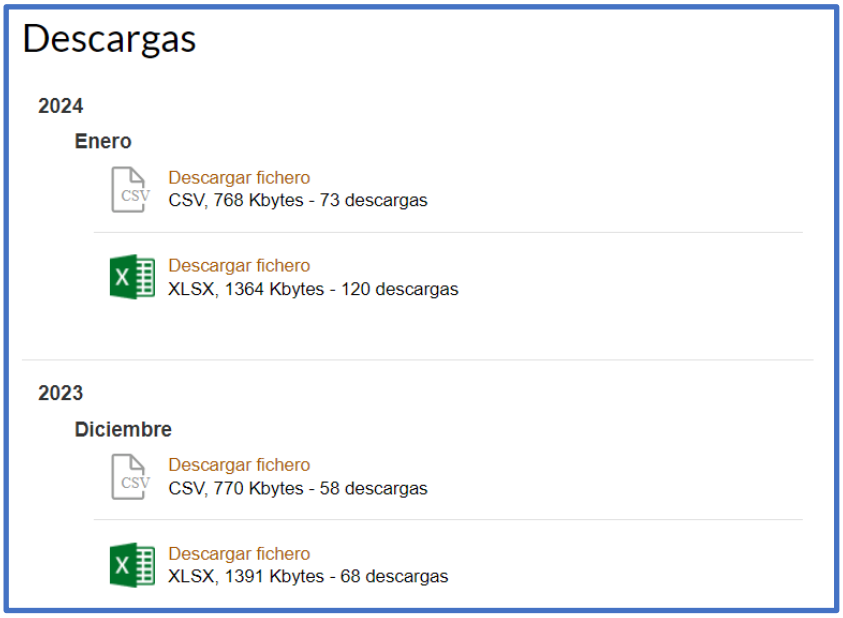
Definición del conjunto de datos +

Publicación de datos +

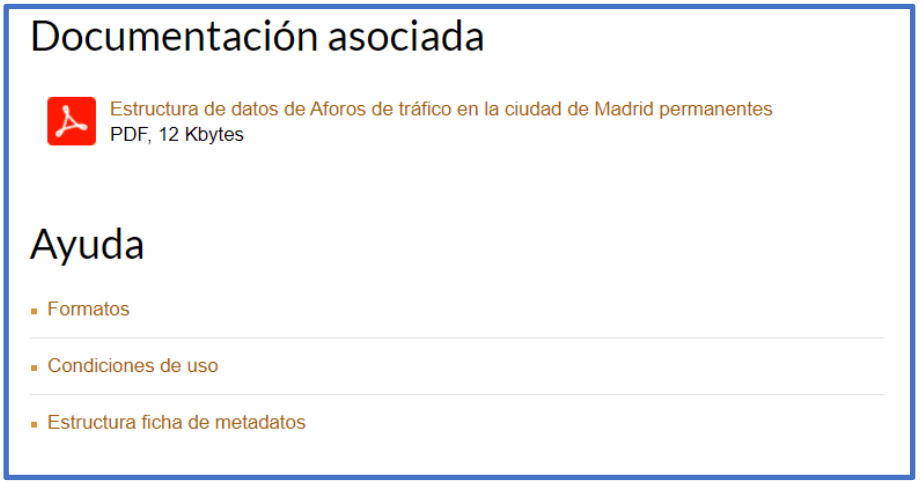
Ver resultados en el mapa −



- **Datos:** Cada conjunto de datos tiene una serie de ficheros que el número dependerá si está diseñado como serie histórica o son diferentes pero relacionados al conjunto de datos en cuestión. Los datos se presentan en una variedad de formatos, de los que se incluyen CSV (valores separados por comas), JSON (notación de objetos JavaScript) y XML (lenguaje de marcado extensible), entre otros. Estos ficheros pueden presentar uno o más formatos.

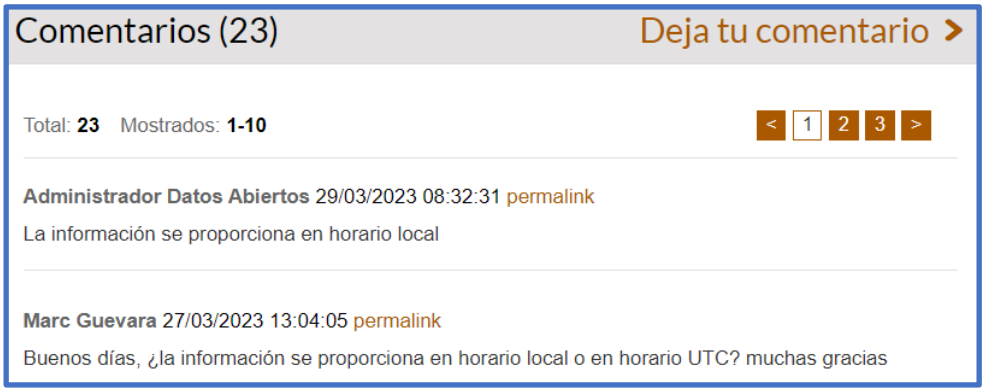


- **Contexto datos:** Los conjuntos de datos tienen información asociada como la ficha de estructura donde se detalla información de los campos de cada uno de los ficheros. Además, incluye otros recursos de interés como otros sitios de interés relacionados al conjunto, e interacciones por parte del usuario como comentarios o valoración del contenido. La documentación asociada aparece en formato de PDFs y links.



La votación del contenido es medida del uno al cinco, y aparece el número de votos totales.

Finalmente, aparece la parte de comentarios, donde los distinto usuarios pueden dejar por escrito preguntas al administrador o comentarios.



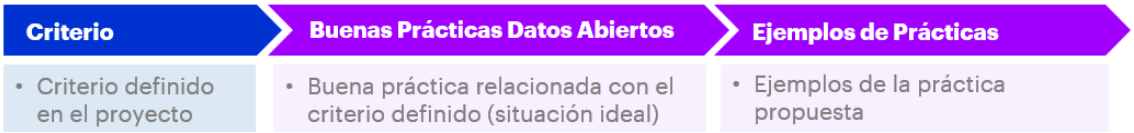
1.5 Beneficios para el propio Ayuntamiento de tener un Portal de Datos Abiertos.

2 ESTRUCTURA DE LA GUÍA METODOLÓGICA

La presentación de la Guía Metodológica se estructura en tres grandes apartados: **Metadatos**, **Formatos reutilizables** y **Datos.**

Para cada uno de los apartados se identifican los **principales criterios** para la publicación de datos de calidad, identificados a lo largo del proyecto Perfilado de Datos Abiertos y a partir de la experiencia de Accenture.

Para cada criterio se indicarán las **mejores prácticas** asociadas y **ejemplos.** Con el objetivo de que las fuentes de datos puedan realizar los controles de una forma más eficiente.



Adicionalmente, también se detallarán los **principales errores detectados en el Portal de Datos Abiertos del Ayuntamiento de Madrid** agrupados en los tres grandes apartados del documento (Metadatos, Formatos reutilizables y Datos).

Sobre estos **errores más comunes** se incluye una **descripción** del propio error, el **impacto** en el Portal de Datos Abiertos (en qué actividad del proyecto se ha detectado, cuantos datasets están

impactados…) y un **ejemplo**. Sumado a esto, se identifica la **buena práctica**, detallada en el bloque anterior, a la que estaría asociada el error.

Por último, además de la identificación de buenas prácticas relativas a criterios trabajados durante el proyecto y del detalle de los errores más comunes, la Guía Metodológica incluye un bloque relativo a otras guías externas de interés para las fuentes de datos del portal de Madrid.

En la formación, se presentará cada guía de forma resumida, destacando los puntos más relevantes de cada una para la publicación de datos abiertos, de modo que pueda ser profundizado por cada fuente, en caso de que sea de su interés.

Algunos ejemplos de guías externas son:

- **Guías de la iniciativa Ciudades Abiertas** (calidad de datos, anonimización de datos, checklists, etc.)
- **Guías de la iniciativa Datos.Gob** (calidad de datos, calidad de formatos)

3 MEJORES PRÁCTICAS PARA LA PUBLICACIÓN DE DATOS ABIERTOS

El concepto de datos abiertos u *open data*, es un concepto que persigue que determinados datos sean accesibles y estén disponibles, sin restricciones legales, financieras o tecnológicas. Para las administraciones públicas, datos abiertos son fuentes de importante valor, ya que fomentan la transparencia y participación ciudadana, la innovación y la mejora de servicios públicos.

A partir de 2010, se observó una rápida expansión de portales de datos abiertos alrededor del mundo, lo que ha implicado diferentes estándares y mejores prácticas alrededor de los mismos.

En la literatura sobre datos abiertos existen diferentes enfoques sobre cuáles son los atributos y características que los conjuntos de datos deben tener para ser considerados de calidad. Sin embargo, existen algunos principios que son comunes a casi todas las principales referencias en el tema.

Una de las iniciativas internacionales más importantes en este contexto es la **Carta Internacional de los Datos Abiertos** creada en 2015, que reúne más de 150 gobiernos y organizaciones, y que propone las mejore prácticas para publicación de datos gubernamentales. Los 6 principios que rigen carta las orientaciones de dicha la carta, son:

- **Abiertos por defecto**
- **Actualizados y comprensibles**
- **Accesibles y fáciles de usar**
- **Comparables e interoperables**
- **Para mejorar la gobernanza y participación ciudadana** y
- **Para la innovación y el desarrollo inclusivo**.

A nivel práctico, los principios se reflejan en atributos que deben ser verificados para evaluar la calidad de los conjuntos de datos. Estos atributos pueden variar de acuerdo con la institución responsable, nacionalidad y/o finalidad, pero Accenture utiliza y recomienda las definiciones encontradas en la **"Guía práctica para la mejora de la calidad de datos abiertos"** (documento de referencia a nivel nacional para España), conteniendo los siguientes 11 atributos:

1. **Exactitud/Precisión**: aunque no son términos equivalentes se refieren a la veracidad que proporcionan los datos - representando el valor verdadero del atributo. Además, las mediciones que son precisas son consistentes y replicables. Ej: dados de fuentes confiables y con procesos bien definidos.
2. **Completitud:** los datos se consideran completos cuando está disponible toda la información requerida para un atributo, con un nivel de detalle y una desagregación adecuado. Ej: campos no nulos y no duplicados.
3. **Consistencia/Coherencia**: los datos deben estar libres de contradicciones y tener coherencia lógica en un contexto específico, por ejemplo, de formato o temporal. Ej: outliers dentro de rangos razonables; media y medianas coherentes con la realidad.
4. **Credibilidad:** Los datos deben ser objetivos, deben estar publicados con los estándares estadísticos apropiados y las prácticas y políticas para su recogida y publicación deben ser transparentes. Ej: tratamientos estadísticos claros desde la fuente.
5. **Actualidad y actualización/Puntualidad:** los datos deben estar disponibles a tiempo y sin retrasos que afecten a su relevancia y se actualizarán regularmente, manteniendo así su valor. Ej: actualizaciones periódicas preferiblemente automatizadas.
6. **Accesibilidad:** los datos deben ser de fácil acceso y estar disponibles para la más amplia gama de usuarios y propósitos. Ej: datos descarga pública sin la necesidad de autenticación del usuario.
7. **Conformidad:** los datos se adhieren a estándares o normativas vigentes. Ej: NTI, ISO, etc.
8. **Confidencialidad:** los datos se deben publicar respetando la privacidad y seguridad de estos. Ej: datos sin informaciones de identificación personal *(no-pii)*, agregados o anonimizados.
9. **Eficiencia:** los datos tienen atributos que pueden ser procesados y proporcionados con unos recursos razonables. Ej: agrupación de años en un mismo fichero para reducir volumen de extracciones.
10. **Trazabilidad:** Los datos tienen atributos que proporcionan un histórico del camino de acceso auditado a los datos o cualquier otro cambio realizado sobre ellos desde la fuente. Ej: metadatos y documentos extra que expliquen el proceso de tratamiento en el caso de estadísticas.
11. **Comprensibilidad /Interpretabilidad:** los datos pueden ser interpretados y leídos por los usuarios. Ej: diccionario de datos claro que indique la definición y el formato de cada campo.

Aparte a estos principios y atributos, como ya se ha comentado en este documento, a partir de la experiencia de Accenture en el proyecto para el Ayuntamiento de Madrid, se identificaron y

se analizaron 36 criterios y 55 buenas prácticas que impactan directamente la calidad de los datos publicados en los portales de datos abiertos.

A la luz de dicha experiencia, se detallan a continuación todos los criterios analizados indicando, para cada uno de ellos, sus mejores prácticas asociadas y ejemplos prácticos con "pantallazos" para facilitar el entendimiento por parte del lector.

Nota importante: Los ejemplos que se plantean en la Guía no siempre existirán en el portal en el momento de lectura, bien porque el dataset ya no exista, o bien porque se haya subsanado el aspecto a mejorar que se indica.

3.1 Metadatos

Los metadatos desempeñan un papel fundamental en la comprensión y la utilidad de los datos abiertos. Son datos adicionales o "etiquetas" que se adjuntan a los datos para proporcionar contexto sobre los mismos, y permitir clasificarlos, catalogarlos, caracterizarlos y facilitar su búsqueda.

3.1.0 Criterio 1 - El nombre del conjunto de datos es correcto y claro

- Definición
  El nombre del conjunto de datos es el identificador que se asigna a un conjunto específico de información. Un nombre de conjunto de datos correcto y descriptivo facilita la identificación rápida y precisa de la información que contiene y la búsqueda del mismo.

- Mejores prácticas asociadas
  1. *Mantener los nombres de los conjuntos de datos lo más cortos y concisos posible, evitando excesiva longitud o ambigüedad.*

- Ejemplos
  1. *El dataset "Puntos Limpios de Proximidad" tiene un nombre corto y conciso*

- Riesgo de no usar un nombre correcto y claro
  1. *La ciudadanía puede tener problemas para identificar rápidamente el contenido del conjunto de datos.*
  2. *Nombres inadecuados pueden dificultar la búsqueda y localización del conjunto de datos en bases de datos y motores de búsqueda.*

3.1.1 Criterio 2 - Existe descripción correcta para el dataset

A continuación del nombre, figura la descripción del dataset o conjunto de datos. Es lo primero que se lee de un conjunto de datos y es fundamental que sea bastante clara y explicativa de los datos que se podrán obtener más abajo.

- Definición
  Una descripción correcta de un dataset es la explicación detallada y precisa sobre el contenido, el propósito y el contexto de los datos. Esta descripción debe permitir a los

usuarios entender rápidamente qué tipo de datos contiene, cómo fueron recopilados, cuál es su alcance y cómo pueden ser utilizados. En el Portal, la descripción se encuentra debajo del nombre del conjunto de datos. Se recomienda que esta descripción sea, al menos, del orden de 1000 caracteres.

- Mejores prácticas asociadas

1. *Proporcionar una descripción detallada del conjunto de datos que incluya su contenido, fuente y propósito, para que los usuarios puedan evaluar su relevancia y comprender su contenido y contexto.*

2. *Redactar la descripción del conjunto de datos en un lenguaje claro y accesible para el público en general.*

- Ejemplos

1. *El dataset "Aparcamientos municipales para residentes (PAR). Listas de espera" tiene una descripción detallada del conjunto de datos, en un lenguaje claro y accesible:*

Aparcamientos municipales para residentes (PAR). Listas de espera

Escuchar

Los ciudadanos que soliciten el derecho de uso en un aparcamiento donde no estén disponibles plazas libres, quedarán en una **lista de espera** con un número de orden riguroso, según la fecha de presentación de la solicitud. En virtud de este número de orden los residentes autorizados serán requeridos por el Ayuntamiento para acceder al uso de las plazas por correo certificado a su domicilio.

"Para cualquier consulta relacionada con su solicitud o aclaraciones sobre gestiones o trámites sobre las listas de espera de Aparcamientos municipales para residentes, debe dirigirse a la Subdirección General de Gestión de Aparcamientos. En el apartado "Comentarios " de este conjunto de datos no se podrá resolver este tipo de cuestiones. Disculpe las molestias"

En este conjunto de datos se relacionan los distintos aparcamientos para residentes con su lista de espera, si existiese.

La información disponible en el fichero de listas de espera (ver apartado *Descargas*) es la siguiente:

- **cod_aparcamiento**: Código único que identifica un aparcamiento para residentes (PAR)
- **aparcamiento**: Denominación del aparcamiento para residentes
- **num_orden_lista**: Marca la preferencia a los efectos de asignar plaza cuando se realicen ofrecimientos de plazas vacantes.
- **nombre_solicitante** Iniciales del nombre o denominación, de la persona en lista de espera.
- **documento_solicitante**: Número de documento identificativo de la persona en lista de espera. Este número se presenta adecuándose a las reglas que se indican en el documento de descripción de la estructura del fichero de datos.
- **fecha_solicitud**: Fecha de inclusión en el sistema de la solicitud.

Las **personas interesadas que desconozcan el número de orden** asociado a su solicitud pueden obtenerlo, previa acreditación de su identidad, a través de la Subdirección General de Gestión de Aparcamientos.

Las listas de espera están sujetas a variaciones debido a modificaciones sobrevenidas por causas administrativas o judiciales.

Las personas con movilidad reducida tienen preferencia en la lista de espera y por tanto pueden alterar el orden de esta.

En el Portal de Datos Abiertos también están disponibles otros conjuntos de datos relacionados con esta información:

- Aparcamientos municipales para residentes (PAR)
- Aparcamientos públicos municipales

A través del Portal de Visualizaciones "Visualiza Madrid con Datos Abiertos", el Ayuntamiento de Madrid pone a tu disposición una visualización realizada con datos abiertos de Aparcamientos.

- Riesgos de que no existe una descripción correcta

1. *Sin una descripción detallada, la ciudadanía puede tener dificultades para determinar si el conjunto de datos es relevante para sus necesidades.*
2. *La falta de una descripción clara puede generar dudas sobre la calidad y la fiabilidad del conjunto de datos.*

3. *Sin una descripción detallada, es difícil evaluar la calidad del conjunto de datos y planificar futuras actualizaciones.*

3.1.2 Criterio 3 - El dataset tiene asignado un sector correctamente

- Definición
  *La asignación de un sector a un dataset es la categorización o etiqueta que se asocia de acuerdo con el contexto o área temática relevante de los datos. Esto permite una organización eficiente de la información para su búsqueda, análisis y uso posterior. En España se sigue la normativa "**La Norma Técnica de Interoperabilidad de Reutilización de recursos de información"** (NTI ó Norma Técnica de Interoperabilidad), que es un conjunto de directrices y estándares técnicos que pretenden promover la interoperabilidad entre las diferentes administraciones públicas y facilitar el intercambio de información. Actualmente hay 22 sectores que utilizan todos los portales de datos abiertos, y no se puede categorizar de otra forma que no sea utilizando uno de los sectores o categorías existentes. La tabla con los sectores de la normativa se encuentra disponible en el Anexo II de este documento.*

- Mejores prácticas asociadas
  1. *Clasificar el conjunto de datos en un sector o categoría apropiada para facilitar su búsqueda y comprensión.*
  2. *Utilizar la taxonomía de la Norma Técnica de Interoperabilidad para la categorización en sectores, de modo que estén consistentes y bien definidos.*

- Ejemplos
  1. *El dataset "Contratos menores" tiene asignado correctamente su sector:*

     Sector
     Hacienda

  2. *El dataset "Placas conmemorativas de Madrid" utiliza la taxonomía de la NTI (pág. 24):*
     https://datos.gob.es/sites/default/files/20160726_guia_de_aplicacion_de_la_nti_reutilizacion_recursos_de_informacion_l.pdf

     Dataset "Placas conmemorativas de Madrid":

     Sector
     Cultura y ocio

Taxonomía de la NTI:

SECTOR	IDENTIFICADOR
Ciencia y tecnología *Incluye: Innovación, Investigación, I+D+i, Telecomunicaciones, Internet y Sociedad de la Información.*	ciencia-tecnologia
Comercio *Incluye: Consumo.*	comercio
Cultura y ocio *Incluye: Tiempo libre.*	cultura-ocio

Los responsables del portal de Datos Abiertos propondrán un sector correcto.

- Riesgos de no tener un sector correcto asignado

1. *Sin una descripción detallada, la ciudadanía puede tener dificultades para determinar si el conjunto de datos es relevante para sus necesidades.*
2. *La falta de una descripción clara puede generar dudas sobre la calidad y la fiabilidad del conjunto de datos.*

3.1.3 Criterio 4 - Existe un conjunto de palabras clave correcto

- Definición
  Conjunto de términos relevantes y descriptivos que se asocian con un conjunto de datos, y que se utilizan para etiquetar y categorizar los datos, con el objetivo de optimizar la organización, búsqueda y accesibilidad de los mismos. Las palabras clave son usadas por los buscadores para permitir encontrar los conjuntos de datos que responden a ciertas palabras buscadas, permitiendo identificar contenido relevante y relacionado con la temática de búsqueda.

- Mejores prácticas asociadas
  1. *Incluir palabras clave relevantes (keywords) que describan el contenido y el tema del conjunto de datos, para mejorar su descubrimiento y clasificación.*
  2. *Utilizar el símbolo "," como separador de palabras clave en lugar de otros caracteres para mantener la consistencia.*
  3. *Comprobar que las palabras clave no están encerradas entre comillas dobles.*

- Ejemplos

1. *En la siguiente imagen se puede observar que el dataset "Centros Municipales de Mayores. Datos sobre número y perfil de socios y socias" tiene un conjunto correctamente representado de palabras clave:*

> Palabras clave:
>
> Centros Municipales de Mayores, envejecimiento activo, servicios sociales, bienestar social, tarjeta madridmayor.es

- Riesgos de que no esté asignado un conjunto de palabras correcto

1. *Si las palabras clave no reflejan adecuadamente el contenido del conjunto de datos, este puede pasar desapercibido para usuarios que podrían beneficiarse de él.*
2. *Los usuarios pueden tener problemas para encontrar el conjunto de datos si las palabras clave no son relevantes o descriptivas.*
3. *Palabras clave inadecuadas dificultan la categorización y organización eficiente de los conjuntos de datos.*

3.1.4 Criterio 5 - Existe la fecha desde/hasta del dataset

- Definición

Las fechas "desde" y "hasta" en un dataset se refieren al período de tiempo en el que los datos fueron recopilados, registrados o son válidos. De esta forma es posible delimitar el contexto temporal de los datos y comprender cuándo son relevantes y aplicables.

*Lógicamente este **aplica únicamente** a conjuntos de datos que publican una serie temporal o cronológica. Este criterio no aplicaría a otros conjuntos de datos, como por ejemplo el de "Centros culturales", que presentan una única foto actual o situación actual de esos centros culturales de la Ciudad, por lo que no tiene lógica tener esos metadatos, y solamente tendría que tener un metadato de fecha de actualización.*

El conjunto de datos de datos "Tráfico. Datos del tráfico en tiempo real" que solo muestra la foto o situación del momento de la medición, la cual va variando cada pocos minutos, tampoco tiene sentido que tenga los metadatos fecha desde y fecha hasta. En este caso si tiene sentido tener igualmente el metadato fecha de actualización.

