1. Introducción

Los datos abiertos son datos que pueden ser utilizados, modificados y compartidos libremente por cualquier persona para cualquier propósito. En los últimos años, la cantidad y variedad de datos abiertos publicados por las administraciones públicas a nivel mundial -uno de los principales usuarios y productores de datos- ha aumentado de manera tangible, de la misma manera que ha aumentado la voluntad política de apertura y con ello las normativas, hojas de ruta y directrices técnicas. La disponibilidad de la información del sector público como datos abiertos puede aportar un considerable valor añadido y satisfacer una demanda creciente que proviene de empresas, organizaciones gubernamentales, desarrolladores y de la sociedad en su conjunto. No obstante, la apertura, desde el punto de vista normativo y técnico, no es suficiente para crear un ecosistema de reutilización de datos prospero puesto que los fallos en la disponibilidad y la calidad de los datos pueden perjudicar, no solo a la reutilización de los datos, sino también a la credibilidad de las instituciones que los publican.

La calidad de los datos es una de las principales trabas para el sector reutilizador y es un aspecto fundamental para garantizar su reutilización y respaldar el ecosistema de la creación de servicios y aplicaciones. La baja calidad de los datos abiertos obstaculiza una reutilización eficiente de los mismos debido a la necesidad de invertir recursos, por parte de los usuarios, en la comprensión de conjuntos de datos mal documentados y la realización adicional de tareas de depuración y procesamiento. Para incentivar la reutilización de los datos abiertos, es necesario que las administraciones inviertan en la mejora de su calidad.

Esta guía se orienta a publicadores de datos y se presenta como un compendio de **directrices para mejorar la calidad de datos** actuando directamente sobre cada una de las características que la definen. Por otro lado, la recopilación de estas pautas pretende orientar a los reutilizadores de datos sobre cómo afrontar las debilidades de calidad que pueden presentar los conjuntos de datos con los que trabajar.

En esta guía se utiliza el término conjunto de datos o dataset indistintamente, asociado al concepto de colección de datos que poseen vínculos entre sí de acuerdo con alguna estructura, esquema o modelo de representación. La disponibilidad de datasets para su reutilización se puede realizar mediante la descarga de archivos o accediendo a servicios de datos (APIs). Los archivos de datos descargables constituyen la materialización completa de un dataset. En cambio, el acceso a datos mediante servicios permite obtener múltiples datasets en función de las consultas que el reutilizador configure. Igualmente, se denominan distribuciones a la forma en la que se representan los datasets es decir, los formatos. Estos son diversos y responden a las necesidades del reutilizador: puede ser una hoja de cálculo CSV o Excel, un archivo XML, un archivo de imagen, datos vinculados en RDF, etc. La disponibilidad de formatos alternativos de cada dataset facilita la reutilización.

La estructura del documento comienza con una introducción a la calidad de datos, la definición de sus características según las diferentes normas técnicas de referencia internacionales, para centrar posteriormente el cuerpo del documento en una descripción de pautas generales y específicas de los formatos de datos abiertos más habituales, que detallan los problemas que es necesario afrontar, las

características de calidad afectadas y las recomendaciones para su resolución con ejemplos prácticos que ayudarán a entender cada caso presentado. El documento cierra con dos aspectos importantes que contribuyen a mejorar la calidad general como son la estandarización, el enriquecimiento y la documentación de datos abiertos.

Este documento toma como referente la guía para la calidad de datos de data.europe.eu, publicada en 2021 por la Oficina de Publicaciones de la Unión Europea[1]. De dicha guía se han recopilado y adaptado a la estructura del documento las pautas y los ejemplos más relevantes para alcanzar un buen nivel de calidad de datos abiertos. Otros ejemplos utilizados se han adaptado del dataset "Auto MGP Dataset" disponible en el popular *UCI Machine Learning Repository* de la Universidad de California.