- Mejores prácticas asociadas
  1. *Incluir la fecha de inicio y fin de la cobertura temporal del conjunto de datos para que los usuarios comprendan el periodo cubierto.*

- Ejemplos

1. *El dataset "Calidad del aire. Datos diarios desde 2001" tiene fecha desde:*

   Datos obtenidos:
   desde 01/01/2001

   Sería más correcto poner también fecha hasta, aunque si no se pone se asume que es hasta la actualidad.
   *Como se puede ver en este conjunto de datos, Madrid pone a disposición los datos de Calidad del aire de los **últimos 23 años**.*

2. *El dataset "Accidentes de tráfico de la Ciudad de Madrid" tiene fechas desde y hasta:*

   Datos obtenidos:
   desde 01/01/2010 hasta 31/03/2024

   *Como se puede ver en este conjunto de datos, Madrid pone a disposición los datos de Accidentes de tráfico, de los **últimos 14 años**.*

3. *El dataset "Madrid Salud. Estadísticas centro de protección animal" tiene fecha desde y hasta:*

   Datos obtenidos:
   desde 01/01/2012 hasta 31/12/2022

- Riesgos de que no exista una fecha desde/hasta

1. *La ciudadanía puede no comprender el período de tiempo al que se refieren los datos.*
2. *Sin fechas desde/hasta, la ciudadanía puede tener dificultades para comparar y analizar los datos a lo largo del tiempo.*
3. *Sin una fecha de inicio y fin, la ciudadanía puede no saber si los datos están actualizados.*

3.1.5 Criterio 6 - Frecuencia de actualización correcta

- Definición

*La frecuencia de actualización es el intervalo de tiempo en el que se actualizan y revisan los datos almacenados en un sistema o base de datos. Es importante mantener una frecuencia de actualización adecuada **para garantizar el valor de los datos**, así como su calidad, precisión y relevancia de los datos, asegurando que la información refleje*

cambios en la realidad, y evitando que los usuarios se basen en datos obsoletos o desactualizados. Esto permitirá una correcta toma de decisiones.

Por ejemplo, la información de tráfico en tiempo real, hay que actualizarla cada 3 – 5 minutos para que tenga valor. Con la ocupación de los aparcamientos públicos de rotación ocurre lo mismo. Hay otros conjuntos de datos que con actualizarla cada hora es suficiente, como la de calidad del aire o datos meteorológicos. Y así, hay distintas frecuencias, diarias, semanal, mensual, trimestral, anual, etc. siempre tomando como criterio, que los datos tengan valor. Como último ejemplo, mostramos el calendario laboral se actualiza solamente una vez al año.

- Mejores prácticas asociadas

1. *Indicar claramente la frecuencia con la que se actualiza el conjunto de datos para que los usuarios estén informados sobre su vigencia.*
2. *Garantizar que el conjunto de datos se actualiza con la frecuencia de actualización indicada.*

- Ejemplos

1. *El dataset "Comercios centenarios" tiene frecuencia de actualización en el portal anual y los datos están actualizados, ya que la próxima fecha de actualización sería el 19/07/2025. Esto es un ejemplo de buena práctica:*

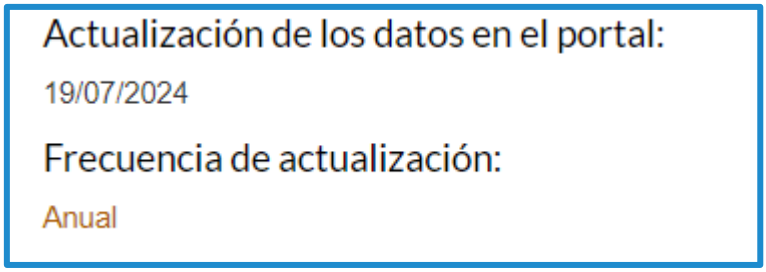


Imagen tomada a 20/07/2024

2. *El dataset "Mercados Municipales: locales comerciales, nombres comerciales y actividades" tiene frecuencia de actualización en el portal (anual). Sin embargo, los datos llevan desactualizados desde el 2020 (e incluso los datos corresponden al 2018). Esto es un ejemplo de práctica a mejorar, ya que los usuarios no pueden acceder a la versión más reciente de los datos.*

Actualización de los datos en el portal:

01/07/2020

Frecuencia de actualización:

Anual

Imagen tomada a 20/07/2024

(Nota, este dataset está en proceso de actualización por la unidad responsable)

- Riesgos de que no exista una actualización frecuente

1. *Si los datos no se actualizan con la frecuencia adecuada, pueden volverse obsoletos y no reflejar la realidad actual.*
2. *La falta de actualización periódica puede generar desconfianza en la calidad y fiabilidad de los datos.*
3. *Las aplicaciones y servicios que dependen de datos desactualizados pueden ofrecer información incorrecta o inútil a los usuarios.*

3.1.6 Criterio 7 – Responsable del conjunto de datos

- Definición

  Identificación clara y precisa de unidad del Ayuntamiento a nivel de Dirección General, Gerencia o similar responsable de la competencia de los datos y de la creación, mantenimiento y distribución del conjunto de datos. Esta información es para establecer la responsabilidad y la autoridad sobre el conjunto de datos, así como para proporcionar un punto de contacto para consultas, comentarios y colaboraciones relacionadas con los datos.

  Los datos publicados en el Portal de Datos Abiertos y su actualización y resolución de posibles consultas ciudadanas son competencia y responsabilidad de la unidad competente de los datos.

- Mejores prácticas asociadas
  1. *Especificar claramente el nombre de la unidad responsable en el Ayuntamiento de esos datos.*
  2. *Mantener la información del responsable del conjunto de datos actualizada para reflejar cualquier cambio en la responsabilidad o en la información de contacto.*

- Ejemplos
  1. *El dataset "Accesibilidad y movilidad en aceras y calzadas. Año 2020" tiene definida la entidad responsable:*

Responsable del conjunto de datos:

Dirección General de Planificación Estratégica

- Riesgos de que no exista un responsable del conjunto de datos

1. *La ausencia de identificación del responsable puede generar confusión sobre quién es responsable de los datos y su mantenimiento.*
2. *Sin un punto de contacto claro, los usuarios pueden tener dificultades para comunicarse con la entidad responsable.*
3. *La falta de identificación del responsable puede percibirse como falta de transparencia y rendición de cuentas.*

3.1.7 Criterio 8 - Existe un documento de estructura correctamente estructurado

- Definición

  Se define documento de estructura como aquel que recoge una descripción de la presentación de la información y de cada una de las columnas que constituyen el fichero de datos, la cual debe ser organizada y coherente, y sigue un formato predefinido que facilita la comprensión, búsqueda y análisis de los datos.

  Aunque no suele variar, tener correctamente actualizada la estructura de este documento es esencial para garantizar la accesibilidad, usabilidad de los datos de manera fluida, así como su consistencia e integridad ***y evitar errores de interpretación en los reutilizadores****. Un documento de estructura debe constar de:*

  - *Nombre del conjunto de datos.*
  - *Descripción: En este conjunto de datos…*
  - *Unidad responsable.*
  - *Frecuencia de actualización.*
  - *Disponibilidad de los datos: (hay que indicar la fecha estimada de actualización)*
  - *Forma de obtener esa información o filtros aplicados: (si fuese necesario especificarlo)*
  - *Estructura del fichero de datos (campos):*
    - *Número de columna*
    - *Nombre columna*
    - *Descripción*
    - *Valores esperados*

  Esta información se encuentra detallada en el Anexo I.

- Mejores prácticas asociadas

1. *Proporcionar documentación clara y completa que explique el significado y la interpretación de los códigos utilizados en los datos. Incluir recursos adicionales, como diccionarios de códigos o glosarios, para facilitar la comprensión de los códigos utilizados en los datos.*
2. *Proporcionar un documento detallado que explique la estructura y el esquema del conjunto de datos, incluyendo la definición de las variables y sus tipos.*
3. *Nombrar el documento de estructura de manera coherente y siguiendo un formato estándar para facilitar su identificación y acceso.*
4. *Verificar que las columnas en los conjuntos de datos coincidan con el archivo de estructura proporcionado. Esto asegurará que la información esté correctamente ubicada y facilitará la comprensión de los datos.*

- Ejemplos
  1. *En el documento de estructura "EstructuraDatos_AgenciaEmpleo_Perfiles.pdf" del dataset "*Agencia para el Empleo. Perfiles de personas inscritas*" todos los códigos están descritos y son interpretables:*

Versión mayo/2024

ESTRUCTURA DEL CONJUNTO DE DATOS

Nombre del conjunto de datos: ATLAS_DF_DS_DATOSABIERTOS_INSCRITOS_AEM

Descripción: Información de las personas inscritas en la Agencia para Empleo de Madrid por fecha (mes / año) a partir de enero 2017: Por edad, nacionalidad (español, comunitario y no comunitario), ocupación y distritos

Unidad responsable: Agencia para el Empleo de Madrid

Frecuencia de actualización: Mensual

Disponibilidad de los datos: 05 cada mes

Forma de obtener esa información o filtros aplicados: Muestra las inscripciones a partir del año 2017 en adelante. Datos ya filtrados, no necesaria ninguna acción extra.

Estructura del fichero de datos:

  2. *El dataset anterior tiene esta información recogida en un documento de estructura que está en el portal:*

Documentación asociada

Estructura de datos. Perfiles inscritos en Agencia para el Empleo
PDF, 184 Kbytes

  3. *El documento del dataset del ejemplo anterior tiene un nombre coherente: "EstructuraDatos_AgenciaEmpleo_Perfiles.pdf". Otro ejemplo puede ser, el dataset "Infraestructuras municipales para el fomento del emprendimiento", con*

documento de estructura "Estructura_ConjuntoDatos_Infraestructuras-dirigidas.a emprendedores-y-pymes.MadridEmprende.pdf".

4. *En el dataset "Relación de situados aislados en vía pública," en el fichero, las columnas del fichero "Relación_Situados_2024.xlsx" coinciden con el fichero de estructura:*

   Fichero de estructura:

Se establecen los siguientes campos que a continuación serán detallados:

- AÑO
- FECHA PUBLIC. AUTORIZ.
- NOMBRE-VIA
- CLASE-VIAL
- TIPO-NUM
- NUM
- COD-DISTRITO
- DISTRITO
- COORDENADA – X
- COORDENADA – Y
- LATITUD
- LONGITUD
- ENCLAVE
- CARÁCTER
- TIPO
- MODALIDAD
- SITUACIÓN
- OBSERVACIONES

   Fichero "Relación_Situados_2024.xlsx"

- Riesgos de no tener una estructura de datos correcta

1. *La falta de una estructura definida puede llevar a la inconsistencia en la presentación de la información entre diferentes conjuntos de datos.*
2. *Sin un documento de estructura claro y coherente, los usuarios pueden tener dificultades para entender la organización y el significado de los datos.*

3.1.8 Criterio 9 - Existe la licencia correcta del dataset

- Definición

  Las licencias asociadas a los conjuntos de datos establecen de manera clara que puede hacer con los datos un reutilizador. Dando garantía así al reutilizador.

Establecer licencias correctas en el manejo de datos, establece términos legales y condiciones claras que regulen cómo los datos pueden ser utilizados, compartidos y redistribuidos, proporcionando un marco legal y ético que protege los derechos de los creadores, usuarios y propietarios de los datos. De este modo se evitan conflictos legales mientras que se fomenta la confianza entre las partes involucradas.

La licencia normal empleada asociada al portal de Datos Abiertos de Madrid indica que se puede hacer cualquier uso de los datos, pero indicando en el producto resultante que se han empleado datos del Ayuntamiento.

- Mejores prácticas asociadas
  1. *Utilizar una licencia estandarizada y reconocible que sea consistente con otras licencias comúnmente utilizadas en conjuntos de datos similares, y especificarla claramente para que los usuarios conozcan los términos y condiciones de uso.*

- Ejemplos
  1. *El dataset "Pagos a terceros" tiene una licencia estandarizada:*

     Licencia:

     https://datos.madrid.es/egob/catalogo/aviso-legal

     Entre los tipos de licencias estandarizadas, se encuentran las Licencias Creative Commons (CC), que permiten a los creadores de obras, incluidos datos, especificar cómo otras personas pueden usar su trabajo. Otros ejemplos son las licencias de Datos Abiertos del Open Data Commons (ODC), diseñadas específicamente para datos abiertos, o las GNU General Public License (GPL, que incluyen también las versiones modificadas, entre otras.
     La licencia más común es la CC-BY o CC-Reconocimiento o CC-Atribución.

     La elección de una licencia u otra depende de la naturaleza de los datos y de los objetivos del creador en términos de uso y distribución.

Los responsables del portal de Datos Abiertos propondrán la licencia habitual.

- Riesgos de no tener una licencia correcta

1. *La falta de una licencia clara puede dejar a los reutilizadores sin saber cómo pueden utilizar los datos de manera legal.*
2. *Sin una licencia adecuada, los reutilizadores pueden incurrir involuntariamente en violaciones de derechos de autor.*
3. *Una licencia restrictiva puede limitar la reutilización y redistribución de los datos, reduciendo su utilidad y valor potencial.*

3.2 Formatos reutilizables de los ficheros

Este bloque de criterios hace mención a como tiene que estar construido el fichero de datos o ficheros de datos, para garantizar su acceso, uso y reutilización.

3.2.1 Criterio 1 – Fichero formalmente bien construido

- Definición
  Se dice que un fichero está formalmente bien construido cuando cumple con estándares y convenciones de estructura, formato y organización. Es importante, ya que garantiza una presentación uniforme y coherente de los datos, facilitando su comprensión.

- Mejores prácticas asociadas
  1. *Evitar dejar columnas sin nombres o por ejemplo hacer CSVs en los que la información de una misma fila no esté partida en varias celdas y así conseguir mejores prácticas de estructuración y organización de los datos para garantizar su calidad y facilidad de uso.*

- Ejemplos
  1. *El fichero "inventario_instalaciones_fotovoltaicas_2022.xlsx" del dataset "Inventario instalaciones fotovoltaicas" está bien construido:*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Número	Código Ayto	Centro	Dirección	Coordenada-X	Coordenada-Y	Latitud	Longitud	Superficie (m2)	Uso del edificio	Adscripción	Distrito	Potencia KW	Empresa instaladora	Puesta en servicio
2	1	2178	BIBLIOTECA ÁNGEL GONZALEZ	C/ GRANJA DE TORREHERMOSA, 2	434954	4472113	40.39707462511590 0	-3.7664885474841500		cultural	AG Cultura Turismo y Deportes	Latina	20	ORTIZ	7/31/2019
3	2	782	BIBLIOTECA IVÁN DE VARGAS	C/ DOCTOR LETAMENDI, 1	439775	4473962	40.41410290844660	-3.7098591212735800		cultural	AG Cultura Turismo y Deportes	Centro	3	ORTIZ	7/24/2019
4	3	2096	BIBLIOTECA EUGENIO TRIAS (CASA FIERAS)	AV/ MENENDEZ PELAYO , 10-3 BJ	442345	4474221	40.4166091081099	-3.6795930368034800		cultural	AG Cultura Turismo y Deportes	Retiro	8.2	ORTIZ	7/24/2019
5	4	1046	PLANETARIO	AV/ PLANETARIO 16	441830	4471375	40.39094408181480 0	-3.6853911554398 90	3062	cultural	AG Cultura Turismo y Deportes	Arganzuela	25	FULTON	9/13/2019

- Riesgos de que el fichero no esté formalmente construido

1. *Un fichero mal estructurado puede dificultar la comprensión de los datos por parte de los usuarios.*
2. *Una estructura inadecuada puede llevar a errores en el análisis de datos.*
3. *Un fichero mal construido puede limitar la reutilización de los datos en diferentes contextos.*

3.2.2 Criterio 2 – No existen filas o columnas en blanco en el fichero

- Definición

Práctica de eliminar cualquier fila o columna que no contenga información significativa en un conjunto de datos. Esto garantiza que los datos proporcionados estén completos y no contengan espacios vacíos que puedan afectar su integridad, precisión y utilidad para los usuarios.
Si existiesen filas o columnas en blanco, dificultaría su ordenación y manipulación e incluso ante operaciones de ordenación o totales y subtotales, podría generar resultados erróneos.

- Mejores prácticas asociadas

1. *Realizar una revisión exhaustiva del conjunto de datos para identificar y eliminar cualquier fila o columna que esté completamente en blanco o que contenga solo información no relevante o redundante.*
2. *Utilizar herramientas automatizadas de limpieza y procesamiento de datos que puedan identificar y eliminar de manera eficiente filas o columnas en blanco, especialmente en conjuntos de datos grandes o complejos.*
3. *Realizar pruebas de calidad adicionales después de la limpieza de datos para garantizar que el conjunto de datos final cumpla con los estándares de calidad establecidos y sea útil y confiable para los usuarios.*

- Ejemplos

1. *No se han encontrado ejemplos.*

- Riesgos de existir columna/filas en blanco

1. *Filas o columnas en blanco pueden afectar la integridad y precisión de los datos, ya que pueden interferir con cálculos y análisis posteriores.*
2. *Los datos incompletos pueden dificultar la identificación de patrones y tendencias relevantes.*
3. *La necesidad de identificar y eliminar filas o columnas en blanco puede aumentar la carga de trabajo y reducir la eficiencia operativa.*

3.2.3 Criterio 3 - Formato correcto del encabezado y en minúsculas y guiones bajos

- Definición

El encabezado de un conjunto de datos debe ser la primera fila que identifica y describe las columnas o campos de un conjunto de datos. Un formato correcto del encabezado facilita la comprensión del conjunto de datos, y, en consecuencia, su análisis y manipulación.

Los nombres de las columnas deben estar en una sola línea y justificados los nombres siempre a la izquierda. No usar acentos ni caracteres auxiliares
Si el nombre del campo tiene más de una palabra se separan con guion bajo:
tipo_vial

nombre_vial
dirección_auxiliar

- Mejores prácticas asociadas
  1. *Hacer que el encabezado de las columnas esté formateado correctamente (evitar anidación), sea coherente con el resto de información y se ajuste a las expectativas y convenciones establecidas.*
  2. *Verificar que la primera fila del fichero sea utilizada exclusivamente para los encabezados de las columnas o que no esté en la segunda fila. Evitar la presencia de datos o cualquier otra información en la primera fila que no esté relacionada con los encabezados de las columnas*
  3. *Utilizar nombres de columnas descriptivos y claros que sean comprensibles para los usuarios. Evitar el uso de abreviaturas o códigos crípticos en los nombres de las columnas y optar por terminología clara y concisa. Asegurarse de que los nombres de las columnas sigan un formato de texto universal que sea compatible con diferentes sistemas y aplicaciones.*

- Ejemplos
  1. *El fichero "Abonados_2022.xlsx" del dataset "Deportes. Abonados en Centros Deportivos Municipales" cumple con las tres buenas prácticas mencionadas:*

	A	B	C	D	E	F
1	Nº de abonados	Sexo	Edad	Tipo de abono	Centro deportivo	Mes
2	1	MUJER	18	ADM ACTIVIDAD DIRIGIDA	Alfredo Goyeneche	Jan-22
3	1	MUJER	38	ADM ACTIVIDAD DIRIGIDA	Alfredo Goyeneche	Jan-22
4	2	MUJER	50	ADM ACTIVIDAD DIRIGIDA	Alfredo Goyeneche	Jan-22

  2. *El fichero "Órdenes de ejecución 2020.xls" del dataset "Ordenes de ejecución para el cumplimiento del deber de conservación y rehabilitación exigidas a los propietarios" tiene encabezados anidados. Es un ejemplo de práctica a mejorar:*

DIRECCIÓN GENERAL DE LA EDIFICACIÓN			
SUBDIRECCIÓN GENERAL DE CONTROL DE LA EDIFICACIÓN			
Órdenes de ejecución dictadas por el Ayuntamiento para el cumplimiento del deber de conservación y rehabilitación exigidas a los propietarios			
DURANTE EL PRIMER TRIMESTRE DE 2020			
NUMERO DE ORDEN	**Emplazamiento principal**	**MOTIVO**	**MES**
1	PEDRO VALDIVIA, 36	ORDEN DE EJECUCIÓN	ENERO
2	RICARDO ORTIZ, 110	ORDEN DE EJECUCIÓN	ENERO
3	SAN LUIS, 144	ORDEN DE EJECUCIÓN	ENERO
4	TRITON, 10	ORDEN DE EJECUCIÓN	ENERO
5	AVDA. ORCASUR, 47	ORDEN DE EJECUCIÓN	ENERO

- Riesgos de que no exista un encabezado correcto

1. *Un encabezado mal formateado puede dificultar la identificación y comprensión de las columnas en el conjunto de datos.*
2. *Un encabezado mal formateado puede provocar errores en la manipulación y procesamiento de datos.*
3. *Un formato de encabezado inadecuado puede ser incompatible con algunas herramientas o sistemas de análisis de datos.*

4. *Nombres de columnas poco descriptivos o que no siguen convenciones comunes pueden no cumplir con las expectativas de los usuarios.*

3.2.4 Criterio 4 - Evitar filas o columnas de totales o subtotales

- Definición
  La presencia de filas o columnas de totales o subtotales puede dar lugar a errores y actualizaciones incorrectas. Dependiendo del caso será conveniente o no que estén incluidas en el conjunto de datos, aunque en términos generales, es preferible calcularlas de manera independiente para mayor exactitud.

- Mejores prácticas asociadas
  1. *Identificar si existen filas o columnas que contengan totales o subtotales en el conjunto de datos. Documentar claramente estas filas o columnas y considerar si es necesario incluirlas o excluirlas en el análisis posterior. (en principio no debería haber columnas totalizadoras).*

- Ejemplos
  1. *El fichero "Agua_Regenerada_Madrid_2022.xlsx" del dataset "Volumen de agua regenerada" tiene columnas totalizadoras, cuando, en principio, no debería tener. Es un ejemplo de práctica a mejorar:*

ERAR	Jan-01	Jan-02	Feb-01	Feb-02	Mar-01	Mar-02	Apr-01	Apr-02
VIVEROS	4,032.00	6,723.00	26,607.00	32,421.00	15,945.00	7,558.00	53,913.00	28,745.00
LA CHINA	107,629.72	106,956.00	75,825.00	38,443.00	33,956.00	36,672.00	40,870.72	45,944.44
LA GAVIA	9,954.00	12,329.00	19,667.00	15,802.00	23,611.00	20,015.00	24,052.00	21,356.00
LAS REJAS	24,079.00	29,213.00	57,184.00	64,595.00	44,740.00	40,212.00	50,180.00	51,649.00
TOTAL	145,694.72	155,221.00	179,283.00	151,261.00	118,252.00	104,457.00	169,015.72	147,694.44
		300,915.72		330,544.00		222,709.00		316,710.16

- Riesgos de tener filas/columnas con totales o subtotales

1. *La presencia de filas o columnas de totales o subtotales en el conjunto de datos puede llevar a errores en los cálculos y análisis posteriores.*
2. *La inclusión de filas o columnas de totales o subtotales puede causar confusión sobre el nivel de granularidad de los datos.*
3. *Los totales o subtotales incluidos en el conjunto de datos pueden dificultar la identificación de errores o discrepancias en los datos individuales.*

3.2.5 Criterio 5 - Sólo un tipo de dato por columna

- Definición
  Mantener un solo tipo de dato por columna (entero, decimal, fecha, texto… etc.), asegura la integridad de los datos y evita errores de cálculo y problemas en aplicaciones.

Es esencial para un correcto análisis y posterior toma de decisiones basadas en cálculos correctamente realizados.

- Mejores prácticas asociadas
  1. *Verificar que los tipos de datos utilizados en los campos sean apropiados y se ajusten a la naturaleza de la información. Esto evitará problemas de interpretación y permitirá un análisis preciso.*

- Ejemplos
  1. *Como ejemplo de práctica a mejorar, si se observa el fichero "MasasParquesHistoricoSingularesForestales_2020.csv" del dataset "Superficie de parques y zonas verdes de Madrid", tiene distintos formatos en la columna "Unidades 2020". Deberían ser todos enteros o todos decimales, según el contexto de los datos.*

PARQUE	ESPECIE PREDOMINANTE	Unidades 2020	SUPERFICIE OCUPADA (m2)	Superficie (ha)	Superficie TOTAL Parque (ha)
PARQUE LINEAL DEL MANZANARES	Pinus halepensis	1726			
PARQUE LINEAL DEL MANZANARES	Pinus pinea	571			
PARQUE LINEAL DEL MANZANARES	Otros	38			
PARQUE LINEAL DEL MANZANARES		**2,335**	**334,491.63**	**33.45**	88.22
PARQUE FORESTAL DE VALDEBEBAS- FELIPE VI	Pinus pinea	**53,241**	**2,900,300.05** (Comprende las zonas en conservación y con bajo mantenimiento)	**290.03**	387.55 Superficie total ejecutada del parque
PARQUE MADRID RÍO	Pinus halepensis	**2,881**	**92,970.17**	**9.30**	105.71

Un caso típico de varios tipos de datos en una columna suele ser en dirección. La dirección debe ir separada conforme a lo que se indica en el documento de estructura: tipo de vial, el nombre de la vía y el número de edificio. Otro caso de información que debiera ir separada es la fecha y la hora. Y por último tampoco es correcto incorporar en la misma columna el código del distrito y el nombre, es mejor que vengan en dos columnas separadas.

- Riesgos de tener diferentes formatos de datos

1. *La presencia de múltiples tipos de datos en una columna puede conducir a inconsistencias en los datos.*
2. *La mezcla de tipos de datos en una columna puede resultar en errores durante cálculos y operaciones.*
3. *Tipos de datos inconsistentes pueden aumentar la complejidad de la manipulación y limpieza de datos.*
4. *La presencia de múltiples tipos de datos puede no cumplir con los estándares de calidad y coherencia de datos.*

3.2.6 Criterio 6 - No incluir hojas vacías y preferentemente una hoja por fichero excel y preferentemente, una hoja por fichero excel

- Definición
  En ocasiones los conjuntos de datos vienen en formatos donde es posible que exista más de una hoja con contenido. Es esencial evitar la inclusión de hojas vacías y sin información útil en los formatos excel, ya que los csv solo pueden tener una hoja. También se debe evitar la existencia de hojas adicionales, procurando que toda la información esté en la misma tabla (punto 3.2.7).

- Mejores prácticas asociadas
  1. *Eliminar cualquier hoja vacía o sin datos en el conjunto de datos. Asegurarse de que todas las hojas contengan información relevante y útil.*

- Ejemplos
  1. *El fichero "Edificios declarados en ruina legal y física 2021.xls" del dataset "Relación edificios declarados en ruina" tiene una hoja en blanco, cuando no debería tener ninguna. Es un ejemplo de práctica a mejorar:*

- Riesgos de tener hojas vacías

1. *La presencia de hojas vacías o no relevantes puede dificultar la identificación de los datos importantes.*
2. *Hojas vacías aumentan el tamaño del archivo, lo que puede causar problemas de almacenamiento y aumentar el riesgo de errores.*
3. *La presencia de múltiples hojas puede complicar la exportación e importación de datos entre diferentes sistemas y plataformas.*
4. *Podrían incorporarse por error datos no publicables en las hojas adicionales como, por ejemplo, algún dato personal.*

3.2.7 Criterio 7 - El fichero debe contener una única tabla de datos por hoja

- Definición
  La existencia de más de una tabla de datos en el mismo fichero puede dar lugar a confusión, e incluso, desorden, de un conjunto de datos. Es fundamental mantener una tabla única por conjunto de datos, garantizando al usuario la calidad y consistencia de los datos. Cabe destacar que los ficheros csv sólo pueden contener una tabla.

- Mejores prácticas asociadas

1. *Asegurarse de que el fichero de datos contenga solo una tabla con la información relevante. Evitar la presencia de múltiples tablas o conjuntos de datos en el mismo fichero, ya que esto puede generar confusión y dificultar el análisis posterior.*

- Ejemplos

1. *Como ejemplo de práctica a mejorar, el fichero "Órdenes de ejecución 2020.xls" del dataset "Ordenes de ejecución" tiene más de una tabla. Sólo debería tener una:*

DURANTE EL PRIMER TRIMESTRE DE 2020			
NUMERO DE ORDEN	Emplazamiento principal	MOTIVO	MES
1	PEDRO VALDIVIA, 36	ORDEN DE EJECUCIÓN	ENERO
2	RICARDO ORTIZ, 110	ORDEN DE EJECUCIÓN	ENERO
3	SAN LUIS, 144	ORDEN DE EJECUCIÓN	ENERO
4	TRITON, 10	ORDEN DE EJECUCIÓN	ENERO
5	AVDA. ORCASUR, 47	ORDEN DE EJECUCIÓN	ENERO
6	FÚCAR, 10	ORDEN DE EJECUCIÓN	ENERO
7	JESÚS DEL VALLE, 15	ORDEN DE EJECUCIÓN	ENERO
DURANTE EL SEGUNDO TRIMESTRE DE 2020			
NUMERO DE ORDEN	Emplazamiento principal	MOTIVO	MES
1	C/ CARLOS II, 14	EJECUTAR OBRAS	ABRIL
2	C/ NAPOLES, 32	EJECUTAR OBRAS	ABRIL
3	C/ PABLO LAFARGUE,16	EJECUTAR OBRAS	ABRIL
4	C/ PICO CEBOLLERA, 31	EJECUTAR OBRAS	MAYO
5	FE, 6	ORDEN DE EJECUCION	MAYO
6	JESUS Y MARIA, 6	ORDEN DE EJECUCION	MAYO
7	ALFONSO XII, 58	ORDEN DE EJECUCION	MAYO
8	Algodonales, 3	xpediente Distrito (Actuación Servicio Municipal, Bombero:	MAYO
9	Ancora, 28	Actuación Servicio Municipal (Bomberos)	MAYO

2. *Otro ejemplo de práctica a mejorar es el fichero "tenis_mesa_2020.xlsx" del dataset "Tenis mesa aire libre" tiene dos hojas de Excel, es decir, dos tablas, cuando debería tener una.*

- Riesgos de tener diferentes tablas de datos

1. *La presencia de múltiples tablas puede causar confusión en la interpretación de los datos.*
2. *Múltiples tablas pueden complicar la extracción y transformación de datos para su posterior análisis.*
3. *La presencia de múltiples tablas puede dificultar la auditoría y validación de los datos.*

3.2.8 Criterio 8 - Mismo orden y número de columnas en todas las filas, series temporales y formatos

- Definición

La organización y estructura de los datos de manera coherente en una tabla o conjunto de datos asegura que todas las filas tengan la misma disposición de columnas y que esta estructura se aplique de manera consistente a lo largo del conjunto de datos. Es esencial para garantizar la integridad y la facilidad de procesamiento de los datos.

- Mejores prácticas asociadas
  1. *Verificar que no se hayan agregado o eliminado columnas en los nuevos conjuntos de datos, ya que esto podría afectar la consistencia y el análisis posterior.*
  2. *Mantener una estructura consistente en todas las series temporales y formatos relacionados, asegurando el mismo orden y número de columnas. Esto facilitará la comparación y el procesamiento de los datos.*

- Ejemplos
  1. *El siguiente ejemplo representa las dos buenas prácticas anteriores: Los ficheros "pmed_ubicacion_04-2023.xlsx" y "pmed_ubicacion_05-2023.xlsx" del dataset "Puntos de medidas de tráfico" tienen el mismo orden y número de columnas:*

A	B	C	D	E	F	G	H	I
tipo_elem	distrito	id	cod_cent	nombr	utm_x	utm_y	longitud	latitud

- Riesgos de no tener el mismo orden y número de filas

1. *La falta de uniformidad en el orden y número de columnas puede provocar inconsistencias en la estructura de datos.*
2. *Diferentes estructuras de datos pueden provocar errores durante el procesamiento y análisis de datos.*
3. *Estructuras de datos inconsistentes dificultan la reproducción de análisis y resultados.*

3.2.9 Criterio 9 - El nombre del fichero de datos es consistente con la serie anterior.

- Definición

  Cuando se trata de una serie de datos relacionados, se debe mantener un formato y estructura uniformes al nombrar los archivos. Facilita al usuario la identificación de los archivos.

  Evitar nombres complicados de ficheros, con signos, espacios, etc.
  Si se trata de ficheros de una serie cronológica, hay que garantizar la coherencia del nombre con los datos. Por ejemplo: es correcto que el fichero de datos de accidentes de 2023, se llame "2023_accidentes.xlsx", pero no sería correcto que se llamase "2021_accidentes.xlsx" ni tampoco sería recomendable que se llamase "accidentes.xlsx" puesto que cuando llegase el año 2024, este fichero no se podría volver a llamar "accidentes.xlsx" si ambos años siguiesen publicados.

- Mejores prácticas asociadas

1. *Mantener nombres coherentes y descriptivos para los ficheros de datos, asegurándose de que sean consistentes con la serie anterior. Esto facilitará la identificación y el seguimiento de los datos a lo largo del tiempo*

- Ejemplos
  1. *Los ficheros del dataset "Accidentes de tráfico de la Ciudad de Madrid", relativos a 2024 y 2023 tienen nombres coherentes. Los de 2024 se llaman "2024_Accidentalidad.xxx" y los de 2023 "2023_Accidentalidad.xxx".*
- Riesgos de no tener consistencia

1. *Nombres inconsistentes dificultan la identificación de archivos relacionados en una serie de datos.*
2. *Nombres de archivo inconsistentes pueden causar confusión en la organización de datos.*
3. *Nombres de archivo inconsistentes dificultan el seguimiento de versiones anteriores y posteriores de los datos.*
4. *Nombres de archivo inconsistentes pueden provocar errores al referenciar datos en documentos o sistemas externos.*

3.2.10 Criterio 10 - Uso de ";" como campo separador de caracteres en los ficheros CSV

- Definición

  *Los separadores de caracteres son utilizados para delimitar y distinguir diferentes elementos o campos en un conjunto de datos. El uso del símbolo ";" como separador de caracteres en ficheros CSV reduce ambigüedades, y asegura la integridad del conjunto de datos, ya que se evitan errores tanto en el cálculo y análisis, como en el procesamiento**. Permite a Excel interpretar correctamente los datos.** En algunas webs y algunas administraciones, a estos CSV separados por ";", se les llama "CSV for Excel" o "CSV para Excel".*

  No es incorrecto usar como separador de campos la coma "," pero en el Ayuntamiento de Madrid y otras administraciones, preferimos no emplearlo porque estos ficheros no se abren correctamente en Microsoft Excel.

- Mejores prácticas asociadas
  1. *Utilizar el símbolo de punto y coma ";" como separador de campos en lugar de otros caracteres como comas o tabulaciones. Mantener una consistencia y verificar que el uso del punto y coma no genere conflictos o ambigüedades con los datos mismos, especialmente si los datos contienen caracteres especiales o texto que también pueda incluir punto y coma.*

- Ejemplos

1. *El fichero "ContenedoresRopa.csv" del dataset "Contenedores de ropa" tiene el símbolo ";" como separador de caracteres:*

ID;TIPO_DATO;LOTE;COD_DISTRITO;DISTRITO;COD_BARRIO;BARRIO;DIRE
1;Contene 52;;;Calle; 3869;4470 875;40.38516785;-3.68677717;RO_024_00
2;Contene 14 frente; 3;4471034 3;40.38785956;-3.68775987;RO_024_002
3;Contene 32 esquin 6;4471046;40.38795031;-3.69021968;RO_024_003
4;Contene 49 ;;;Calle 7;4470976;40.38730993;-3.69184964;RO_024_004
5;Contene 4 ;;"Instal 0658;4470 429;40.3854922;-3.69214483;RO_024_005

- Riesgos de no cumplir con el ";"

1. *El uso de un separador de caracteres incorrecto puede provocar errores en la interpretación de los datos.*
2. *El uso de un separador de caracteres no estándar puede causar incompatibilidades con ciertas herramientas de software.*
3. *La falta de un estándar de separador de caracteres puede provocar inconsistencias en la forma en que se estructuran y procesan los datos.*

3.2.11 Criterio 11 - El conjunto de datos está dividido en distintas distribuciones de manera que cada una de ellas sea suficientemente tratable con programas informáticos habituales

- Definición

  Siempre que no haya problemas de volumen, se debe tender a tener un único fichero para simplificar, evitar problemas al reutilizador, y de cara a la generación de APIs. Sin embargo, ***si se supera el millón de registros en Excel****, se debe dividir el fichero en ficheros más pequeños, por años o por meses. Dividir un conjunto de datos en distribuciones más pequeñas reduce tiempos de procesamiento y optimiza el rendimiento, permitiendo al usuario aprovechar las capacidades de las herramientas y programas informáticos. Tener un conjunto de datos correctamente dividido permite la ejecución de operaciones paralelas, y facilita la identificación de errores en el proceso de análisis y procesamiento.*

  En caso de división de las series cronológicas, cada trozo debe tener la fecha o periodo en una columna.

- Mejores prácticas asociadas

  1. *Si el conjunto de datos es extenso o complejo, considerar dividirlo en distribuciones más pequeñas y manejables. Proporcionar la estructura y la documentación adecuadas para cada distribución, de modo que puedan ser utilizadas fácilmente con programas informáticos comunes.*

- Ejemplos
  Un ejemplo de práctica a mejorar es el fichero "1er trimestre.csv" del dataset "SER. Tiques de aparcamiento", cuyo tamaño es tan grande que no se puede procesar con python:

1er trimestre		11/28/2022 3:08 PM	Microsoft Excel Com...	1,207,107 KB

- Riesgos de tener datasets demasiado extensos

1. *Conjuntos de datos extensos pueden sobrecargar los recursos del sistema y ralentizar el rendimiento de los programas informáticos.*
2. *Conjuntos de datos grandes pueden ser difíciles de manejar y analizar con herramientas informáticas convencionales.*
3. *Conjuntos de datos extensos dificultan la identificación de errores y anomalías en los datos.*

3.2.12 Criterio 12 – No existe demasiada anidación en los datos

- Definición
  La anidación en los datos es la estructura de niveles de jerarquía que se ha podido crear en un conjunto de datos. Es importante evitar una anidación excesiva en los datos para mantener la claridad, la accesibilidad, y la eficiencia en su manipulación y análisis, afectando lo mínimo posible al rendimiento de las aplicaciones.

- Mejores prácticas asociadas
  1. *Evaluar si la anidación excesiva de los datos dificulta su análisis y comprensión con herramientas de análisis de datos (Excel, Python, etc.). Considerar si es necesario simplificar la estructura de los datos para facilitar su uso.*

- Ejemplos
  1. *El fichero "descuentos202404.json" del dataset "Deportes. Descuentos en instalaciones deportivas" apenas tiene anidación:*

```
[{
  "GRUPO_DESCUENTO" : "FAMILIA NUMEROSA",
  "MES" : "2024.03.MARZO",
  "CENTRO_DEPORTIVO" : "Centro Integrado Arganzuela",
  "DISTRITO" : "ARGANZUELA                                        ",
  "GRUPO_ACTIVIDAD_DEPORTIVA" : "AUSENCIA DE ACTIVIDAD DEPORTIVA",
  "SEXO" : "FEMENINO",
  "NUM_DESCUENTOS" : 125,
  "FX_CARGA" : "2024-04-03T07:00:47",
  "FX_DATOS_INI" : "2024-03-01",
  "FX_DATOS_FIN" : "2024-03-31"
},{
  "GRUPO_DESCUENTO" : "EDAD INFANTIL",
  "MES" : "2024.03.MARZO",
  "CENTRO_DEPORTIVO" : "Gallur",
  "DISTRITO" : "LATINA                                            ",
  "GRUPO_ACTIVIDAD_DEPORTIVA" : "AUSENCIA DE ACTIVIDAD DEPORTIVA",
  "SEXO" : "FEMENINO",
  "NUM_DESCUENTOS" : 398,
  "FX_CARGA" : "2024-04-03T07:00:47",
  "FX_DATOS_INI" : "2024-03-01",
  "FX_DATOS_FIN" : "2024-03-31"
},{
```


Un ejemplo de fichero con demasiada anidación es "ocupacion_rodajes_2022.xml". Esto es una práctica a mejorar:


```
egobfiles_MANUAL_300311_ocupacion_rodajes_2022.xml
- Perfilado Opendata - Perfilado Opendata > Ficheros Datasets > 300311 - Ocupacion via publica > 2022 > egobfiles_MANUAL_300311_ocupacion_rodajes_2022.xml
1  <?xml version="1.0" encoding="UTF-8"?>
2  <soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
3     <soap:Body>
4        <ns2:ObtenerOpenDataResponse xmlns:ns2="http://www.ocuvi.iam.es/WSObtencionOpenData/ObtenerOpenDa
5           <ns2:resultado>OK</ns2:resultado>
6           <ns2:numOcupaciones>9630</ns2:numOcupaciones>
7           <ns2:Ocupaciones>
8              <idSolicitud>21/RAC/5969</idSolicitud>
9              <tipoSolicitud>Acto comunicado de Rodaje</tipoSolicitud>
10             <origenSolicitud>Externa</origenSolicitud>
11             <departamentoTramitacion>Régimen Jurídico y Autorizaciones - Rodajes</departamentoTramitaci
12             <estadoTramitacion>TRAMITADA</estadoTramitacion>
13             <estadoResolucion>AUTORIZADA</estadoResolucion>
14             <fechaCreacion>2021-04-29+02:00</fechaCreacion>
15             <fechaUltimaModificacion>2021-04-29+02:00</fechaUltimaModificacion>
16             <fechaEntradaRegistro>2021-04-26+02:00</fechaEntradaRegistro>
17             <fechaRecibidaDepartamento>2021-04-29+02:00</fechaRecibidaDepartamento>
18             <solicitante>
19                <tipoSolicitante>EMPRESA</tipoSolicitante>
20                <documento>A86821444</documento>
21                <razonSocial>ALEA MEDIA S.A</razonSocial>
22             </solicitante>
23             <datosEspecificos>
24                <pregunta>Motivo</pregunta>
25                <respuesta>Serie Televisión</respuesta>
```


2. *El fichero "Órdenes de ejecución 2020.xls" del dataset "Ordenes de ejecución para el cumplimiento del deber de conservación y rehabilitación exigidas a los propietarios" tiene demasiada información anidada:*

DIRECCIÓN GENERAL DE LA EDIFICACIÓN
SUBDIRECCIÓN GENERAL DE CONTROL DE LA EDIFICACIÓN

Órdenes de ejecución dictadas por el Ayuntamiento para el cumplimiento del deber de conservación y rehabilitación exigidas a los propietarios

DURANTE EL PRIMER TRIMESTRE DE 2020

NUMERO DE ORDEN	Emplazamiento principal	MOTIVO	MES
1	PEDRO VALDIVIA, 36	ORDEN DE EJECUCIÓN	ENERO
2	RICARDO ORTIZ, 110	ORDEN DE EJECUCIÓN	ENERO
3	SAN LUIS, 144	ORDEN DE EJECUCIÓN	ENERO
4	TRITON, 10	ORDEN DE EJECUCIÓN	ENERO
5	AVDA. ORCASUR, 47	ORDEN DE EJECUCIÓN	ENERO

- Riesgos de tener excesiva anidación

1. *Una anidación excesiva puede dificultar la comprensión de la estructura de los datos y la relación entre diferentes elementos.*
2. *La anidación excesiva puede afectar negativamente la eficiencia y el rendimiento de las aplicaciones que manipulan los datos.*
3. *Una estructura de datos excesivamente anidada puede dificultar la escalabilidad y el mantenimiento a medida que el conjunto de datos crece o cambia con el tiempo.*

3.2.13 Criterio 13 - Codificación correcta de caracteres

- Definición

La codificación de caracteres es el método utilizado para representar y almacenar caracteres y símbolos en un sistema informático.

Mantener una codificación correcta de caracteres garantiza la perfecta visualización de los datos y que no se generen problemas cuando se pasa de un ordenador a otro.

La incorrecta codificación de datos puede dar lugar a problemas de interpretación y lectura, de análisis de la información, o, incluso, a la pérdida de esta.

*Las codificaciones más utilizadas incluyen **UTF-8, UTF-16, ASCII y ISO-8859-1**.*

***La codificación deseable para los ficheros publicados en el Portal de Datos Abiertos de Madrid es la UTF-8**. es ampliamente preferido por su capacidad de representar caracteres de múltiples idiomas y su eficiencia en el almacenamiento*

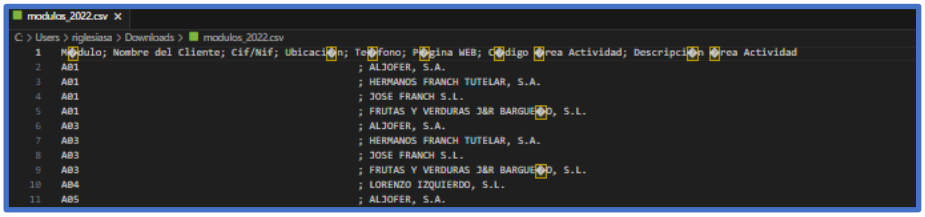
Este tema de la codificación, es algo a indicar al personal informático que genera el fichero.