[1] Publications Office, *Data.europa.eu data quality guidelines*, Publications Office, 2021, https://data.europa.eu/doi/10.2830/79367

2. Definición y requisitos mínimos de la calidad de los datos abiertos

1.1. ¿Qué es la calidad de los datos abiertos?

Existen múltiples definiciones para la calidad de datos, pero la más extendida la define como la **idoneidad de un conjunto de datos para servir a su propósito específico.** Esta definición implica que la calidad de los datos debe ser gestionada en base al **cumplimiento de unos requisitos determinados por las características que los definen**. La gestión de la calidad de los datos, además, debe proporcionar métodos y herramientas para evaluar y establecer procesos de mejora siempre que sean precisos.

A lo largo de estos últimos años, la mayoría de las iniciativas de datos abiertos aplican métodos y herramientas para evaluar y establecer procesos de mejora de la calidad de sus datos. Muchas de estas iniciativas han utilizado el método de 5 estrellas para los datos abiertos publicado por Tim Berners-Lee en 2006[1]. Aunque este esquema es ampliamente utilizado para medir la calidad de los datos, sólo cubre un aspecto específico de la calidad, la codificación utilizada para publicar los datos, por lo que un conjunto de datos publicado puede alcanzar el nivel de 5 estrellas, pero al mismo tiempo mostrar una calidad deficiente ya que puede presentar otros tipos de errores, como errores de sintaxis, duplicidad de registros o datos obsoletos, entre otros que se irán revisando en esta guía práctica.

Igualmente es común identificar iniciativas que ligan la calidad de los datos con los principios que debe regir toda política de apertura de datos. Estos principios nos indican que los datos deben ser completos, primarios, actuales accesibles, procesables por maquinas, no discriminatorios, no propietarios y sin restricciones de utilización. Pero estos principios, aunque guardan relación, no ponen el foco en la calidad de los datos, sino en una serie de propiedades que deben poseer los conjuntos de datos para considerarlos abiertos y reutilizables.

En la literatura existen diferentes enfoques sobre qué características deben tener los datos de calidad y esto conlleva a que existan diferentes modelos de medición y aseguramiento. Algunas referencias que establecen características imprescindibles para conseguir una alta calidad en los datos se citan a continuación.

- Los principios definidos por la carta internacional de los Datos Abiertos (Open Data Charter) establecen ya una serie de cualidades que los datos de calidad deben cumplir, tales como completitud, exhaustividad, puntualidad, oportunidad, comparabilidad e interoperabilidad. Sin embargo, hay otros aspectos que también definen la calidad de los datos, y deben tenerse en cuenta a la hora de producir y valorar cualquier tipo de datos.
- La OCDE, en su informe "Marco de calidad y directrices para las actividades estadísticas de la OCDE" publicado en 2011, considera la calidad en términos de 7 dimensiones: pertinencia, precisión, credibilidad, actualidad, accesibilidad, interpretabilidad y coherencia, y además,

establece que está ampliamente vinculada con las perspectivas, necesidades y prioridades de los reutilizadores.

- Por otro lado, la norma ISO/IEC 25012 también establece un modelo de Calidad del Producto de Datos compuesto por 15 características, clasificadas en dos grandes categorías, 12 de ellas relacionadas directamente con los datos como producto: exactitud, completitud, consistencia, credibilidad, actualidad, accesibilidad, conformidad, confidencialidad, eficiencia, precisión, trazabilidad y comprensibilidad.

Otro enfoque distinto, aunque en línea con la calidad que deben presentar los datos, es la propuesta que, en 2016, publica la revista Scientific Data de Nature, sobre los "Principios FAIR para la gestión y administración de datos científicos". Los principios FAIR, son un conjunto de directrices precisas y medibles que debe seguir cualquier científico para la publicación de sus datos con la mayor calidad posible. Estos principios hacen hincapié en la capacidad de acción de los sistemas informáticos para **encontrar, acceder, interoperar y reutilizar** los datos con la mínima intervención humana, ya que los seres humanos dependen cada vez más del apoyo informático para tratar los datos como resultado del aumento del volumen, la complejidad y la velocidad de creación de los datos. En 2018, se creó la comunidad GO FAIR con el fin de ayudar e implementar en la comunidad científica los principios FAIR.