- Mejores prácticas asociadas
    1. *Utilizar una codificación estándar y ampliamente aceptada, como UTF-8, para garantizar la compatibilidad y evitar problemas de interpretación de caracteres especiales.*
- Ejemplos

1. *El fichero Actuaciones_UDC.csv tiene codificación utf-8.*

NOMBRE FICHERO	FORMATO FICHERO	PATH FICHERO	CODIFICACION
Actuaciones_UDC.csv	csv	300635 - Actuaciones UDC	utf-8

Un ejemplo de práctica a mejorar es el fichero "modulos_2022.csv" perteneciente al dataset "Mercamadrid. empresas y ubicación", que no tiene codificación utf-8. Se pueden observar caracteres no válidos en las tildes, en el uso de la letra "ñ", etc.:



- Riesgos de no tener una codificación correcta

1. *Una codificación incorrecta puede resultar en la visualización incorrecta de caracteres especiales y símbolos.*
2. *Diferentes sistemas y plataformas pueden interpretar la codificación de caracteres de manera diferente.*
3. *Una codificación incorrecta puede resultar en la pérdida o corrupción de datos durante el proceso de almacenamiento o transferencia.*

3.2.14 Criterio 14 – Organización vertical de la información, en vez de horizontal

- Definición

  La organización vertical de la información pretende estructurar y presentar datos de manera que se priorice la claridad, la estructura y la profundidad en la presentación de datos. ***En una organización vertical, el fichero crece siempre hacia abajo en nuevas filas de datos.***

  En una organización horizontal, el fichero crece siempre apareciendo nuevas columnas de datos, *lo cual genera que va cambiando la estructura del documento, y además dificulta la explotación de la información. Aunque visualmente para un ciudadano puede resultar más atractivo, hay que recordar que el portal de Datos Abiertos está focalizado en la reutilización de los datos y no tanto en la visualización por un ciudadano.*

  En otras palabras, en lugar de dispersar la información en múltiples dimensiones o categorías horizontales, este criterio busca organizar los datos de manera que reflejen una estructura vertical clara, y que permita un mayor nivel de detalle.

- Mejores prácticas asociadas

  1. *Agrupar los datos en categorías principales que representen conceptos más generales o de alto nivel, y luego desglosar estos conceptos en subcategorías más específicas o detalladas.*

- Ejemplos
  1. *El fichero del ejemplo del criterio anterior, "Agua_Regenerada_Madrid_2022.xlsx" del dataset "Volumen de agua regenerada", tiene una organización horizontal. Es un ejemplo de práctica a mejorar:*

ERAR	Jan-01	Jan-02	Feb-01	Feb-02	Mar-01	Mar-02	Apr-01	Apr-02
VIVEROS	4,032.00	6,723.00	26,607.00	32,421.00	15,945.00	7,558.00	53,913.00	28,745.00
LA CHINA	107,629.72	106,956.00	75,825.00	38,443.00	33,956.00	36,672.00	40,870.72	45,944.44
LA GAVIA	9,954.00	12,329.00	19,667.00	15,802.00	23,611.00	20,015.00	24,052.00	21,356.00
LAS REJAS	24,079.00	29,213.00	57,184.00	64,595.00	44,740.00	40,212.00	50,180.00	51,649.00
TOTAL	145,694.72	155,221.00	179,283.00	151,261.00	118,252.00	104,457.00	169,015.72	147,694.44
		300,915.72		330,544.00		222,709.00		316,710.16

  2. *El fichero "Abonados_2022.xlsx" del dataset "Deportes. Abonados en Centros Deportivos Municipales" tiene una organización vertical. Es un ejemplo de buena práctica.*

	A	B	C	D	E	F
1	Nº de abonados	Sexo	Edad	Tipo de abono	Centro deportivo	Mes
2	1	MUJER	18	ADM ACTIVIDAD DIRIGIDA	Alfredo Goyeneche	Jan-22
3	1	MUJER	38	ADM ACTIVIDAD DIRIGIDA	Alfredo Goyeneche	Jan-22
4	2	MUJER	50	ADM ACTIVIDAD DIRIGIDA	Alfredo Goyeneche	Jan-22

- Riesgos de no seguir la organización vertical

1. *La organización horizontal puede dificultar la comprensión de la información al dispersarla en múltiples dimensiones o categorías.*
2. *Al presentar los datos en múltiples columnas, la estructura del documento cambia constantemente, lo que dificulta la extracción y análisis de la información de manera eficiente.*
3. *Una organización horizontal puede limitar esta reutilización al dificultar el acceso y la comprensión de los datos, lo que reduce su valor para la comunidad.*

3.2.15 Criterio 15 – Identificación en los datos, del año y mes a que hacen referencia

- Definición

  La inclusión de información clara y explícita dentro de un conjunto de datos que indique el período de tiempo al que corresponden los datos presentados. Este criterio exige que cada registro o conjunto de datos esté etiquetado con la fecha precisa, incluyendo el año y el mes al que se refieren.

- Mejores prácticas asociadas
  1. *El uso de un formato estándar de fecha, preferiblemente YYYY-MM (año-mes), asegurando que la fecha de referencia esté claramente identificada y visible en cada registro de datos o conjunto de datos.*

     Aunque estamos muy habituados a ver el mes en letra, si lo expresamos en número, permitirá una ordenación cronológica mejor.

2. *Se recomienda actualizar regularmente la fecha de referencia para reflejar la información más reciente disponible, y proporcionar documentación detallada sobre cómo se determina y se registra esta fecha, asegurando así la compatibilidad con estándares y prácticas comunes en el ámbito de aplicación de los datos. Esta buena práctica se relaciona con la mejor práctica número 2 del criterio 1: Apartado 3.1.1.*

- Ejemplos
  1. *El fichero "ActuacionesBomberos_2024.xlsx" del dataset "Actuaciones del Cuerpo de Bomberos" tiene cada registro etiquetado con la fecha, y la fecha de referencia está incluida en el nombre del fichero.*

AÑO	MES	DISTRITO	FUEGOS	DAÑOS EN CONSTRUCCION	SALVAMENTOS Y RESCATES	DAÑOS POR AGUA	INCIDENTES DIVERSOS	SALIDAS SIN INTERVENCION	SERVICIOS VARIOS	TOTAL
2024	marzo	CENTRO	51	56	75	25	85	13	14	319
2024	marzo	ARGANZUELA	8	16	28	9	23	0	2	86
2024	marzo	RETIRO	7	8	30	8	17	2	1	73

- Riesgos de no cumplir con el criterio

1. *Sin la inclusión clara del año y mes a los que corresponden los datos, los usuarios pueden enfrentarse a dificultades para entender el contexto temporal de la información.*
2. *La falta de etiquetado temporal adecuado dificulta la capacidad de realizar análisis temporales significativos.*
3. *Los conjuntos de datos sin identificación temporal pueden perder relevancia con el tiempo, ya que los usuarios no pueden determinar cuándo fueron recopilados los datos y si aún son válidos o actualizados.*
4. *La falta de etiquetas de tiempo estandarizadas dificulta la comparación de datos entre diferentes períodos.*

3.2.16 Criterio 16 – Mes mejor en formato numérico en vez de texto, para permitir una ordenación cronológica de los meses, en vez de alfabética

- Definición

  Establece la preferencia por representar los meses del año utilizando valores numéricos en lugar de texto. Esta práctica se adopta con el propósito de facilitar la ordenación cronológica de los datos, en lugar de depender de una ordenación alfabética que podría generar confusiones. En este enfoque, cada mes se designa con un número del 01 al 12, correspondiente a enero hasta diciembre, lo que permite una disposición secuencial y sistemática de los datos según su temporalidad.

- Mejores prácticas asociadas
  1. *Mantener la consistencia en la representación numérica de los meses en todos los conjuntos de datos relacionados, así como documentar claramente esta convención para garantizar su comprensión y aplicación adecuada.*

2. *Se recomienda validar la precisión de los valores numéricos asignados a los meses para evitar errores y garantizar una ordenación cronológica precisa y confiable de los datos.*

- Ejemplos
  1. *El fichero "reservas_moto.xlsx" del dataset "Moto. Reservas Moto" tiene la columna de fechas como valor de fecha:*

Gis_X	Gis_Y	Fecha de Alta	Distrito	Barrio
439587.2800000000	4473509.2200000000	8/9/2023	1 CENTRO	11 PALACIO
439587.7200000000	4473915.0900000000	7/25/2023	1 CENTRO	11 PALACIO
439677.9300000000	4474298.7700000000	7/25/2023	1 CENTRO	11 PALACIO
439706.5200000000	4473415.0100000000	8/3/2023	1 CENTRO	11 PALACIO
439798.1600000000	4474465.9700000000	11/18/2022	1 CENTRO	11 PALACIO
439833.8600000000	4474547.6300000000	6/3/2023	1 CENTRO	11 PALACIO
439725.6100000000	4473693.4100000000	7/25/2023	1 CENTRO	11 PALACIO

Este fichero tendría una segunda mejora y es separar en distintas columnas los códigos de distrito y barrio, con las denominaciones, como se ha indicado en apartados previos.

- Riesgos de no seguir el formato numérico

1. *Al utilizar representaciones textuales de los meses, se depende de una ordenación alfabética que puede generar confusiones.*
2. *La falta de consistencia en la representación de los meses puede complicar el análisis temporal de los datos.*
3. *Al no utilizar un formato numérico estándar para los meses, se puede dificultar la integración de los datos en herramientas de análisis que dependen de una ordenación cronológica precisa para generar visualizaciones o informes.*
4. *La representación textual de los meses puede aumentar la probabilidad de errores en la entrada de datos, especialmente si se utilizan diferentes convenciones de escritura o abreviaturas.*

3.2.17 Criterio 17 - Utilizar el mayor número de formatos posibles, para que los datos sean más accesibles.

- Definición

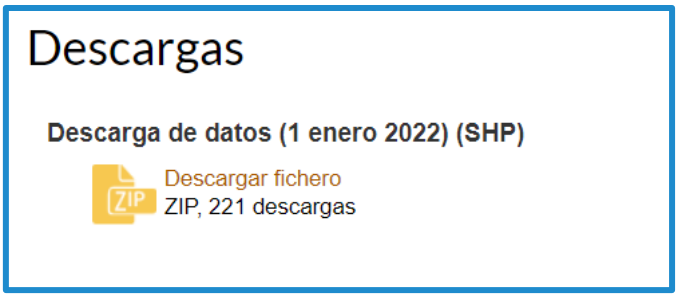
  Proporcionar datos en una variedad de formatos para aumentar su accesibilidad y utilidad para una amplia gama de usuarios. Esto implica ofrecer los datos en diferentes formatos, como CSV, JSON, XML, entre otros, para adaptarse a las preferencias y necesidades de los usuarios y facilitar su integración en diferentes sistemas y aplicaciones.

  Si estamos mostrando datos espaciales en formato SHAPE, es recomendable que también exista un CSV o XLS (con el mayor número de campos posibles), para que aquellas personas que no sean capaces de interpretar un SHAPE o no tienen las herramientas adecuadas, puedan utilizar o analizar la información.

- Mejores prácticas asociadas
  1. *Proporcionar los datos en múltiples formatos, incluyendo CSV, JSON, XML, RDF, entre otros, para adaptarse a las diferentes necesidades y preferencias de los usuarios. Esto aumenta la accesibilidad de los datos y permite su uso en una variedad de contextos y aplicaciones.*
  2. *Priorizar el uso de formatos estándar y abiertos que sean ampliamente reconocidos y compatibles con una variedad de herramientas y plataformas. Esto garantiza la interoperabilidad y la reutilización de los datos en diferentes entornos.*
  3. *Facilitar la conversión entre diferentes formatos mediante el uso de herramientas y servicios que permitan la transformación de los datos de un formato a otro de manera rápida y sencilla. Esto mejora la interoperabilidad y la flexibilidad en el uso de los datos.*

- Ejemplos
  1. *El dataset "Actividades Culturales y de Ocio Municipal en los próximos 100 días" tiene varios formatos para descargar.*

  2. *El fichero "Defunciones clasificadas por nacionalidad y sexo según Distrito y Barrio. 1 Enero 2022" el formato está exclusivamente en formato SHAPE sin existir un formato tabular asociado.*



- Riesgos de no tener diferentes formatos

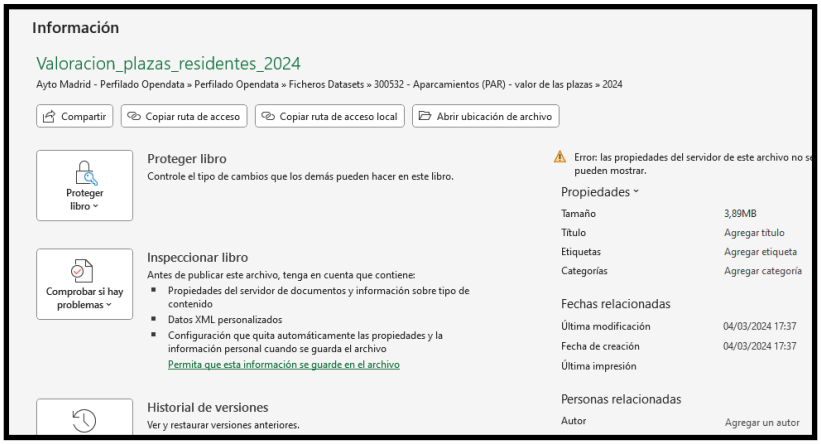
1. *Al ofrecer los datos en un único formato o en un número limitado de formatos, se excluye a usuarios que puedan preferir o necesitar otros formatos para trabajar con los datos.*
2. *Al no ofrecer una variedad de formatos, se puede excluir a usuarios que no cuenten con las herramientas o habilidades necesarias para trabajar con el formato proporcionado.*
3. *Al no ofrecer datos en una variedad de formatos, se reduce la posibilidad de que los datos sean reutilizados en diferentes contextos y por diferentes usuarios.*

3.2.18 Criterio 18 – No existen metadatos de autor

- Definición
  La ausencia de metadatos específicos que identifiquen personalmente al autor o autores de un fichero de datos. La eliminación de estos metadatos busca preservar la privacidad de la persona que genera el documento. Aunque el conocimiento del autor puede proporcionar un contexto valioso, no es una práctica recomendable en el marco de los datos abiertos. Ya se indica asociado al conjunto de datos, que unidad del Ayuntamiento es la responsable de ese conjunto de datos.

- Mejores prácticas asociadas
  1. *Revisión y limpieza de metadatos: Antes de publicar cualquier conjunto de datos, realizar un proceso de revisión para identificar y eliminar cualquier metadato de autor.*

- Ejemplos
  2. *El dataset "Aparcamientos municipales para residentes (PAR). Valor de las plazas", no tiene metadatos de autor para el fichero de 2024, mientras que si tiene en el de 2023:*



Importante: al final del documento hay un anexo con los pantallazos paso a paso, de como eliminar los metadatos de autor asociado a un fichero Excel o un fichero de office.

Estos metadatos si no se eliminasen, saldrían publicados y serían accesibles por cualquier ciudadano. Por otro lado, si a partir de un documento office se genera un documento en pdf, estos metadatos de autorías y fechas quedarían también presentes en ese documento en pdf.

- Riesgos de incluir metadatos

1. *La inclusión de metadatos que identifiquen personalmente al autor o autores puede comprometer la privacidad.*

3.3 Datos

Corresponden a la información que se comparte en el mundo de los datos abiertos. Pueden representar todo tipo de observaciones sobre todo tipo de temas. Los datos, que pueden adoptar la forma de números, texto, imágenes o cualquier otro formato, son el "producto" que se comparte con los usuarios.

Los siguientes criterios permiten una mejor reutilización de los datos.

3.3.1 Criterio 1 – Orden lógico de las columnas

- Definición
  El criterio de orden lógico de las columnas en el contexto de datos abiertos se refiere a la organización sistemática y coherente de las columnas en un conjunto de datos para facilitar su comprensión, análisis y reutilización. Esta organización debe seguir un orden que refleje una estructura lógica y jerárquica de la información.

- Mejores prácticas asociadas

1. *Mantener siempre el mismo orden de las columnas en todos los conjuntos de datos similares. Esto facilita la automatización y el análisis por parte de los usuarios.*

2. *Colocar primero las columnas que contienen identificadores únicos y datos temporales. Esto permite una rápida referencia y seguimiento de cada registro.*

3. *Ordenar las columnas siguiendo una estructura jerárquica lógica, desde divisiones administrativas más amplias (como distrito) hasta las más específicas (como barrio y dirección).*

4. *Adoptar estándares reconocidos para la organización de datos. Por ejemplo, utiliza formatos y convenciones comúnmente aceptados para fechas, coordenadas y otros datos específicos.*

- Ejemplos

1. *El dataset "Tráficos. Cabezas de semáforo" tiene los campos ordenados de manera lógica:*

CAMPO
tipo_elem
distrito
id
Id_cruce
fecha_inst
utm_x
utm_y
longitud
latitud

- Riesgos

1. *Si las columnas no están organizadas de manera lógica, los usuarios pueden tener dificultades para entender la estructura y el significado de los datos.*
2. *Un orden desordenado de las columnas puede dificultar el proceso de análisis de datos, ya que los usuarios pueden tener que buscar y reorganizar la información antes de poder realizar cualquier análisis significativo.*
3. *Un orden inconsistente o ilógico de las columnas puede aumentar la probabilidad de errores en la manipulación y el procesamiento de los datos.*

3.3.2 Criterio 2 - Los tipos de campos se ajustan a lo esperado

- Definición

  Los tipos de campos en el contexto del manejo de datos se refieren a las categorías predefinidas que describen el contenido y formato de los datos en una columna, ya sea entero, decimal, fecha, texto... etc. Es crucial que los tipos de campos se ajusten a lo esperado, y que se alineen con las características reales de los datos, garantizando que sean precisos y coherentes.

- Mejores prácticas asociadas
  1. *Verificar que los tipos de datos utilizados en los campos sean apropiados y se ajusten a la naturaleza de la información. Esto evitará problemas de interpretación y permitirá un análisis preciso.*

- Ejemplos
  1. *En el fichero "plan_operativo_gobierno_02062023.xls" los tipos de datos se ajustan a los esperado según los nombres de las columnas:*

Código actuación	Denominación actuación	Situación de la actuación	Código Área de Gobierno
45791	Creación de un Centro de Competencias Tecnológicas	Cumplida	5
45807	Promoción de la participación activa de la mujer en la vida laboral	Cumplida	5
45835	Salvaguarda del empleo, formando a personas "activas" que necesiten reciclaje en diferentes materi	Cumplida	5
45799	Impulso a la formación digital - Digitalización de los cursos para crear eficiencia en los recursos. Redu	Cumplida	5
45803	Puesta en marcha de talleres de formación técnica profesional	Cumplida	5
45783	Orientación laboral y entrenamiento de competencias transversales y de acceso al empleo mediante	Cumplida	5
45787	Favorecimiento de la formación y el desarrollo de las competencias profesionales necesarias para fac	Cumplida	5
45795	Impulso de la formación con una oportunidad de experiencia en un entorno laboral real para la mejo	Cumplida	5
45811	Casación inteligente de las ofertas de empleo y los perfiles de personas desempleadas	Cumplida	5

- Riesgos

1. *Si los tipos de campos no se ajustan adecuadamente a la naturaleza de la información, puede resultar en una pérdida de precisión en los datos.*
2. *Los tipos de campos incorrectos pueden dificultar el análisis de los datos, ya que ciertas operaciones y funciones pueden no ser aplicables o pueden producir resultados incorrectos debido a la falta de coherencia en los tipos de datos.*
3. *Si los tipos de campos no se ajustan correctamente, puede resultar en una presentación inconsistente de los datos, lo que dificulta la comprensión y el uso de la información por parte de los usuarios.*

3.3.3 Criterio 3 - Asignación de un ID único

- Definición

  Un ID (identificador) único es un valor exclusivo asignado a cada registro de un conjunto de datos. La asignación de un ID único es esencial para evitar duplicados y errores. Además, permite establecer relaciones entre conjuntos de datos.

- Mejores prácticas asociadas
  1. *Asignar un identificador único a cada registro o entidad en los datos (columna) para facilitar su identificación y seguimiento. Utilizar un esquema de ID claro y consistente para evitar duplicados o confusiones.*
- Ejemplos
  1. *El fichero "servicios_sociales_personasatendidas_2023.xlsx" del dataset "Personas atendidas" tiene un identificador único. La columna "Secuencia":*

Secuencia	Código Centro	Centro	Código Distrito Centro	Distrito Centro	Código Distrito	Distrito	Código Barrio	Barrio	Sección Censal	Tramo Edad	Nacionalidad	Sexo	Año Atención
1	104	CSS Infante Don Juan	09	MONCLOA-ARAVACA	02	ARGANZUELA	0204	LEGAZPI	101	80 - 84	Española	H	2,023
2	69	CSS José Villarreal	02	ARGANZUELA	02	ARGANZUELA	0202	ACACIAS	015	60 - 64	Española	H	2,023
3	69	CSS José Villarreal	02	ARGANZUELA	02	ARGANZUELA	0202	ACACIAS	018	65 - 69	Extranjera	H	2,023
4	69	CSS José Villarreal	02	ARGANZUELA	02	ARGANZUELA	0202	ACACIAS	028	30 - 34	Española	M	2,023
5	69	CSS José Villarreal	02	ARGANZUELA	02	ARGANZUELA	0202	ACACIAS	087	>= 85	Española	M	2,023
6	69	CSS José Villarreal	02	ARGANZUELA	02	ARGANZUELA	0202	ACACIAS	096	>= 85	Española	M	2,023

- Riesgos

1. *La ausencia de un identificador único puede llevar a la presencia de registros duplicados en el conjunto de datos, lo que dificulta la integridad y la precisión de la información.*
2. *Sin un ID único, puede resultar difícil identificar y distinguir entre diferentes registros en el conjunto de datos, especialmente cuando los registros comparten características similares.*
3. *La falta de un ID único puede dificultar la establecimiento de relaciones entre diferentes conjuntos de datos, lo que limita la capacidad de realizar análisis y consultas que involucren múltiples conjuntos de datos.*

3.3.4 Criterio 4 - Los valores de datos de tipo fecha y fecha/hora deben describirse en formato ISO 8601

- Definición

  El formato ISO 8601 es una norma internacional para representar fechas y horas de manera consistente y comprensible. Es importante representar los valores fecha y fecha/hora de forma estandarizada, ya que garantiza la uniformidad en la interpretación, evita ambigüedades y facilita la comparación y el análisis de datos en diferentes sistemas y aplicaciones.

  *Así la fecha podría ir de la siguiente forma**: AAAAMMDD o AAAA/MM/DD o AAAA-MM-DD**. Esta forma de mostrar la fecha permitirá también ordenaciones cronológicas más fáciles.*

- Mejores prácticas asociadas

  1. *Representar las fechas y horas en formato v para facilitar la interoperabilidad y la comprensión de los datos temporales. Evitar ambigüedades y asegurarse de que las fechas se interpreten correctamente en diferentes sistemas y aplicaciones.*

- Ejemplos

  1. *Como ejemplo de práctica a mejorar en el uso del formato ISO 8601, que establece que las fechas completas deben representarse: YYYY-MM-DD (Año-Mes-Día), se tiene el fichero "TAXI_Flota_Diario.xls", el cual no tiene la columna "Fecha inicio de prestación del servicio de taxi" de acuerdo a este formato. Por ejemplo, la primera fila debería ser "2021-06-18":*

Fecha inicio de prestación del servicio de taxi
18/06/2021
21/06/2021
16/09/2014

- Riesgos

1. *Si los valores de fecha y fecha/hora no siguen un formato estándar como ISO 8601, puede llevar a interpretaciones erróneas o ambiguas por parte de los usuarios.*
2. *La falta de un formato estandarizado dificulta la comparación y el análisis de datos temporales entre diferentes conjuntos de datos o sistemas.*
3. *Los sistemas y aplicaciones pueden tener dificultades para interpretar correctamente los valores de fecha si no siguen un formato estándar.*

3.3.5 Criterio 5 - Cumplimiento de codificación para información de barrios y distritos

- Definición

La codificación estándar de barrios y distritos se refiere asignación de códigos o identificadores únicos a cada área geográfica localizada. La correcta codificación de barrios y distritos tiene aplicación en aspectos relacionados con la planificación urbana, la toma de decisiones basadas en datos y la implementación de políticas específicas para áreas geográficas particulares, entre otros. La información de barrios y distritos se puede encontrar en los datasets "Barrios municipales de Madrid" y "Distritos municipales de Madrid". En el Anexo III se encuentra una tabla abreviada con algunos distritos.

Desgraciadamente, todavía siguen apareciendo tablas de datos con distritos como: San Blas ó Moncloa o Fuencarral, cuando desde hace años es "San Blas – Canillejas", "Moncloa – Aravaca" o "Fuencarral – El Pardo".

- Mejores prácticas asociadas

1. *Asegurarse de que la información relacionada con los distritos y barrios cumpla con las pautas y estándares de codificación establecidos. Esto garantizará la coherencia y la interoperabilidad de los datos. Utilizar el estándar definido en el siguiente dataset del portal:*

   "Barrios municipales de Madrid" y "Distritos municipales de Madrid".

- Ejemplos

1. *El dataset "Barrios.xlsx" del dataset "Barrios Municipales de Madrid" tiene los códigos de distrito de acuerdo con el estándar. Se observa como ejemplo de práctica a mejorar, que no utiliza codificación utf-8 (criterio 3.2.1):*

OID_	OBJECTID	CODDIS	NOMDIS	COD_BAR	NOMBRE
0	132	17	Villaverde	172	San CristÃ³bal
1	133	12	Villaverde	173	Butarque
2	134	18	Villaverde	175	Ãngeles
3	135	11	Villaverde	174	Los Rosales
4	136	10	Villaverde	171	Villaverde Alto - Casco HistÃ³rico de Villaverde
5	137	13	Usera	121	Orcasitas
6	138	2	Villa de Vallecas	183	Ensanche de Vallecas
7	139	14	Carabanchel	116	Buenavista

- Riesgos

1. *La falta de cumplimiento con las pautas y estándares de codificación puede llevar a la presencia de inconsistencias en los datos, como nombres de distritos o barrios mal escritos o no actualizados.*
2. *Si los nombres de los distritos y barrios no siguen un estándar de codificación, puede ser difícil comparar y analizar datos entre diferentes conjuntos de datos o fuentes, lo que limita la capacidad de tomar decisiones informadas basadas en los datos.*
3. *Los sistemas y aplicaciones pueden tener dificultades para interpretar correctamente los valores de fecha si no siguen un formato estándar.*

3.3.6 Criterio 6 – Formato de dirección válida.

- Definición

  Proporcionar orientación o instrucciones claras sobre cómo representar las direcciones dentro de un conjunto de datos. Esta recomendación se basa en los campos disponibles en el conjunto de datos, como el tipo de vial, el nombre de la vía y el número de edificio, y tiene como objetivo estandarizar la forma en que se presentan las direcciones para garantizar su consistencia y comprensión.

  Los campos que se recomiendan tener son los siguientes:

  - tipo_vial
  - nombre_vial
  - numero_vial
  - cirección_auxiliar (si la hubiese)
  - ndp (Código NDP o número de policía)
  - coordenada_x
  - coordenada_y
  - latitud
  - longitud

- Mejores prácticas asociadas

  1. *Establecer un formato estándar para la representación de direcciones dentro del conjunto de datos. Esto puede incluir el orden de los campos (tipo de vial, nombre de la vía, número de edificio), el uso de separadores (como comas o espacios) y la capitalización de las palabras.*

2. *Aprovechar al máximo los campos disponibles en el conjunto de datos para incluir toda la información relevante para una dirección completa y precisa. Esto puede incluir no solo el tipo de vial, el nombre de la vía y el número de edificio, sino también campos adicionales como el código postal, el distrito o la ciudad.*

- Ejemplos
  1. *En las estructuras de datos de los datasets, ofrecer una guía en los campos relacionados a direcciones. Un ejemplo es la ficha de estructura del dataset "Mobiliario Urbano. Papeleras".*

NOMBRE_VIAL	Varchar	Nombre del Vial.
TIPO_NÚMERO	Varchar	Tipo de numeración: número, kilómetro….
NÚMERO	Entero	Número del vial.
COD_BARRIO	Entero	Código del distrito.
BARRIO	Varchar	Denominación del Barrio.
COD_DISTRITO	Entero	Código del Distrito .
DISTRITO	Varchar	Denominación del Distrito.

- Riesgos

1. *La presencia inesperada de valores nulos o la falta de ellos cuando se esperan puede comprometer la integridad y calidad de los datos, afectando su utilidad y fiabilidad para el análisis.*
2. *La falta de claridad sobre qué representa un valor nulo frente a un valor cero o un valor no nulo puede llevar a confusiones y malas interpretaciones de los datos.*
3. *En bases de datos relacionales, los valores nulos en campos clave pueden romper la integridad referencial, dificultando la correcta vinculación entre diferentes conjuntos de datos.*

3.3.7 Criterio 8 - Los valores nulos y no nulos se ajustan a lo esperado

- Definición

Se define como valor no nulo a aquel valor que está informado en el conjunto de datos. En consecuencia, el valor nulo es aquel valor faltante, que no está informado. Es importante analizar si estos valores se ajustan a lo esperado, bien analizando si los valores se ajustan a las expectativas del conjunto de datos, o bien estudiando posibles variaciones significativas entre series, y si esta variación es coherente con el contexto de los datos. ***Es importante diferenciar entre valor nulo y valor cero. Un valor cero es un valor informado, mientras que un valor nulo indica ausencia de información.***

- Mejores prácticas asociadas
  1. *Verificar que los valores no nulos sean consistentes con las expectativas y requisitos de los datos. Esto ayudará a garantizar la integridad y calidad de los datos.*
  2. *Comprobar si hay una disminución significativa en el número de celdas no vacías y verificar si el número de celdas no vacías sigue siendo coherente con la serie anterior y si cumple con los requisitos esperados.*

- Ejemplos
  1. *Mientras que el fichero "mupis_pilas.xlsx" se ajusta a lo esperado, todos los campos están correctamente informados, en el fichero "datos_de_gestion.xls", del dataset "Multas de circulación" hay mayor número de celdas vacías. Es un ejemplo de práctica a mejorar:*

NOMBRE DATASET	NOMBRE FICHERO	COLUMNA	Número de celdas vacías	Número de celdas no vacías
Multas de circulación	datos_de_gestion.xls	mes	13	181
Multas de circulación	datos_de_gestion.xls	número total de boletines de denunc	9	185
Multas de circulación	datos_de_gestion.xls	boletines no iniciables o de conclusió	163	31
Multas de circulación	datos_de_gestion.xls	pagados en vía voluntaria	9	185
Multas de circulación	datos_de_gestion.xls	recaudación en vía voluntaria	9	185
Multas de circulación	datos_de_gestion.xls	importe por boletín pagado en volun	9	185
Multas de circulación	datos_de_gestion.xls	boletines pagados en voluntaria resp	9	185
Multas de circulación	datos_de_gestion.xls	boletines transferidos a la vía ejecuti	16	178
Multas de circulación	datos_de_gestion.xls	importe boletines transferido a la vía	16	178
Multas de circulación	datos_de_gestion.xls	denuncias a extranjeros	110	84

NOMBRE FICHERO	PATH FICHERO	COLUMNA	Número de celdas vacías	Número de celdas no vacías
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	marca temporal	4	587
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	sexo	11	580
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	edad	8	583
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	distrito donde asistes a marcha nórdic	31	560
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	¿cómo le llegó la información sobre la	43	548
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	¿practicaba la marcha nórdica antes de	8	583
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	la atención recibida cuando se inscribi	6	585
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	información del programa en el centro	23	568
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	procedimiento de inscripción y reserv	66	525
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	organización de la actividad.	4	587
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	lugares elegidos para la realización de	5	586
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	cumplimiento del calendario anual de	18	573
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	adquisición de la técnica básica para e	13	578
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	adquisición de conocimientos de princ	10	581
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	puntualidad del profesorado.	4	587
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	calidad de las clases recibidas.	5	586
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	satisfacción global con el profesorado.	8	583
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	satisfacción global con las clases.	13	578
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	se producen ausencias de profesorado	6	585
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	en caso de ausencias ¿se recuperan es	319	272
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	relación calidad precio del servicio rec	7	584
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	satisfacción global de la actividad mar	10	581
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	¿recomendaría usted el programa a fa	8	583
NORDICA_RESULTADOS_2021.xlsx	300267 - Encuesta participantes marcha nordica\2021	para finalizar, indíquenos, si lo cree op	293	298

- Riesgos

1. *La falta de un formato estándar para las direcciones puede llevar a inconsistencias en cómo se ingresan y se presentan las direcciones dentro de un conjunto de datos.*
2. *Sin un formato consistente, validar y normalizar las direcciones se vuelve un proceso complejo y propenso a errores.*
3. *Los servicios de geocodificación, que convierten direcciones en coordenadas geográficas, dependen de formatos de dirección estandarizados.*

3.3.8 Criterio 7 - Las coordenadas latitud y longitud, correctamente representadas

- Definición

Las coordenadas expresadas en latitud y longitud son valores numéricos que indican la ubicación geográfica precisa de un punto en la Tierra. Un formato estándar facilita la integración de datos de diferentes fuentes y la correcta representación en mapas y visualizaciones, además de prevenir errores en la interpretación de la ubicación.

- Mejores prácticas asociadas
  1. *Utilizar el sistema de referencia de coordenadas ETRS89 y asegurarse de que las longitudes, latitudes, y coordenadas X e Y estén representadas en formato decimal. Esto garantizará la consistencia y facilitará el análisis geoespacial.*

- Ejemplos

  El fichero "ContenedoresRopa.xlsx" del dataset "Contenedores de ropa" no tiene ningún valor fuera de rango. El rango de valores normal para Madrid ronda el -3,xxx y el 40,xxx en longitud y latitud, respectivamente. Este rango de valores viene dado por https://www.ign.es/web/coordenadas-de-estaciones-ergnss. En el anexo al final del documento se indica valores normales para la longitud y latitud.

  Este es un ejemplo de buena práctica en cuanto al criterio, pero en cuanto al uso de punto como separador decimal sería una práctica mejorable, de acuerdo al criterio anterior:

LATITUD	LONGITUD
40.38516785	-3.68677717
40.38785956	-3.68775987
40.38795031	-3.69021968
40.38730993	-3.69184964
40.3854922	-3.69214483

Con respecto a utilizar el punto o la coma, como separador de decimales en las coordenadas, **no hay unanimidad**. Depende de la configuración de las aplicaciones que utilicen estas coordenadas, y actualmente hay algunas que interpretan solo correctamente el punto y otras que interpretan solo correctamente la coma. Por ejemplo, Google Maps, prefiere como separador de decimales el punto.

- Riesgos

1. *Coordenadas mal representadas pueden llevar a errores en la ubicación geográfica, afectando la precisión de análisis geoespaciales y la visualización en mapas.*
2. *El uso de sistemas de referencia o formatos incorrectos puede dificultar la integración de datos de diferentes fuentes, limitando la capacidad para combinar y analizar información geográfica de manera coherente.*

3. *La falta de un formato estándar y coherente puede dificultar el uso de los datos en diferentes plataformas y aplicaciones, especialmente aquellas que requieren un formato específico para las coordenadas.*

3.3.9 Criterio 9 – Las coordenadas X e Y, correctamente representadas

- Definición

  Es fundamental que las coordenadas X e Y (plano bidimensional) estén correctamente representadas y se ajusten al sistema de coordenadas esperado, ya que esto garantiza la precisión y la interpretación adecuada de la ubicación espacial de los datos. Si las coordenadas se expresan en metros o milímetros, es esencial que esta unidad se indique claramente para evitar ambigüedades en la interpretación de los datos. Además, si se utilizan decimales para representar fracciones de una unidad, es importante que se utilice la coma (",") como separador decimal, especialmente en contextos donde este es el estándar utilizado.

  Las coordenadas de la ciudad de Madrid deben estar entre estos dos rangos:

  X: tiene que estar entre las cifras 433000 a 391000 en metros.

  Y: tiene que estar entre las cifras 4960000 a 4910000 en metros.

  Si no estuviese entre esos rangos, probablemente estaría fuera del término de Madrid.

  En anexo al final del documento, se explica un poco más estos rangos.

- Mejores prácticas asociadas
  1. *Verificar que las coordenadas X e Y estén correctamente representadas y se ajusten al sistema de coordenadas esperado. Esto asegurará la precisión y la coherencia espacial de los datos.*
  2. *Indicar claramente si las coordenadas se presentan en metros o milímetros para evitar confusiones. Esto se puede hacer mediante la inclusión de una etiqueta explícita junto a las coordenadas (por ejemplo, "Coordenadas en metros" o "Coordenadas en milímetros").*

- Ejemplos

  Cualquier fichero con coordenadas X e Y debe tener valores entre esos rangos.

- Riesgos

1. *La incorrecta representación de las coordenadas X e Y puede resultar en errores de ubicación, afectando la precisión de los datos geoespaciales y llevando a interpretaciones incorrectas sobre la posición de puntos específicos.*

2. *La falta de coherencia en el sistema de coordenadas utilizado puede complicar la integración de datos de diferentes fuentes.*
3. *Sistemas y aplicaciones que dependen de la exactitud de las coordenadas para realizar funciones como la visualización en mapas o el análisis espacial pueden fallar o producir resultados incorrectos si las coordenadas no están correctamente representadas.*

3.3.10 Criterio 10 - Decimales representados con coma en números.

- Definición

  Utilizar la coma como separador decimal, es decir, para separar la parte entera de la parte decimal de un valor numérico. Es fundamental para garantizar la coherencia y consistencia en los datos, así como para prevenir errores de cálculo.

- Mejores prácticas asociadas
  1. *Utilizar la coma como separador decimal en lugar de puntos u otros caracteres. Mantener la consistencia en la representación de los valores decimales en todo el conjunto de datos.*

- Ejemplos
  1. *La columna del fichero de 2015 "modificaciones inscritas 2015.xls" del dataset "Actividad Contractual" es incorrecta. Hay un ejemplo marcado en naranja de cómo se debe utilizar la coma como separador decimal:*

IMPORTE DE LA MODIFICACIÓN
0.00 €
57,957.14 €
6229,67
-67,115.43 €

- Riesgos

1. *El uso inconsistente de separadores decimales puede llevar a dificultades en la interpretación y manipulación de los datos.*
2. *Sistemas y aplicaciones que esperan un formato específico pueden interpretar incorrectamente los valores decimales, lo que puede resultar en cálculos incorrectos y análisis erróneos.*
3. *La falta de un separador decimal claro y consistente puede causar confusión, especialmente en contextos internacionales donde las convenciones de separadores decimales varían (punto vs. coma).*

3.3.11 Criterio 11 - No se deben utilizar caracteres de formato de “miles”.

- Definición
  El uso de caracteres de formato de "miles" se refiere a la representación de grandes cantidades numéricas con un punto o una coma como separador. La inclusión de separadores puede generar ambigüedad y errores de interpretación y cálculo.

- Mejores prácticas asociadas
  1. *Evitar el uso de caracteres como puntos o comas para separar miles en números. Mantener la coherencia en el formato numérico utilizado en todo el conjunto de datos.*

- Ejemplos
  1. *El fichero “centroproteccionanimal.xls” del dataset “Madrid Salud. Protección animal” no utiliza puntos ni comas para separar miles. El campo ratios colonias utiliza el punto como separador decimal. Esto es una práctica mejorable.*

2021			
Total 2021	Acumulado total 2016/21	Porcentaje sobre total 2016/21	Ratio colonias (10.000h) 2016/21
193	1826	100.0%	5.49

3.3.12 Criterio 12 - No se deben incluir ceros a la izquierda.

- Definición
  Los ceros a la izquierda son ceros que preceden a un número en su representación, hablando de campos numéricos. Es importante no incluirlos para evitar errores y confusiones, especialmente entre sistemas que pueden interpretarlos como valores octales o en otras bases numéricas.

- Mejores prácticas asociadas
  1. *Eliminar cualquier cero no significativo a la izquierda de los números. Asegurarse de que los números se representen de manera consistente y sigan un formato numérico estándar.*

- Ejemplos
  1. *Un ejemplo de práctica a mejorar, teniendo ceros a la izquierda, sería “0198”. En el caso de que esa celda fuese formato texto, no sería lo mismo que si se tiene “198”.*
- Riesgos

1. *Algunos sistemas pueden interpretar números con ceros a la izquierda como valores en una base numérica diferente, como octal, lo que puede llevar a errores de cálculo y procesamiento.*
2. *La inclusión de ceros a la izquierda puede crear inconsistencias en la representación de los datos, dificultando la comparación y el análisis de los mismos.*
3. *Diferentes aplicaciones y sistemas pueden manejar los ceros a la izquierda de manera diferente, lo que puede complicar la integración de datos y su uso en múltiples plataformas.*

3.3.13 Criterio 13 - Las unidades de medida y monedas deben indicarse por separado, o en el nombre de las columnas.

- Definición
  *La correcta indicación de qué **unidades de medida** y moneda se están utilizando evita ambigüedad, posible malinterpretación de los datos, y, por lo tanto, errores en los cálculos.*

- Mejores prácticas asociadas
  1. *Especificar claramente las unidades de medida y las monedas utilizadas en los datos. Evitar ambigüedades proporcionando etiquetas o metadatos claros para las columnas que contienen información de unidades o monedas.*

- Ejemplos
  1. *El fichero "EstadoParquesHistoricoSingularesForestales2017.xlsx" del dataset "Arbolado en parques y zonas verdes de Madrid" tiene correctamente especificadas las unidades en su documento de estructura:*

Contenido XLS Nombre	Tipo	Descripción
PARQUE	Varchar	Denominación del parque.
Altura Promedio (m)	Numérico con decimales	Se obtiene como media aritmética de todos los ejemplares arbóreos de cada parque integrado en los parques históricos, singulares y forestales de Madrid
Perimetro Promedio (cm)	Numérico con decimales	Se obtiene como media aritmética de todos los ejemplares arbóreos de cada parque integrado en los parques históricos, singulares y forestales de Madrid
Recien plantado y no consolidado	Numéricos	Árboles de 1 a 5 años desde que se realizó su plantación
Joven	Numéricos	Árbol que todavía no ha alcanzado su máximo desarrollo
Maduro	Numéricos	Árbol en pleno vigor, árbol que ha alcanzado su tamaño máximo
Viejo	Numéricos	Árbol en regresión
Otros	Numéricos	Agrupa a árboles decrépitos, muertos, tocones y con edad fenológica sin definir
Total general	Numéricos	Suma de cada tipo de árboles según su edad fenológica y distrito.

- Riesgos

1. *La falta de especificación de las unidades de medida y las monedas puede llevar a una interpretación errónea de los datos, ya que los usuarios pueden asumir incorrectamente las unidades o la moneda utilizada.*
2. *La omisión de las unidades de medida o de la moneda puede resultar en errores en los cálculos, especialmente si los datos se utilizan en operaciones matemáticas o financieras.*
3. *Sin una indicación clara de las unidades de medida o la moneda, puede ser difícil comparar y analizar los datos, lo que afecta la calidad y utilidad de la información.*

3.3.14 Criterio 14 - Valores de distribución de cada columna coherentes con la serie anterior

- Definición

  Se entiende por coherencia en los valores de distribución de cada columna con respecto a la serie anterior a la relación lógica y consistente entre los datos de cada columna del conjunto de datos que se está analizando con los datos de las mismas columnas en la serie anterior. Se analiza esta relación lógica estudiando posibles variaciones y patrones, y su alineación con la estructura y contexto general de los datos. Este estudio es esencial

para garantizar la precisión e integridad de los datos, aportando al usuario confianza en las conclusiones que extraiga de su análisis.