Principio	Descripción
Encontrables (Findable)	Los datos y metadatos pueden ser encontrados por la comunidad de manera sencilla después de su publicación, mediante herramientas de búsqueda. • Asignar un identificador único y persistente. • Describir los datos con metadatos de manera prolija. • Indexar los datos y metadatos en el recurso de búsqueda. • En los metadatos se debe especificar el identificador de los datos que se describen.

Accesibles (Accessible)	Los datos y metadatos están accesibles y por ello pueden ser descargados por otros investigadores utilizando sus identificadores.
	• Los datos y metadatos son recuperables por su identificador utilizando protocolos de comunicación estandarizados.
	  ○ El protocolo es abierto, gratuito y de aplicación universal.
	  ○ El protocolo permite un procedimiento de autenticación y autorización cuando sea necesario.
	• Los metadatos son accesibles, incluso cuando los datos ya no están disponibles.

Principio	Descripción
Interoperables (Interoperable)	Los datos deben poder integrarse con otros datos. Además, los datos deben poder interoperar con aplicaciones o flujos de trabajo para su análisis, almacenamiento y procesamiento. • Los datos y metadatos deben utilizar un lenguaje formar, accesible, compartido y ampliamente aplicable. • Los datos y metadatos utilizan vocabularios que siguen los principios FAIR. • Los datos y metadatos incluyen referencias cualificadas a otros datos o metadatos.
Reutilizables (Reusable)	Los datos y metadatos deben estar bien descritos para que puedan ser replicados y/o combinados en diferentes entornos. • Los datos y metadatos están bien descritos con una pluralidad de atributos precisos y relevantes. • Los datos y metadatos están asociados a una procedencia detallada. • Los datos y metadatos cumplen con las normas de la comunidad pertinentes para el sector.

Figura 1. Principios FAIR para la gestión y administración de datos

1.2. Características de calidad de datos

Los referentes mencionados anteriormente definen un amplio conjunto de características comunes que se espera que cumplan los conjuntos de datos considerados de alta calidad

Figura 2. Características de calidad de los datos según la norma ISO 25012

A continuación, se describe someramente cada una de las características:

- **Exactitud/Precisión**, aunque no son términos equivalentes se refieren a la veracidad que proporcionan los datos. Los datos que presentan esta característica representan correctamente el valor verdadero del atributo al cual simbolizan en el mundo real. Además, las mediciones que son precisas son consistentes y replicables.
- **Completitud:** los datos se consideran completos cuando está disponible toda la información requerida para un atributo. Los datos deben presentar un nivel de detalle y una desagregación adecuada para ser relevantes y reutilizables.
- **Consistencia/Coherencia,** los datos deben estar libres de contradicciones y tener coherencia lógica en un contexto específico, por ejemplo, de formato o temporal.
- **Credibilidad** tanto para los datos en sí como para la fuente de información. 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. La credibilidad, también incluye el concepto de autenticidad (la veracidad de los orígenes de datos, atribuciones y compromisos).
- **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.
- **Accesibilidad**, referida a la facilidad de acceso a los datos. Los datos deben estar disponibles para la más amplia gama de usuarios y propósitos.
- **Conformidad**, los datos se adhieren a estándares o normativas vigentes.
- **Confidencialidad,** los datos se deben publicar respetando la privacidad y seguridad de estos. En contextos específicos, los datos sólo serán accedidos e interpretados por usuarios autorizados. La confidencialidad es un aspecto fundamental de la seguridad de la información.
- **Eficiencia,** los datos tienen atributos que puede ser procesados y proporcionados con unos recursos razonables.
- **Trazabilidad** respecto a la fuente u origen de los datos. Los datos tienen atributos que proporcionan un histórico del camino de acceso auditado a los datos o cualquier otro cambio realizado sobre ellos.
- **Comprensibilidad /Interpretabilidad** los datos pueden ser interpretados y leídos por los usuarios y ser expresados utilizando lenguajes, símbolos y unidades coherentes con el contexto de los datos. Cierta información sobre la comprensibilidad puede ser expresada mediante metadatos.