- Mejores prácticas asociadas

1. *Verificar que el número de filas en los nuevos conjuntos de datos sea igual o similar al número de filas en la serie anterior. Esto garantizará que no se haya perdido o agregado información incorrectamente.*
2. *Realizar una verificación de consistencia en los valores de las columnas, asegurándose de que estén dentro del rango esperado y sean coherentes con la serie anterior. Esto ayudará a identificar posibles errores o discrepancias en los datos.*
3. *Verificar que los valores máximos en cada columna no excedan los límites establecidos en la serie anterior. Esto ayudará a identificar valores atípicos o errores en los datos.*
4. *Verificar que los valores mínimos en cada columna no sean inferiores a los límites establecidos en la serie anterior. Esto ayudará a identificar posibles errores o discrepancias en los datos.*
5. *Verificar que no haya duplicados en los valores únicos de cada columna. Comprobar si hay nuevos valores únicos que no estaban presentes en la serie anterior. Si es así, investigar y comprender la razón detrás de la aparición de estos nuevos valores.*
6. *Analizar cualquier cambio drástico en la frecuencia del valor más común y determinar si hay razones válidas para ese cambio. Verificar si la frecuencia del valor más común sigue siendo consistente con la serie anterior.*
7. *Comprobar si hay un aumento o disminución significativa en el valor de la desviación estándar y verificar si el valor de la desviación estándar sigue siendo coherente con la variabilidad de la serie anterior.*
8. *Verificar si hay cambios drásticos en el valor de la media y validar si la media se encuentra dentro de un rango esperado y coherente con la serie anterior.*
9. *Comprobar si los tipos de frecuencia asignados (+ para las más frecuentes y - para las menos frecuentes) se corresponden con las expectativas y la serie anterior. Validar si la asignación de frecuencias es coherente y sigue una lógica adecuada en función de los datos y las tendencias observadas.*

- Ejemplos

2. *El número de filas entre series en el dataset "Aparcamientos (PAR) - valor de las plazas" es similar:*

PATH FICHERO	NUM FILAS
300532 - Aparcamientos (PAR) - valor de las plazas\2022	93049
300532 - Aparcamientos (PAR) - valor de las plazas\2023	93302

3. *Los valores entre columnas en el dataset "Presupuestos. Ejecución mensual ejercicio 2023" están dentro de la variación esperada de un mes al siguiente, en este caso de abril a mayo. La media mantiene su tendencia. Lo mismo ocurre con la desviación estándar "std":*

NOMBRE FICHERO	PATH FICHERO	COLUMNA	mean	std
Ejecucion_2023_04_gastos_001_OOAA.csv	300618 - Presupuestos 2023\Abril 2023	centro	37.97617	131.4683
Ejecucion_2023_04_gastos_001_OOAA.csv	300618 - Presupuestos 2023\Abril 2023	capitulo	2.13895	1.636422
Ejecucion_2023_05_gastos_001_OOAA.csv	300618 - Presupuestos 2023\Mayo 2023	centro	39.74968	134.309
Ejecucion_2023_05_gastos_001_OOAA.csv	300618 - Presupuestos 2023\Mayo 2023	capitulo	2.161406	1.655842

4. *Los valores máximos y mínimos no exceden la serie anterior; coinciden de un mes al siguiente:*

NOMBRE FICHERO	PATH FICHERO	COLUMNA	mean	min	max
Ejecucion_2023_04_gastos_001_OOAA.csv	300618 - Presupuestos 2023\Abril 2023	centro	37.97617	1	509
Ejecucion_2023_04_gastos_001_OOAA.csv	300618 - Presupuestos 2023\Abril 2023	capitulo	2.13895	1	9
Ejecucion_2023_05_gastos_001_OOAA.csv	300618 - Presupuestos 2023\Mayo 2023	centro	39.74968	1	509
Ejecucion_2023_05_gastos_001_OOAA.csv	300618 - Presupuestos 2023\Mayo 2023	capitulo	2.161406	1	9

5. *En el dataset "Convenios", el valor más frecuente disminuye en un 21% de 2022 a 2023. Tal vez resulte de interés analizar por qué ocurre esto.*

NOMBRE FICHERO	columna	valor	frecuencia	tipo
ConveniosTransparencia_2022.xlsx	Área de Gobierno o Distrito	Área de Gobierno de Cultura, Turismo y Depo	224	+
ConveniosTransparencia_2023.xlsx	Área de Gobierno o Distrito	Área de Gobierno de Cultura, Turismo y Depo	86	+

6. *Para el fichero anterior, se comprueba que los tipos de frecuencia asignados se mantienen de una serie a otra:*

PATH FICHERO	columna	valor	tipo
300295 - Convenios\2022	Área de Gobierno o Distrito	Área de Gobierno de Cultura, Turismo y Deporte	+
300295 - Convenios\2022	Área de Gobierno o Distrito	Área de Gobierno de Familias, Igualdad y Bienestar Social	+
300295 - Convenios\2023	Área de Gobierno o Distrito	Área de Gobierno de Familias, Igualdad y Bienestar Social	+
300295 - Convenios\2023	Área de Gobierno o Distrito	Área de Gobierno de Cultura, Turismo y Deporte	+
300295 - Convenios\2022	Área de Gobierno o Distrito	Distrito Fuencarral-El Pardo	-
300295 - Convenios\2023	Área de Gobierno o Distrito	Distrito Fuencarral-El Pardo	-

- Riesgos

1. *Si el número de filas en los nuevos conjuntos de datos no coincide con la serie anterior, podría indicar pérdida o agregación incorrecta de información.*
2. *Si los valores mínimos y máximos en cada columna no están dentro de los límites esperados, podría indicar la presencia de valores atípicos o errores en los datos, lo que afectaría la precisión de los análisis.*
3. *Si no se detectan duplicados, nuevos valores únicos o cambios drásticos en la frecuencia de los valores más comunes, se podría pasar por alto información importante que afecta la interpretación de los datos.*
4. *La presencia de cambios drásticos en la desviación estándar o la media podría afectar la precisión de los cálculos y análisis realizados con los datos.*

3.3.15 Criterio 14 - Confidencialidad y anonimización de los datos

No se pueden publicar datos identificativos de una persona directamente o que permitan la identificación de una manera indirecta. Solo se pueden publicar datos personales, cuando alguna ley habilite a ello o se tenga autorización expresa de la persona implicada.

Siempre que detrás de los datos publicados existan datos personales, aunque no se publiquen las columnas identificativas **hay que hacer un análisis de protección de datos** para garantizar que a partir de la información publicada e incluso otras fuentes públicas, no se pueda llegar a identificar a personas.

- Definición
  La confidencialidad y anonimización de los datos se refieren a la protección de la identidad y la información personal de los individuos en conjuntos de datos. Se lleva a cabo a través de diferentes prácticas como enmascaramiento, codificación u omisión de datos que vinculen a personas físicas, poniéndolas en riesgo de discriminación o robo de identidad, entre otros. El uso y tratamiento de este tipo de datos sensibles se regula en Europa por el Reglamento General de Protección de Datos (GDPR).

- Mejores prácticas asociadas
  1. *Garantizar la confidencialidad y privacidad de los datos sensibles. Aplicar técnicas adecuadas de anonimización de datos para proteger la identidad de las personas o entidades involucradas. No se puede publicar nada si no hay una norma o ley que lo habilite.* ***Tampoco se pueden publicar campos que puedan identificar a una persona****.*

- Riesgos

1. *Publicar datos identificativos de personas directa o indirectamente puede violar su privacidad y exponer información confidencial sin su consentimiento.*
2. *La divulgación de información personal puede llevar a la discriminación o estigmatización de las personas afectadas, especialmente si se revelan detalles sensibles sobre su salud, orientación sexual, religión, origen étnico, etc.*
3. *No cumplir con las regulaciones de protección de datos, como el Reglamento General de Protección de Datos (GDPR) en Europa, puede resultar en multas significativas y daños a la reputación de la entidad responsable.*

3.3.16 Criterio 15 – Dato único. Consistencia entre datos del portal de Madrid y fuentes externas (Madrid.es y Banco de Datos de Estadística)

- Definición
  La armonía y congruencia en los valores, definiciones y estructura de los datos compartidos entre las diferentes fuentes o webs municipales son clave para la confiabilidad, uniformidad y consistencia de los datos por parte de los usuarios. Esta consistencia garantiza que los análisis realizados y las conclusiones obtenidas están bien fundados. Una buena filosofía a seguir es la del dato único, que se basa en la introducción de la información una sola vez y en su reaprovechamiento o

retroalimentación en múltiples procesos, sin incurrir en posibles errores o en duplicidad de los datos.

- Mejores prácticas asociadas
  1. *Asegurar la coherencia y consistencia entre los datos del portal del Ayuntamiento de Madrid y otras fuentes municipales (como la web www.Madrid.es, el Banco de Datos de Estadística del Ayto. de Madrid, el Geoportal, etc.). Si se detectan cualquier discrepancia o inconsistencia entre los conjuntos de datos o datos ofrecidos por las diferentes webs municipales, hay que tomar medidas para corregirlas o aclararlas.*

- Ejemplos

  Los datos de 2023 del dataset "Aparcamientos municipales. Número de plazas" entre datos abiertos y el portal del ayuntamiento de Madrid son consistentes y coherentes:

ID_DATASET	NOMBRE DATASET	URL OpenData Madrid	URL Madrid.es	Atributo analizado	Fecha_OpenData_Madrid	Fecha_Madrid.es	Valor OpenData Madrid	Valor Madrid.es
300542	Aparcamientos municipales. Número de	https://datos.madrid.es/s	Aparcamiento mixto. Alcá	PARplazasRotacion	NA	NA	121	121

  Se pueden consultar en los links:
  1. *Portal del Datos Abiertos:*
     https://datos.madrid.es/sites/v/index.jsp?vgnextoid=8edd9863d48bd710VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

  2. *Madrid.es:*
     https://www.madrid.es/portales/munimadrid/es/Inicio/Movilidad-y-transportes/Aparcamiento-mixto-Alcantara/?vgnextfmt=default&vgnextoid=79df252aa267d710VgnVCM1000001d4a900aRCRD&vgnextchannel=220e31d3b28fe410VgnVCM1000000b205a0aRCRD

- Riesgos

1. *Si los datos proporcionados en el portal de Madrid no coinciden con los datos de otras fuentes municipales, los usuarios pueden perder la confianza en la precisión y la fiabilidad de la información presentada.*
2. *Si los datos son inconsistentes entre diferentes fuentes, los análisis realizados pueden conducir a conclusiones incorrectas o incompletas.*
3. *La duplicación o discrepancia de datos entre diferentes fuentes puede llevar a la ineficiencia en el uso de los recursos municipales.*

4 GUÍAS EXTERNAS

4.1 Guías de la iniciativa Ciudades Abiertas

La iniciativa Ciudades Abiertas es un convenio de ayuntamientos en España para la colaboración, reutilización e interoperabilidad de portales públicos de datos abiertos. Consiste en la implementación de políticas y tecnologías que facilitan el acceso a la información pública, fomentan la colaboración entre el gobierno y los ciudadanos, y promueven la innovación en el ámbito urbano. Esta iniciativa busca crear entornos más inclusivos, eficientes y sostenibles, donde los ciudadanos puedan involucrarse activamente en la toma de decisiones y contribuir al desarrollo de sus comunidades.

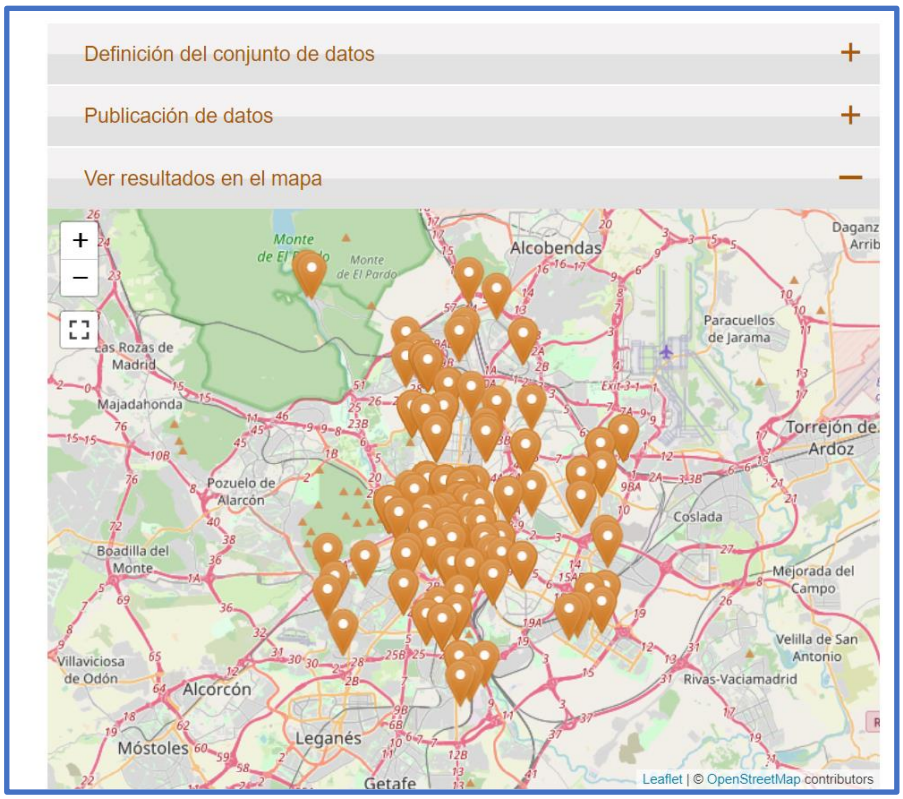
Dicha iniciativa posee algunos documentos con orientaciones sobre cómo mejorar la calidad de los datos abiertos. El más completo es el "*Adaptación de estructuras de conjuntos de datos para asegurar su calidad y anonimización*", que además posee el "*Checklist para la comprobación de criterios de calidad aplicables a distribuciones de conjuntos de datos*", el cual ha servido de guía para la mayoría de los controles de calidad realizados durante el proyecto.

A continuación, se destacan otros criterios de calidad interesantes indicados en el checklist de Ciudades Abiertas, pero que no han sido mencionados anteriormente en este documento por no haber sido analizados en el proyecto:

FORMATOS:

- **GEO-2. En el caso de representar polígonos u otras formas geométricas, se recomienda usar el formato WKT**. El formato WKT es ampliamente utilizado por muchas aplicaciones para la representación de polígonos y otras formas geométricas (por ejemplo, para representar la parcela de un edificio público. Es el formato que se debería usar en caso de que sea necesario exportar este tipo de información, que normalmente está ya presente en formatos como shapefiles o GeoJSON.

Ejemplo dataset "*Actividades Culturales y de Ocio Municipal en los próximos 100 días*":

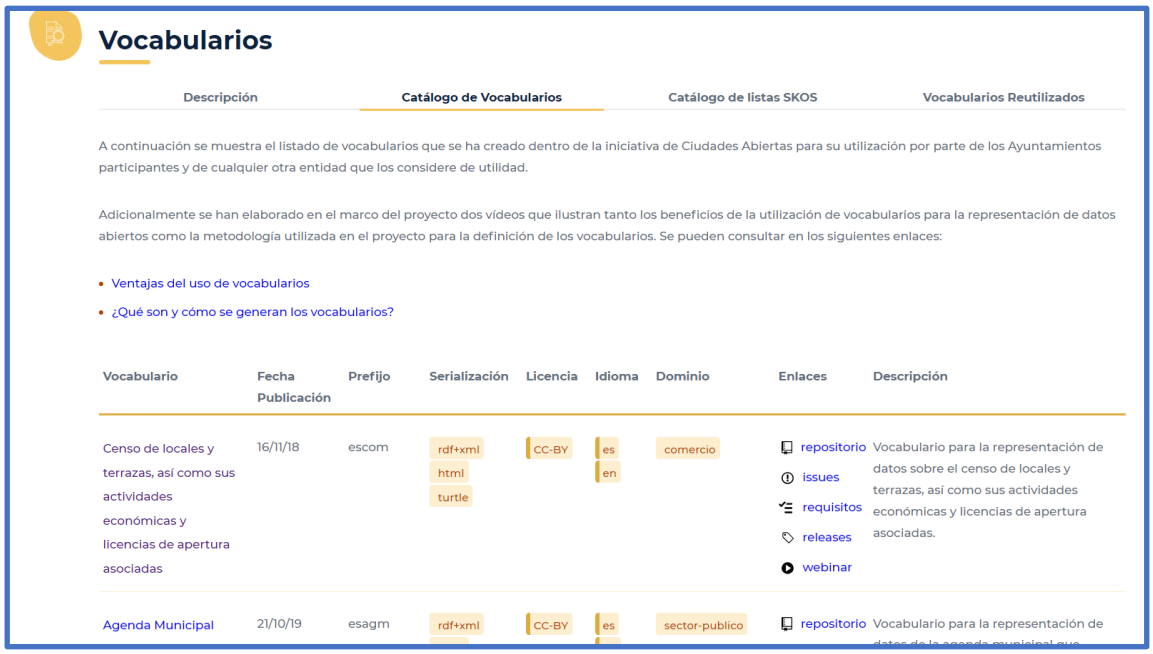


DATOS

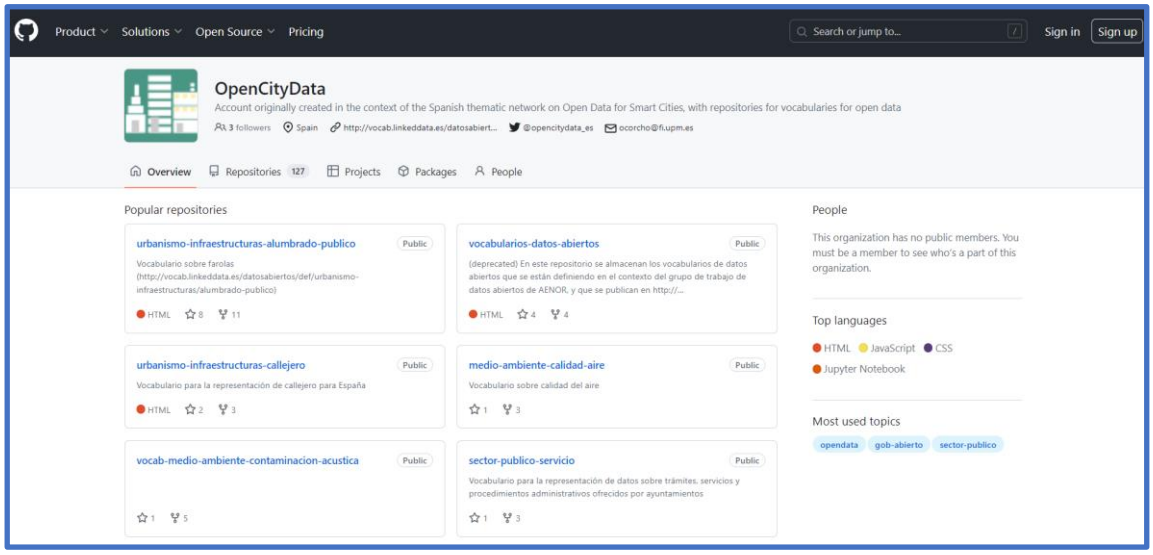
- **ESTAD-5. Siempre que sea posible, utilizar valores en los campos que estén estandarizados.** Siempre que los valores pertenezcan a un conjunto controlado de valores, es útil utilizar los códigos correspondientes a vocabularios, tesauros o listas de códigos SKOS correspondientes, pues esto puede facilitar la reutilización, así como la ausencia de errores de codificación, etc. Esta propuesta también se recoge en la guía de datos.gob.es, como pautas P8 y P9. Por ejemplo, en el ejemplo anterior esto podría ser aplicable a las categorías de locales comerciales, en caso de estar identificadas por códigos, como por ejemplo los códigos CNAE o los códigos de actividad económica de un ayuntamiento concreto.

Ejemplos de vocabularios:

1. Ciudades Abiertas - https://ciudades-abiertas.es/vocabularios/



2. Github OpenCityData - https://github.com/opencitydata



- **FECHA-3. Si no se tiene información completa sobre una fecha, no debe representarse como tal, sino en diversas columnas**. Esto ocurre, por ejemplo, cuando para algún conjunto de datos no se tiene el día del mes, o un valor se refiere a un trimestre del año, o sólo se tiene el año. En este caso, la representación de estos datos debe dividirse en tantas columnas como información de la que se disponga (por ejemplo, año y mes, año y trimestre, año simplemente, etc.), en lugar de en una única columna, donde inevitablemente habría que incluir datos que potencialmente son erróneos. En guías

como la de datos.gob.es se propone que en el caso de que falte algo de información esta sea rellenada por defecto, pero no se considera una buena práctica (por ejemplo, se propone que, si solo se dispone de un dato mensual, la mejor opción es incluir una fecha completa ajustada al último día del mes - por ejemplo, para septiembre, 2019-09-30).

Mes	Año
1	2023
2	2023
3	2023
4	2023
5	2023
6	2023
7	2023
8	2023
9	2023
10	2023
11	2023
12	2023

Fecha
ene-23
feb-23
mar-23
abr-23
may-23
jun-23
jul-23
ago-23
sep-23
oct-23
nov-23
dic-23

Además, el documento también realiza un análisis sobre los metadatos de algunos datasets del Ayuntamiento de Madrid, puntúa la falta de estandarización en el llamado 'Documento de Estructura' en sus datasets y propone una estructura para tal, haciendo referencia a cada elemento del DCAT (lo que facilita la definición de los vocabularios para cada dataset).

4.2 Guías de la iniciativa Aporta (Datos.Gob)

La Iniciativa Aporta es una iniciativa promovida por el Gobierno de España con el objetivo de fomentar la apertura y reutilización de la información del sector público. Aporta busca establecer un marco común para la apertura de datos en todas las administraciones públicas del país, con el fin de garantizar la transparencia y la participación ciudadana en la gestión pública. A través de esta iniciativa, se promueve la adopción de estándares técnicos y jurídicos para la apertura de datos, así como la colaboración entre las diferentes entidades públicas.

Además, Aporta trabaja en la sensibilización y formación de los empleados públicos, promoviendo la cultura de la apertura y la reutilización de la información como un elemento clave para la mejora de la eficiencia y la calidad de los servicios públicos.

Aporta también posee diversos documentos con orientaciones acerca de la publicación de datos abiertos. Uno de los más completos y que, juntamente con el checklist de ciudades abiertas, ha servido de base para la definición de los criterios para el proyecto, ha sido el documento *"Guía práctica para la mejora de la calidad de datos abiertos"*.

A continuación, se destacan otros criterios de calidad interesantes indicados en la guía, pero que no han sido mencionados en apartados anteriores por no haber sido analizados en el proyecto:

METADATOS

- **Utilizar vocabularios controlados siempre que sea posible -** En el contexto de los datos enlazados la reutilización de los vocabularios controlados existentes es un requisito para garantizar la interoperabilidad de datos y deben utilizarse siempre que sea posible. La Oficina de Publicaciones de la Unión Europea proporciona numerosos vocabularios para su uso en el ámbito de las administraciones públicas. DCAT-AP es el vocabulario controlado que se utiliza para la interoperabilidad de catálogos de datos abiertos en Europa. Otras referencias útiles son los vocabularios aplicables en contexto de ciudades abiertas o para un propósito más general, los relacionados en el sitio web Linked Open Vocabularies (LOD). Ejemplo: En el ejemplo se observan dos formas -validas sintácticamente en ambos casos- pero en el segundo caso el uso del vocabulario controlado "license" es más eficiente.

  Ejemplo extraído de la Guía:

```
<dct:accrualPeriodicity>
    <dct:Frequency rdf:about="http://dataset/Frequency">
        <rdf:value>
            <time:DurationDescription rdf:about="htpp://dataset/DurationDescription">
                <time:years rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal"1</time:years>
                        </time:DurationDescription>
        </rdf:value>
    </dct:Frequency>
</dct:accrualPeriodicity>
```

```
<dct:accrualPeriodicity>
        <dct:Frequency rdf:about="http://publications.europa.eu/resource/authority/frequency/DAILY"/>
</dct:accrualPeriodicity>
```

FORMATOS

- **Evitar formatos de datos no procesables -** Cualquier informe, mapa, gráfico, infografía, tabla, o cualquier otra representación visual elaborada a partir de datos, debe ir siempre acompañada de la serie de archivos en formato abierto y reutilizable que faciliten el acceso a los datos en los que se basan o a los que hacen referencia en dicha representación Los publicadores deben proporcionar los datos en un número razonable de formatos alternativos.

  Es recomendable, además, dar preferencia a aquellos formatos con mayor nivel de compatibilidad (por ejemplo, CSV frente a XLS), pero sin relegar a otros formatos populares entre los usuarios o adecuados para determinados tipos de datos, por ejemplo, el formato SHP para datos espaciales.

  Así mismo, siempre que se dispongan datos de alto valor, dinámicos o con alta frecuencia de actualización, además de la posibilidad de descarga masiva, cuando proceda, se deben proporcionar a través de interfaces de programación de aplicaciones (APIs) o de puntos de consulta SPARQL. Con ello, los publicadores conseguirán abrir los datos a un rango más amplio de reutilizadores facilitando, además, la utilización de éstos por maquinas u otras aplicaciones implementando así soluciones basadas en datos más sofisticadas.

DATOS

- **Estandarizar valores de datos -** La duplicidad de los datos es un problema grave y en ocasiones difícil de detectar, que se debe atajar para incrementar la calidad de los datos y garantizar su confiabilidad. Para reducir el número de duplicados es aconsejable estandarizar la recogida de datos y almacenamiento de los mismos, centralizando el proceso en un único sistema de información, de tal forma que sean fácilmente detectables y puedan ser eliminados automáticamente.