A lo largo de esta guía veremos cómo los diferentes problemas que se relatan afectan a una o más de las características indicadas.

1.3. ¿Por qué es importante disponer datos de la mayor calidad posible?

Generalmente, la solución de los problemas de calidad de los datos en conjuntos de datos abiertos implica una inversión significativa, a veces con rendimientos decrecientes, tanto por parte de los

publicadores como por parte de los usuarios. Ya la OCDE en el "Marco de calidad y directrices para las actividades estadísticas de la OCDE" mencionado con anterioridad, determina que la rentabilidad (costes + beneficios) es un factor que se debe tener en cuenta en cualquier análisis de calidad, ya que puede afectar a todas las características antes mencionadas.

Haug et al., en 2011, llevaron a cabo un análisis de las causas y una estimación de los costes asociados a la deficiente calidad de los datos. Los autores afirman que, en la práctica, la baja calidad de los datos puede implicar perjuicios económicos a la organización que publica los datos de múltiples formas. Además, determinan que las organizaciones tienden a sobreestimar la calidad de sus datos e infravalorar el coste generado por los errores derivados. En numerosas ocasiones, los publicadores de datos gastan más dinero en solucionar errores derivados de datos deficientes que en evitar con antelación los potenciales problemas que su uso a largo plazo puede generar en los reutilizadores.

Este análisis sostiene que el nivel óptimo de mantenimiento de calidad de los datos no es lograr datos perfectos, si no alcanzar un equilibrio entre el coste de las tareas asociadas con el aseguramiento de la calidad de los datos y el ahorro del coste causado por datos de baja calidad. Además, los autores, clasifican los tipos de costes como:

- **Costes ocasionados por la deficiente Calidad de los Datos ocasionados por**:
  - Costes directos asociados a la verificación, re-entrada y compensación de datos
  - Costes indirectos derivados de decisiones o acciones incorrectas y pérdida de inversiones.
- **Costes en mejora o de aseguramiento de la Calidad de los Datos derivados de**:
  - Costes de **prevención** asociados a la formación, monitorización, desarrollo e implementación estándar.
  - Costes de **detección** asociados a la creación de análisis e informes asociados a los datos
  - Costes de **reparación** asociados a la planificación de reparaciones y la reparación de implementaciones.

En el concepto de los datos abiertos es importante tener en cuenta que los costes derivados de la mala calidad de los datos, trasciende a la repercusión que tiene sobre la propia organización que los gestiona y en primera instancia utiliza, dado que se trasladan al sector reutilizador, multiplicando así el efecto negativo que conlleva.

Por último, aunque no es el objetivo de esta guía, es esencial que los conjuntos de datos estén acompañados de unos metadatos de excelente calidad, ya que es el primer contacto que el usuario tiene con el conjunto de datos. Para llevar a cabo una reutilización efectiva de los datos, es imprescindible que estén acompañados de una documentación que ayude a los reutilizadores a comprender el contenido de los datasets y cómo usarlos, incluidos los defectos ya detectados. Los metadatos forman parte del conjunto de datos y proporcionan información crucial sobre el origen de los datos, el grado de actualización, las restricciones legales sobre su uso y otra información relevante. Además, los metadatos deberían de informar sobre la calidad que presentan los datos. Unos metadatos deficientes disminuyen la probabilidad de reutilización de los datos.