Ejemplo extraído de la Guía:

year; visitors; viewing-time
2014; 768954;00:03:18
2013;822101;00:02:59
2013;822101;00:02:59
2011;707402;00:03:44
2010;707402;00:03:50
2010;707402;00:03:50

year; visitors; viewing-time
2014; 768954;00:03:18
2013;822101;00:02:59
2012;792967;00:02:52
2011;707402;00:03:44
2010;707402;00:03:50
2009;429430;00:03:16

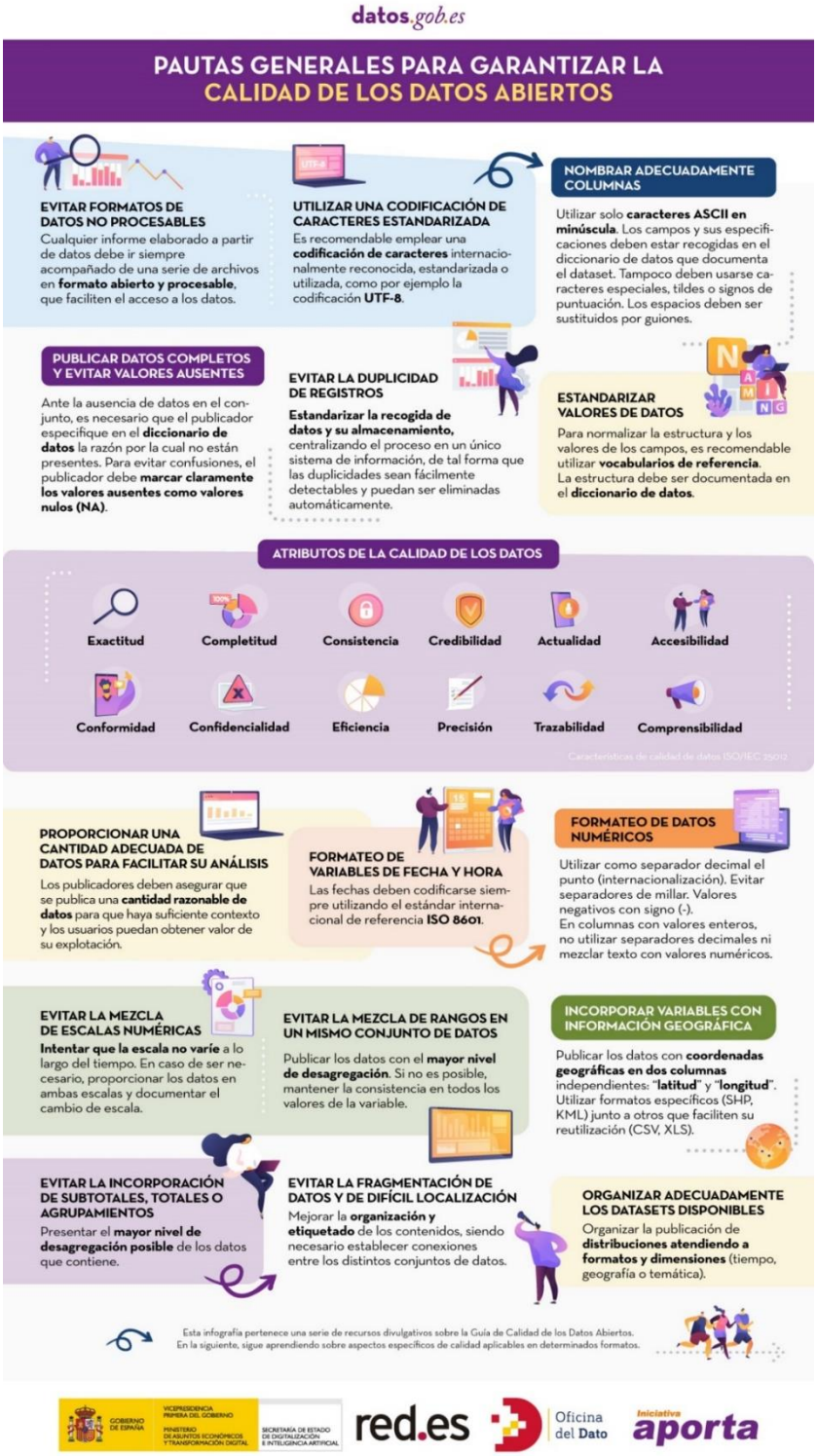
- **Evitar la mezcla de rangos en un mismo conjunto de datos -** Es aconsejable que los conjuntos de datos se publiquen con el mayor nivel de desagregación posible evitando el uso de rangos de datos, facilitando a los usuarios el mayor detalle de toda la información. Si no es posible representar los datos con el máximo nivel de desagregación es imprescindible mantener la consistencia en todos los valores de la variable y evitar la combinación de texto e información numérica añadiendo columnas adicionales que representen el valor inicial y final de rango.

  Ejemplo extraído de la Guía:

marca	año	consumo	ventas	potencia	aceleracion
ford torino	1970	Alto	2.50	De 100 a 150	12
buick skylark 320	1970	Medio	2.63	De 150 a 200	11.5
plymouth satellite	1970	Medio	2.37	De 150 a 200	11
chrevrolet chevelle malibu	1970	Bajo	2.40	Mas de 200	13

marca	año	consumo	ventas	potencia	aceleracion
ford torino	1970	Alto	2.50	130	12
buick skylark 320	1970	Medio	2.63	165	11.5
plymouth satellite	1970	Medio	2.37	150	11
chrevrolet chevelle malibu	1970	Bajo	2.40	210	13

A continuación, se puede visualizar un infográfico con los principales criterios (pautas) para garantizar la calidad de los datos abiertos, indicado por la guía de datos.gob:



4.3 Normas UNE de la Gestión del Dato y de la Calidad del Dato

Las normas UNE de gestión y calidad del dato son un conjunto de estándares establecidos por la Asociación Española de Normalización (UNE) que se enfocan en proporcionar directrices y mejores prácticas para garantizar la gestión eficaz y la calidad de los datos en las organizaciones. Estas normas buscan asegurar que los datos sean precisos, completos, consistentes, accesibles y oportunos, lo que es fundamental para la toma de decisiones informadas y el funcionamiento eficiente basadas en datos.

Las normas UNE de gestión y calidad del dato abarcan diferentes aspectos relacionados con la gestión de datos, incluyendo la planificación estratégica de la gestión de datos, la definición de políticas y procedimientos, la recolección y almacenamiento de datos, la seguridad y privacidad de la información, así como la evaluación y mejora continua de la calidad de los datos.

Estas normas proporcionan un marco de referencia para que las organizaciones establezcan sistemas de gestión de datos sólidos y efectivos, lo que les permite maximizar el valor de sus activos de datos y minimizar los riesgos asociados con la mala gestión de la información. Además, al adoptar las reglas UNE de gestión y calidad del dato, las organizaciones pueden mejorar la confianza de los clientes, cumplir con los requisitos regulatorios y aumentar su competitividad en el mercado.

A continuación, se destacan estrategias de calidad interesantes indicados en la guía, pero que no han sido mencionados en apartados anteriores por no haber sido analizados en el proyecto:

- **Definición de políticas y procedimientos -** Establecer políticas claras y procedimientos documentados para la gestión de datos, incluyendo la identificación de responsabilidades, roles y procesos para la recolección, almacenamiento, procesamiento y distribución de datos.

- **Calidad de los datos -** Implementar controles de calidad de datos para garantizar la precisión, integridad, consistencia y relevancia de la información. Esto puede incluir la validación de datos en tiempo real, la limpieza periódica de bases de datos y la corrección de errores.

- **Gestión de metadatos -** Desarrollar un sistema de gestión de metadatos que documente de manera exhaustiva la estructura, el significado y el contexto de los datos. Esto facilita la comprensión y el uso efectivo de la información por parte de los usuarios.

- **Seguridad de la información -** Implementar medidas de seguridad adecuadas para proteger los datos contra accesos no autorizados, alteraciones y pérdidas. Esto puede incluir el uso de firewalls, encriptación de datos, controles de acceso y auditorías de seguridad.

- **Privacidad de los datos -** Cumplir con la privacidad de datos aplicables y proteger la información confidencial según las leyes y estándares relevantes. Esto puede implicar la anonimización de datos sensibles y la obtención de consentimiento para el uso de información personal.

- **Gestión del ciclo de vida de los datos -** Establecer políticas y procedimientos para el ciclo de vida completo de los datos, desde su creación y captura hasta su archivado y eliminación. Esto garantiza la disponibilidad y relevancia de los datos cuando se necesitan, así como el cumplimiento de los requisitos de retención y eliminación.

- **Formación y concienciación -** formación regular a los empleados sobre las políticas y procedimientos de gestión de datos y las mejores prácticas de seguridad y privacidad. Fomentar una cultura organizacional que valore la importancia de la gestión y calidad del dato.

Las normas UNE proporcionan pautas claras sobre cómo estructurar y formatear los datos de manera consistente. Esto es especialmente relevante al trabajar con múltiples archivos Excel o CSV que contienen datos similares. Al seguir estas normas, se garantiza que los datos estén organizados de manera coherente, lo que facilita su análisis y uso por parte de diferentes usuarios. Además, se asegura que los datasets sean compatibles y puedan integrarse fácilmente con otros sistemas y herramientas. Esto es crucial cuando se necesita combinar datos de diferentes fuentes o compartir datasets con colaboradores externos. La adherencia a estas normas facilita la interoperabilidad y evita problemas de compatibilidad entre diferentes plataformas y aplicaciones.

Lo anterior implica la validación de datos, la detección y corrección de errores, y la eliminación de valores duplicados o incorrectos. Al seguir estas normas durante la creación y estructuración de datasets en Excel y CSV, se reduce la probabilidad de errores y se mejora la confiabilidad de los datos para su uso en análisis y toma de decisiones. Buenas prácticas concretas en la creación y estructuración de los datos para el portal de Datos Abiertos son los siguientes:

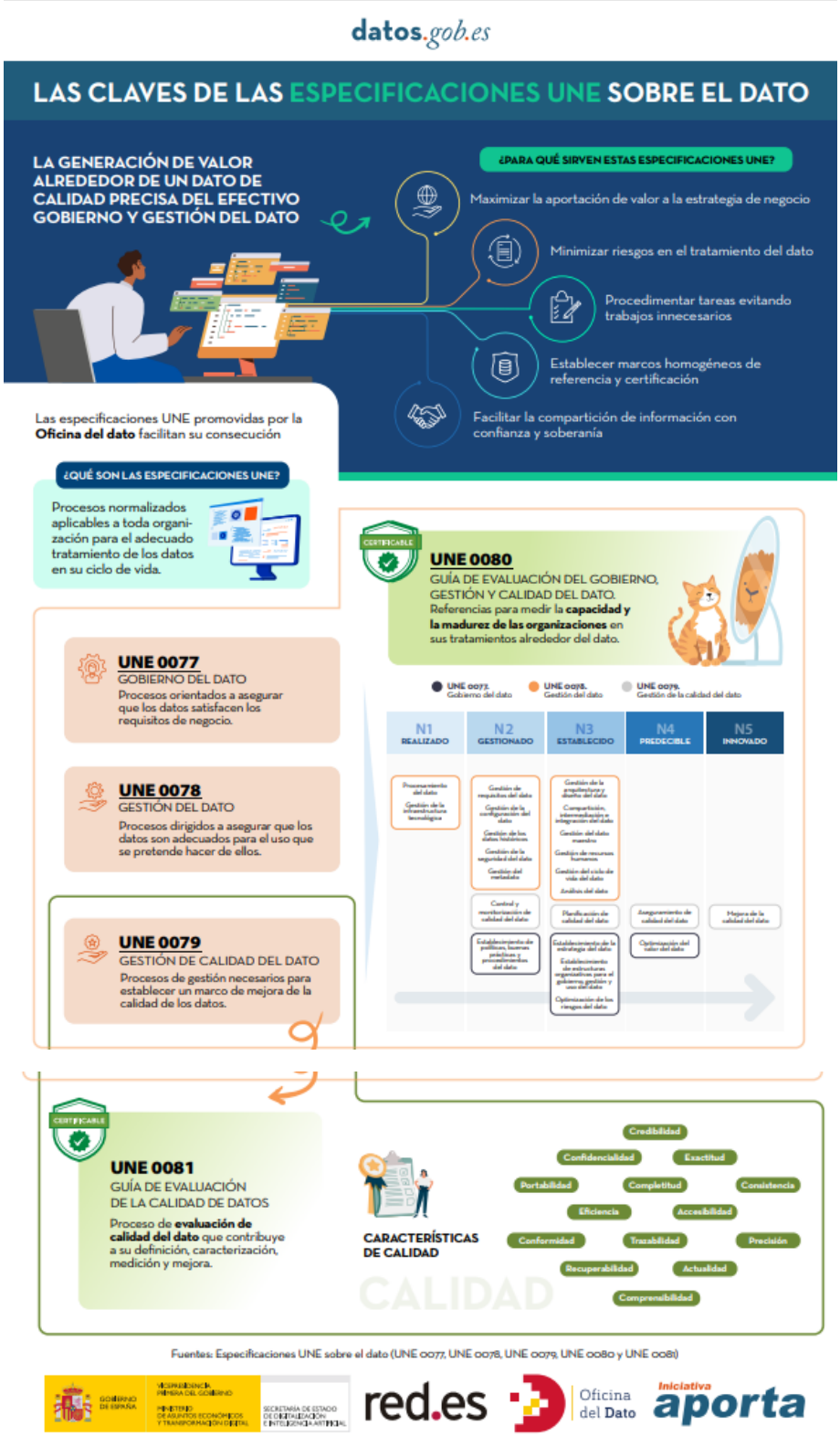
- **Estandarización de Formatos**: Al crear un dataset que combina datos de múltiples fuentes, es importante estandarizar los formatos de datos para asegurar la coherencia y facilidad de uso. Por ejemplo, en el dataset "Declaración de bienes y actividades de los directivos del Ayuntamiento de Madrid", se aplica las normas UNE para estandarizar el formato de fechas, en el caso para el apartado de fecha de inscripción, y monedas en los registros como el importe de deudas e información tributaria.

- **Limpieza y Normalización de Datos**: Antes de incluir datos en un dataset, es crucial realizar procesos de limpieza y normalización para corregir errores, eliminar duplicados y estandarizar la estructura de los datos. Por ejemplo, al recopilar datos de diferentes sistemas, aplicar las normas UNE para normalizar la capitalización de nombres, corregir errores de ortografía y estandarizar los formatos de direcciones.

- **Privacidad y Seguridad**: Las normas UNE también se aplican para garantizar la privacidad y seguridad de los datos en un dataset. Por ejemplo, al recopilar información personal, es importante aplicar las normas UNE relacionadas con la protección de datos para garantizar el cumplimiento de las regulaciones de privacidad, como el RGPD en la Unión Europea.

- **Documentación y Metadatos**: Es fundamental documentar adecuadamente un dataset y proporcionar metadatos claros que describan la fuente, el formato y el significado de los datos incluidos. Al aplicar las normas UNE, se puede seguir pautas específicas para documentar el dataset y los metadatos de manera que sean comprensibles y útiles para otros usuarios que puedan utilizar los datos.

- **Validación y Verificación**: Antes de publicar un dataset, es importante realizar pruebas de validación y verificación para asegurar la calidad y precisión de los datos. Al aplicar las normas UNE, se puede seguir procedimientos específicos para verificar la integridad de los datos, identificar posibles errores y garantizar que los datos sean confiables y precisos para su uso previsto.


datos.gob.es
LAS CLAVES DE LAS ESPECIFICACIONES UNE SOBRE EL DATO
LA GENERACIÓN DE VALOR ALREDEDOR DE UN DATO DE CALIDAD PRECISA DEL EFECTIVO GOBIERNO Y GESTIÓN DEL DATO
¿PARA QUÉ SIRVEN ESTAS ESPECIFICACIONES UNE?
Maximizar la aportación de valor a la estrategia de negocio
Minimizar riesgos en el tratamiento del dato
Procedimentar tareas evitando trabajos innecesarios
Establecer marcos homogéneos de referencia y certificación
Facilitar la compartición de información con confianza y soberanía
Las especificaciones UNE promovidas por la Oficina del dato facilitan su consecución
¿QUÉ SON LAS ESPECIFICACIONES UNE?
Procesos normalizados aplicables a toda organización para el adecuado tratamiento de los datos en su ciclo de vida.
UNE 0077
GOBIERNO DEL DATO
Procesos orientados a asegurar que los datos satisfacen los requisitos de negocio.
UNE 0078
GESTIÓN DEL DATO
Procesos dirigidos a asegurar que los datos son adecuados para el uso que se pretende hacer de ellos.
UNE 0079
GESTIÓN DE CALIDAD DEL DATO
Procesos de gestión necesarios para establecer un marco de mejora de la calidad de los datos.
CERTIFICABLE
UNE 0080
GUÍA DE EVALUACIÓN DEL GOBIERNO, GESTIÓN Y CALIDAD DEL DATO.
Referencias para medir la capacidad y la madurez de las organizaciones en sus tratamientos alrededor del dato.
UNE 0077. Gobierno del dato
UNE 0078. Gestión del dato
UNE 0079. Gestión de la calidad del dato
N1 REALIZADO
N2 GESTIONADO
N3 ESTABLECIDO
N4 PREDECIBLE
N5 INNOVADO
CERTIFICABLE
UNE 0081
GUÍA DE EVALUACIÓN DE LA CALIDAD DE DATOS
Proceso de evaluación de calidad del dato que contribuye a su definición, caracterización, medición y mejora.
CARACTERÍSTICAS DE CALIDAD
Credibilidad
Confidencialidad
Exactitud
Portabilidad
Completitud
Consistencia
Eficiencia
Accesibilidad
Conformidad
Trazabilidad
Precisión
Recuperabilidad
Actualidad
Comprensibilidad
CALIDAD
Fuentes: Especificaciones UNE sobre el dato (UNE 0077, UNE 0078, UNE 0079, UNE 0080 y UNE 0081)
red.es
Oficina del Dato
Iniciativa aporta

Anexo I: Documento de estructura

Versión mes(texto)/Año(NNNN)

Por ejemplo: Versión mayo/2024

estructura del conjunto de datos

Un documento de estructura es necesario en la mayoría de los casos, y sirve para poder explicar mejor los datos proporcionados o alguna de sus columnas, para que sean interpretados correctamente y así evitar posibles confusiones para las personas que no están familiarizados con esa información.

Nombre del conjunto de datos:

Descripción: En este conjunto de datos…

Unidad responsable:

Frecuencia de actualización:

Disponibilidad de los datos: (hay que indicar la fecha estimada de actualización)

Forma de obtener esa información o filtros aplicados: (si fuese necesario especificarlo)

Estructura del fichero de datos:

Número de columna	Nombre columna	Descripción	Valores esperados

Cuestiones a tener en cuenta:

- Si los datos tienen información de distrito y/o barrio, deben estar codificados correctamente con respecto a las tablas oficiales de distritos y barrios:
    - **Distritos** municipales de Madrid
    - **Barrios** municipales de Madrid
- Si los datos contienen coordenadas, tendrá que indicarse el sistema de referencia (Normalmente es ETRS89)
- Si un campo tiene una tipificación de valores posibles, hay que dar un anexo con el código de ese tipo de valor y su descripción (si no fuese autoexplicativo).

- Aparte de la tabla anterior se puede dar cualquier explicación que se considere oportuna, para esa información sea comprendida correctamente.

- Disponibilidad de los datos: se refiere a cuándo estarán disponibles los datos en el Portal de Datos Abiertos. Por ejemplo, si hablamos de información:

    - Anual: "Estará disponible en la segunda quincena de enero" o "estará disponible al final del primer trimestre" o "estará disponible al final del primer semestre" o cuando se estime que puede estar disponible.

    - Mensual: "Estará disponible sobre el día 15 del mes siguiente a su finalización" o "estará disponible mes y medio después a su finalización, con lo cual la información de marzo estaría disponible a mitad de mayo"

    - …

- Advertencias: cualquier posible advertencia que sea necesaria indicar.

- Forma de obtener esa información o filtros aplicados.

Unos ejemplos de ficheros de estructura que ya hay publicados podrían ser los siguientes:

- Accidentes de tráfico: https://datos.madrid.es/FWProjects/egob/Catalogo/Seguridad/Ficheros/Estructura_ConjuntoDatos_Accidentesv2.pdf

- Placas Stolpersteine: https://datos.madrid.es/FWProjects/egob/Catalogo/SociedadBienestar/Ficheros/Estructura_Placas_Stolpersteine.pdf

EJEMPLO

Se muestra un ejemplo de cómo podría ser la tabla de datos de un fichero de estructura.

Número de columna	Nombre columna	Descripción	Valores esperados
1	num_expediente	Número único que identifique al registro. Ejemplo: AAAASNNNNNN, donde: • AAAA es el año del accidente. • S cuando se trata de un expediente con accidente.	Texto plano
2	fecha	Fecha en formato ISO 8601: (aaaammdd o aaaa-mm-dd o aaaa/mm/dd)	Fecha
3	hora	La hora se establece en rangos horarios de 1 hora	Hora
4	direcci	Tipos de viales existentes en la ciudad de madrid	Lista de valores
5	Nombre_vial	Nombre del vial (según BDC, base de datos ciudad)	Texto plano
6	numero	Número de la calle, cuando tiene sentido	Número
7	dirección_auxiliar	Si fuese necesario (calle xxxxx c/v calle yyyy)	Número
8	cod_dis	Código del distrito (numérico)	Lista de valores (1)
9	distri_may	Nombre de distrito en mayúsculas y sin acento	Lista de valores (1)
10	num_bar	Código de barrio (numérico)	Lista de valores (1)
11	barrio_may	Nombre de barrio en mayúscula y sin acento	Lista de valores (1)
12	Latitud	En decimal, y con punto como separador.	Número
13	Longitud	En decimal, y con punto como separador.	Número
14	X	En metros (6 dígitos), y con punto como separador de decimales si fuesen necesarios.	Número
15	Y	En metros (7 dígitos), y con punto como separador de decimales si fuesen necesarios.	Número

(1) Lista de valores que tienen que corresponder con la denominación oficial de distritos y barrios indicada más arriba de este documento.

Anexo II: Sectores según la NTI

SECTOR	IDENTIFICADOR
Ciencia y tecnología *Incluye: Innovación, Investigación, I+D+i, Telecomunicaciones, Internet y Sociedad de la Información.*	ciencia-tecnologia
Comercio *Incluye: Consumo.*	comercio
Cultura y ocio *Incluye: Tiempo libre.*	cultura-ocio
Demografía *Incluye: Inmigración y Emigración, Familia, Mujeres, Infancia, Mayores, Padrón.*	demografia
Deporte *Incluye: Instalaciones deportivas, Federaciones, Competiciones.*	deporte
Economía *Incluye: Deuda, Moneda y Banca y finanzas.*	economia
Educación *Incluye: Formación.*	educacion
Empleo *Incluye: Trabajo, Mercado laboral.*	empleo
Energía *Incluye: Fuentes renovables*	energia
Hacienda *Incluye: Impuestos.*	hacienda
Industria *Incluye: Minería.*	industria
Legislación y justicia *Incluye: Registros.*	legislacion-justicia
Medio ambiente *Incluye: Meteorología, Geografía, Conservación fauna y flora.*	medio-ambiente
Medio Rural *Incluye: Agricultura, Ganadería, Pesca y Silvicultura.*	medio-rural-pesca
Salud *Incluye: Sanidad.*	salud
Sector público *Incluye: Presupuestos, Organigrama institucional, Legislación interna, Función pública.*	sector-publico
Seguridad Incluye: Protección civil, Defensa.	seguridad

SECTOR	IDENTIFICADOR
Sociedad y bienestar *Incluye: Participación ciudadana, Marginación, Envejecimiento Activo, Autonomía personal y Dependencia, Invalidez, Jubilación, Seguros y Pensiones, Prestaciones y Subvenciones.*	sociedad-bienestar
Transporte *Incluye: Comunicaciones y Tráfico.*	transporte
Turismo *Incluye: Alojamientos, Hostelería, Gastronomía.*	turismo
Urbanismo e infraestructuras *Incluye: Saneamiento público, Construcción (infraestructuras, equipamientos públicos).*	urbanismo-infraestructuras
Vivienda *Incluye: Mercado inmobiliario, Construcción (viviendas).*	vivienda

Anexo III: Tabla de distritos (versión abreviada)

La información de barrios y distritos se puede encontrar en los datasets "Barrios municipales de Madrid" y "Distritos municipales de Madrid". Existe varias formas de representarlos, como por ejemplo código de distrito numérico (1) o código de distrito en texto (01), o nombre de distrito en mayúsculas con acento o sin acento, o incluso en letra capital: CHAMARTIN, CHAMARTÍN, Chamartin o Chamartín. En el conjunto de datos se expresa las distintas opciones siendo todas ellas válidas. Lo mismo ocurriría para barrios.

A continuación se refleja una tabla resumida con menos opciones:

COD_DIS	COD_DIS_TX	DISTRI_MAY	DISTRI_MT
1	01	CENTRO	CENTRO
2	02	ARGANZUELA	ARGANZUELA
3	03	RETIRO	RETIRO
4	04	SALAMANCA	SALAMANCA
5	05	CHAMARTIN	CHAMARTÍN
6	06	TETUAN	TETUÁN
7	07	CHAMBERI	CHAMBERÍ
8	08	FUENCARRAL - EL PARDO	FUENCARRAL - EL PARDO
9	09	MONCLOA - ARAVACA	MONCLOA - ARAVACA
10	10	LATINA	LATINA
11	11	CARABANCHEL	CARABANCHEL
12	12	USERA	USERA
13	13	PUENTE DE VALLECAS	PUENTE DE VALLECAS
14	14	MORATALAZ	MORATALAZ
15	15	CIUDAD LINEAL	CIUDAD LINEAL
16	16	HORTALEZA	HORTALEZA

17	17	VILLAVERDE	VILLAVERDE
18	18	VILLA DE VALLECAS	VILLA DE VALLECAS
19	19	VICALVARO	VICÁLVARO
20	20	SAN BLAS - CANILLEJAS	SAN BLAS - CANILLEJAS
21	21	BARAJAS	BARAJAS

Anexo IV: Listado de criterios o checklists a revisar

A continuación, se recogen a modo de check-list o listado de criterios a revisar los diferentes criterios referenciados en la guía para poder realizar revisiones rápidas y que no se olvide ningún criterio de los indicados.

1. Metadatos del conjunto de datos

CRITERIO	Correcto
Criterio 1 - El nombre del conjunto de datos es correcto	
Criterio 2 - Existe descripción correcta para el dataset	
Criterio 3 - El dataset tiene asignado un sector correctamente	
Criterio 4 - Existe un conjunto de palabras clave correcto	
Criterio 5 - Existe la fecha desde/hasta del dataset	
Criterio 6 - Frecuencia de actualización correcta	
Criterio 7 – Responsable del conjunto de datos	
Criterio 8 - Existe un documento de estructura correctamente estructurado	
Criterio 9 - Existe la licencia correcta del dataset	

2. Formatos reutilizables de los ficheros

CRITERIO	Correcto
Criterio 1 - Formalmente bien construido	
Criterio 2 – No existen filas o columnas en blanco	
Criterio 3 - Formato correcto del encabezado	
Criterio 4 - Evitar filas o columnas de totales o subtotales	
Criterio 5 - Sólo un tipo de dato por columna	
Criterio 6 - No incluir hojas vacías	
Criterio 7 - Mismo orden y número de columnas en todas las filas, series temporales y formatos	
Criterio 8 - El fichero debe contener una única tabla de datos	
Criterio 9 - El nombre del fichero de datos es consistente con la serie anterior	
Criterio 10 - Uso de “;” como campo separador de caracteres en los ficheros CSV	
Criterio 11 - El conjunto de datos está dividido en distintas distribuciones de manera que cada una de ellas sea suficientemente tratable con programas informáticos habituales	
Criterio 12 – No existe demasiada anidación en los datos	
Criterio 13 - Codificación correcta de caracteres	
Criterio 14 – Organización vertical de la información, en vez de horizontal	
Criterio 15 – Identificación en los datos, del año y mes a que hacen referencia	

Criterio 16 – Mes mejor en formato numérico en vez de texto, para permitir una ordenación cronológica de los meses, en vez de alfabética	
Criterio 17 - Utilizar el mayor número de formatos posibles, para que los datos sean más accesibles	
Criterio 18 – No existen metadatos de autor	

3. Datos contenidos en los ficheros

CRITERIO	Correcto
Criterio 1 – Orden lógico de las columnas	
Criterio 2 - Los tipos de campos se ajustan a lo esperado	
Criterio 3 - Asignación de un ID único	
Criterio 4 - Los valores de datos de tipo fecha y fecha/hora deben describirse en formato ISO 8601	
Criterio 5 - Cumplimiento de codificación para información de barrios y distritos	
Criterio 6 – Formato de dirección válida	
Criterio 7 - Los valores nulos y no nulos se ajustan a lo esperado	
Criterio 8 - Las coordenadas latitud y longitud, correctamente representadas	
Criterio 9 – Las coordenadas X e Y, correctamente representadas	
Criterio 10 - Decimales representados con coma	
Criterio 11 - No se deben utilizar caracteres de formato de “miles”	
Criterio 12 - No se deben incluir ceros a la izquierda	
Criterio 13 - Valores de distribución de cada columna coherentes con la serie anterior	
Criterio 14 - Las unidades de medida y monedas deben indicarse por separado, o en el nombre de las columnas	
Criterio 15 - Confidencialidad y anonimización de los datos	
Criterio 16 - DATO UNICO. Consistencia entre datos del portal de Madrid y fuentes externas (Madrid.es y Banco de Datos de Estadística)	

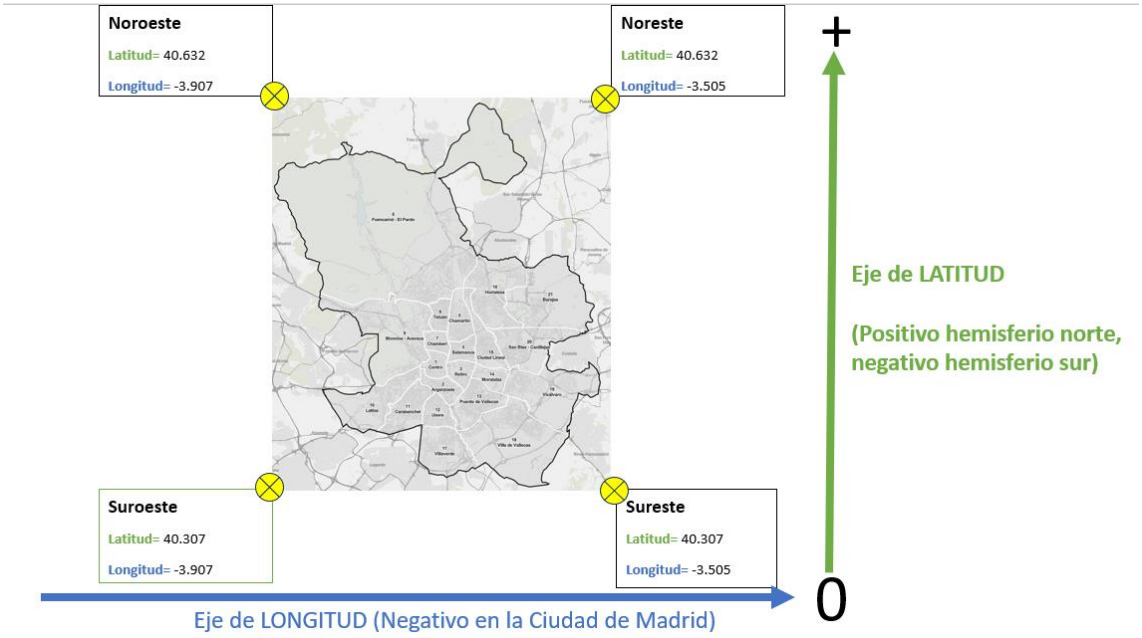
Anexo V: Rangos orientativos de las coordenadas X-Y para la Ciudad de Madrid

Importante, esto es una simplificación de las coordenadas y se proporcionan unos rangos amplios. Lo que sí es seguro es que sí:

- Una coordenada X en metros, está fuera del rango: 423250 – 457000, ese punto estará fuera del término de Madrid.
- Una coordenada Y en metros, está fuera del rango: 4462000 – 4498356, ese punto estará fuera del término de Madrid.

Pero ojo, se podría dar la situación que unas coordenadas dentro de los rangos indicados estuviesen fuera de Madrid. Esto es porque, como se ve en la figura, el termino de Madrid no es un cuadrado perfecto y los limites indicados son los de las esquinas de ese hipotético cuadrado

5 Anexo VI: Rangos posibles para Latitud y Longitud en la Ciudad de Madrid (ETRS89)



Importante, esto es una simplificación de las coordenadas y se proporcionan unos rangos amplios. Lo que sí es seguro es que sí:

- La latitud está fuera del rango 40.307 ---- 40.632 ese punto estará fuera del término de Madrid.
- La longitud está fuera del rango -3.907 --- 3.505 ese punto estará fuera del término de Madrid.

Pero ojo, se podría dar la situación que unas coordenadas dentro de los rangos indicados estuviesen fuera de Madrid. Esto es porque,como se ve en la figura, el termino de Madrid no es un cuadrado perfecto y los limites indicados son los de las esquinas de ese hipotético cuadrado

Anexo VII: Como eliminar metadatos de autor en los ficheros excels

Es una buena práctica quitar los metadatos de autor, para así evitar que figuren los mismos cuando se publica el Excel. Por otra parte, si a partir de estos ficheros excels se genera un pdf, el mismo contendrá también esos metadatos (ocurre lo mismo con cualquier fichero office).

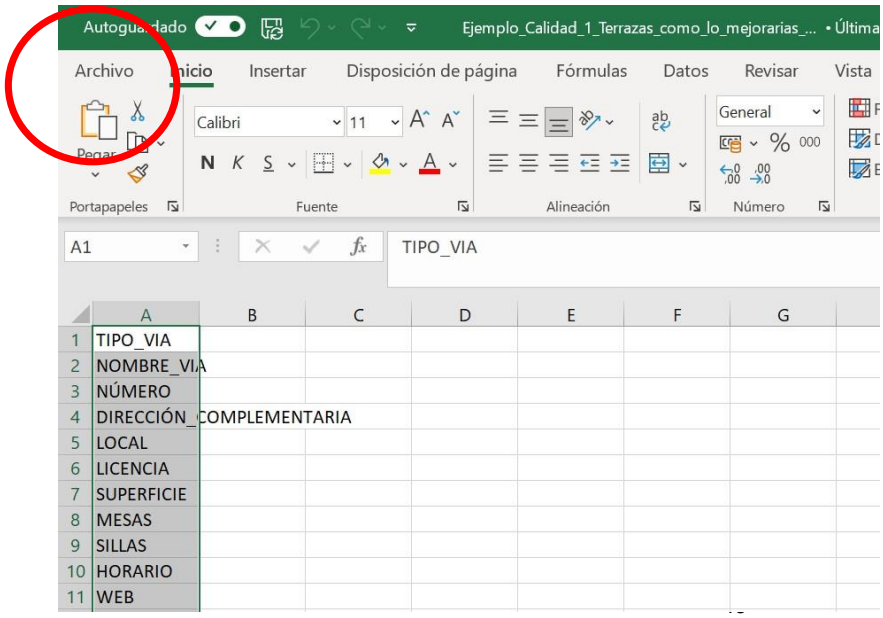
En el INCIBE, instituto nacional de ciberseguridad, se dan recomendaciones y técnicas para eliminar metadatos:

https://www.incibe.es/empresas/blog/son-los-metadatos-y-eliminarlos#:~:text=Para%20eliminar%20los%20metadatos%20de%20archivos%20ya%20existentes%20abre%20el,%22Utiliza%20datos%20de%20usuario%22

INSTITUTO NACIONAL DE CIBERSEGURIDAD

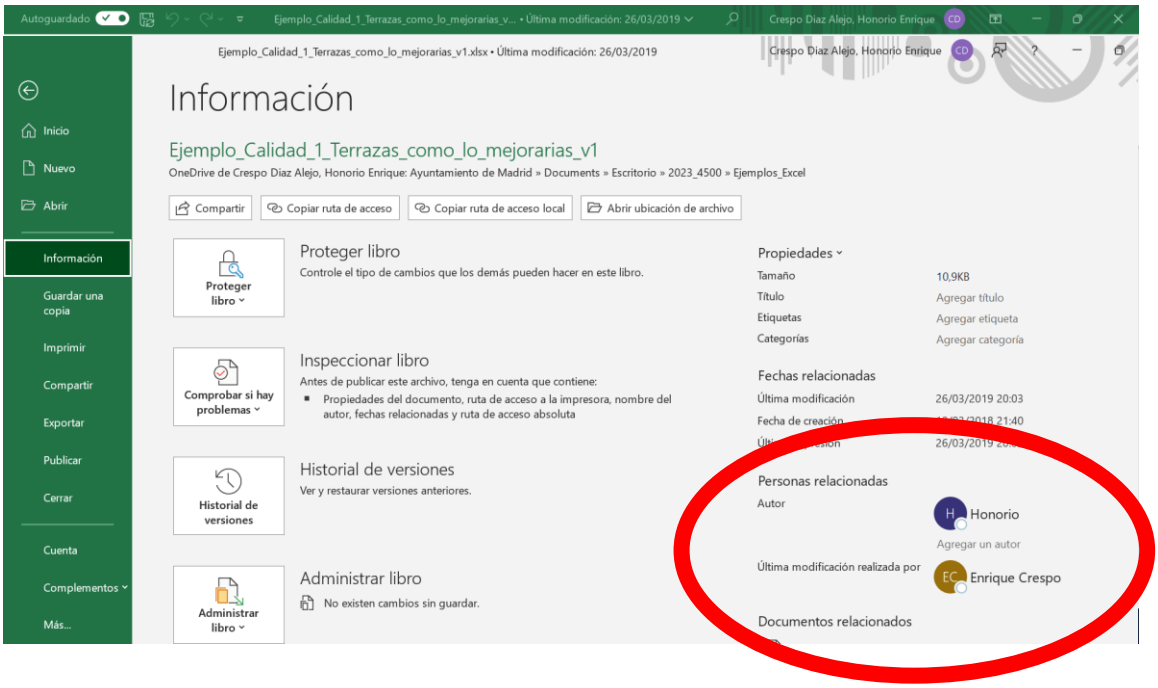
A continuación, se representan los pantallazos y pasos necesarios para eliminar los metadatos de un fichero Excel:

1. **Pulsar sobre el menú "Archivo"**

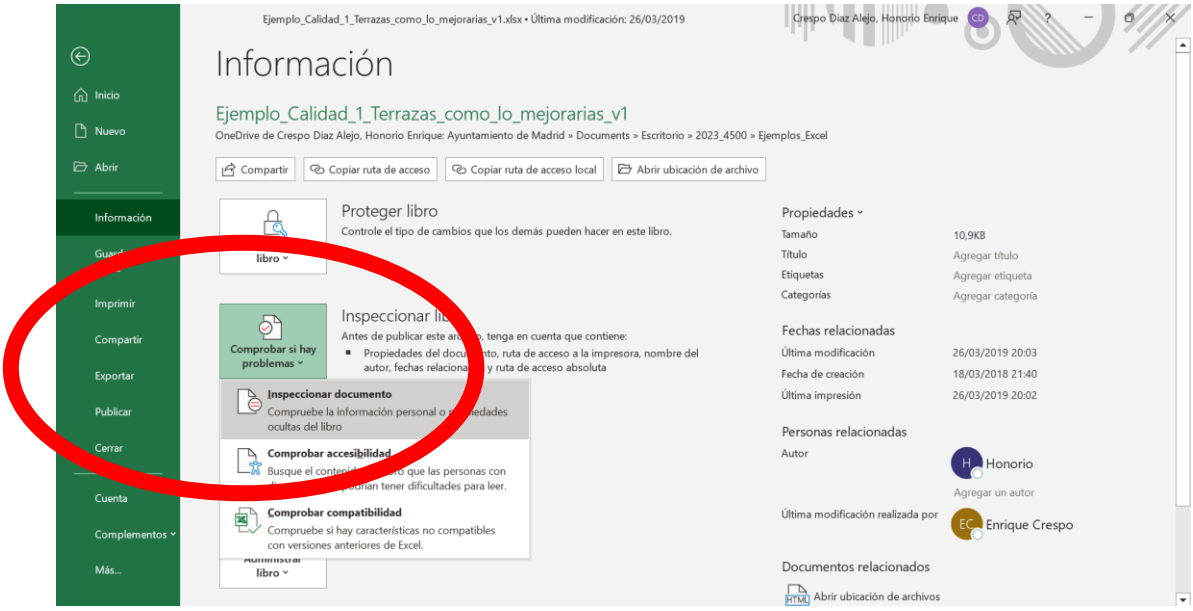


2. Pulsar sobre la opción "Información"

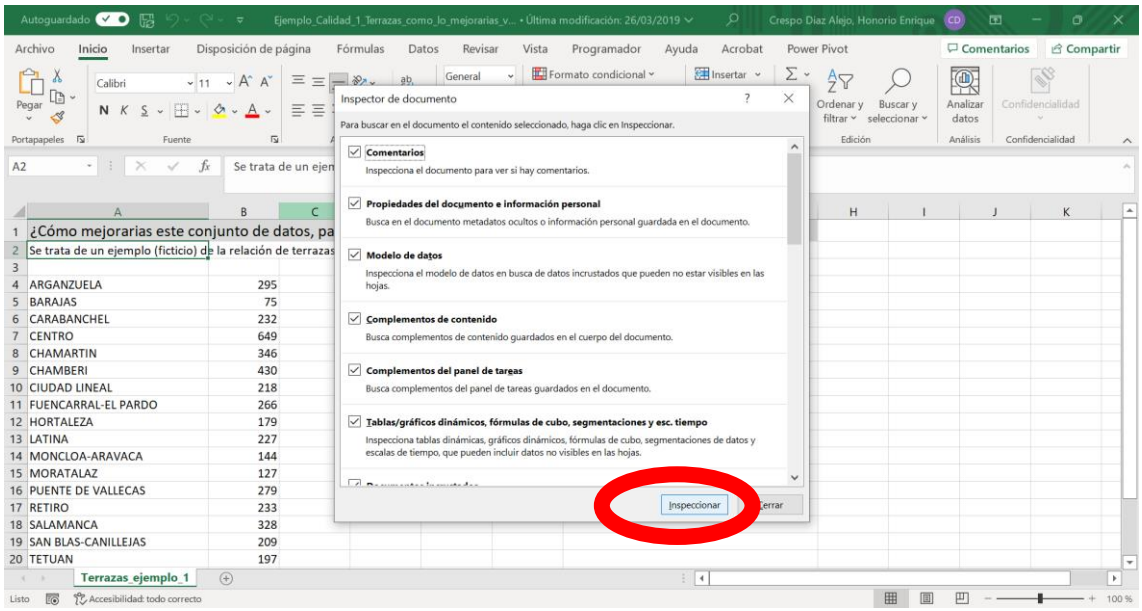
Al pulsar sobre esta opción, saldrán los metadatos de autor y fechas asociadas al fichero.



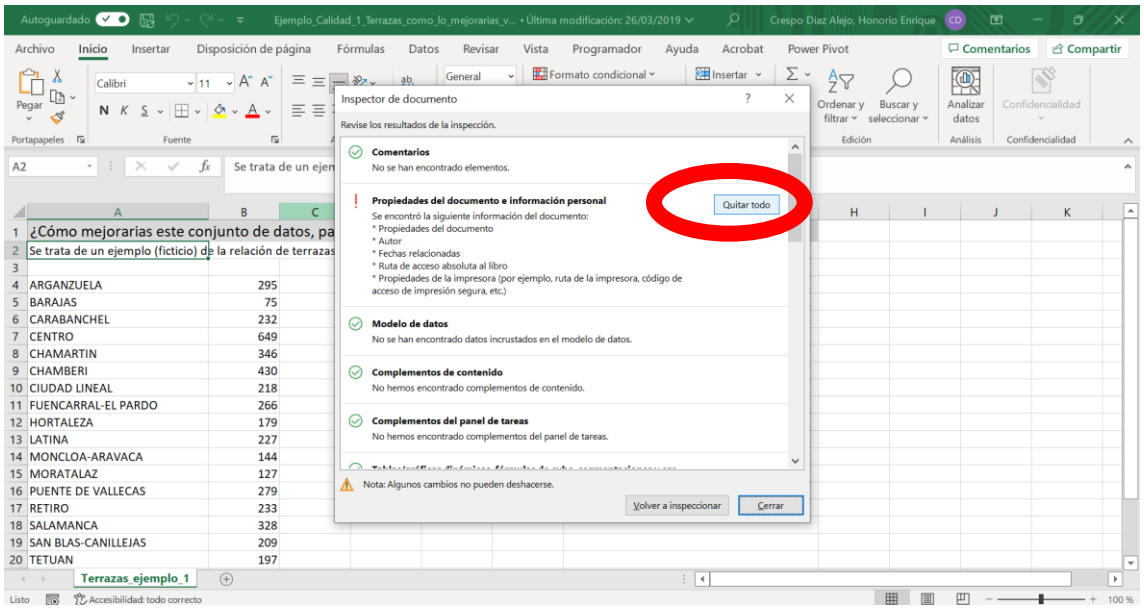
3. Pulsar sobre la opción "Comprobar si hay problemas" y después "Inspeccionar documento"



4. **Pulsar sobre la opción "Inspeccionar"**



5. **Pulsar sobre la opción "Quitar todo" y después cerrar**



De esa forma, se habrán quitado los metadatos. Se podría volver a pulsar Archivo / Información, para asegurarse que ya no figuran los metadatos.