CIUDADES ABIERTAS

Adaptación de estructuras de conjuntos de datos para asegurar su calidad y anonimización

ACTUACIÓN D5 – Implantación de Soluciones para la Publicación de los Conjuntos de Datos Abiertos

INICIATIVA PLATAFORMA DE GOBIERNO ABIERTO, COLABORATIVA E INTEROPERABLE (121/17-SP)

Citar como: Corcho O, De Pablo V (2021) Adaptación de estructuras de conjuntos de datos para asegurar su calidad y anonimización.
doi: 10.5281/zenodo.5942552

Este documento ha sido elaborado en el marco de la iniciativa 'Plataforma de Gobierno Abierto, Colaborativa e Interoperable' cofinanciada por el Ministerio de Economía y Empresa, a través de la Entidad Pública Empresarial Red.es, y por los ayuntamientos de A Coruña, Madrid, Santiago de Compostela y Zaragoza y con la cofinanciación del Fondo Europeo de Desarrollo Regional (FEDER), dentro de la 'II Convocatoria de Ciudades Inteligentes'.

RESUMEN EJECUTIVO

Este documento recoge, por una parte, una serie de guías relacionadas con el aseguramiento de la calidad de los conjuntos de datos publicados en los portales de datos abiertos, considerando tanto el contenido de las distribuciones disponibles para cada conjunto de datos como los metadatos correspondientes a estos conjuntos de datos.

El proceso de aseguramiento de la calidad se enmarca en el contexto de un marco más general sobre gobernanza de datos, fundamentado en buenas prácticas que son habituales en todo tipo de organizaciones, incluyendo empresas privadas y administraciones públicas. Basándose en este contexto general, se realizan recomendaciones generales sobre cómo asegurar una buena calidad en la publicación de los datos abiertos, con recomendaciones específicas relacionadas con cada uno de los tipos de datos que normalmente son publicados como datos abiertos. También se realizan propuestas sobre los metadatos asociados a dichos conjuntos de datos, comenzando con recomendaciones generales, basadas en fuentes bibliográficas existentes y en un análisis de fichas metodológicas asociadas a conjuntos de datos ya publicados, y realizando una propuesta específica de cómo deberían describirse estas fichas metodológicas en el futuro.

Este documento también incluye un informe con recomendaciones sobre cómo abordar la anonimización de datos en el caso de que esta sea necesaria atendiendo al contenido de un conjunto de datos (por ejemplo, por contener datos sensibles como lo son los datos personales). Se describen distintas estrategias de anonimización que están siendo ampliamente utilizadas en el estado del arte, así como herramientas existentes y que podrían ser de aplicación previo a la publicación de determinados conjuntos de datos por las administraciones públicas.

Finalmente, se incluyen recomendaciones sobre las cláusulas tipo que podrían incluirse en los pliegos de prescripciones técnicas de los procesos de contratación pública para poder asegurar la apertura y la calidad de los datos, y checklists que pueden ser utilizados por los responsables de los portales de datos abiertos para asegurar una calidad mínima en los datos que se les proporcionan como resultado de la ejecución de estos contratos.

TABLA DE CONTENIDOS

RESUMEN EJECUTIVO 2
TABLA DE CONTENIDOS 3
1. CALIDAD EN LA PUBLICACIÓN DE DATOS ABIERTOS 4
1.1 ¿QUÉ ES LA CALIDAD DE DATOS? 5
1.2 LA CALIDAD DE DATOS EN EL CONTEXTO GLOBAL DE LA GOBERNANZA DE DATOS 6
1.3 DIMENSIONES DE CALIDAD DE LOS DATOS 9
1.4 DIMENSIONES DE CALIDAD DE LOS METADATOS 11
1.4.1 HERRAMIENTAS DE VALIDACIÓN DE METADATOS 12
2. ASPECTOS ORGANIZATIVOS Y PROCESOS PARA ASEGURAR LA CALIDAD DE LOS DATOS 14
2.1 EQUIPO DE TRABAJO PARA EL ASEGURAMIENTO DE LA CALIDAD DE DATOS 14
2.2 PROTOCOLO DE ASEGURAMIENTO DE LA CALIDAD DE DATOS 15
2.2.1 PASO 1. DEFINICIÓN DE ESTÁNDARES DE DATOS 16
2.2.2 PASO 2. DEFINICIÓN DE PROCESOS Y PROCEDIMIENTOS INTERNOS PARA EL TRATAMIENTO Y PUBLICACIÓN DE DATOS ABIERTOS 18
2.2.3 PASO 3. DEFINICIÓN DE CRITERIOS COMUNES PARA EL ASEGURAMIENTO DE LA CALIDAD DE DATOS 21
3. PROPUESTAS PARA ASEGURAR LA CALIDAD DE DATOS 22
3.1 PROPUESTAS GENERALES POR TIPOS DE FORMATOS 22
3.1.1 PROPUESTAS GENERALES PARA DATOS EN FORMATOS TABULARES: CSV Y XLS(X) 23
3.1.2 PROPUESTAS GENERALES PARA DATOS EN FORMATO JSON 24
3.2 PROPUESTAS PARA CONJUNTOS DE DATOS ESPECÍFICOS 25
3.2.1 PROPUESTAS PARA CONJUNTOS DE DATOS ESTADÍSTICOS 25
3.2.2 PROPUESTAS PARA CONJUNTOS DE DATOS GEOGRÁFICOS 27
3.2.3 PROPUESTAS PARA CONJUNTOS DE DATOS MUY GRANDES (NORMALMENTE DE OBSERVACIONES DE SENSORES) 28
3.3 PROPUESTAS ESPECÍFICAS SEGÚN EL TIPO DE DATOS DE UNA COLUMNA O PROPIEDAD 29
3.3.1 PROPUESTAS PARA LA REPRESENTACIÓN DE FECHAS 29
3.3.2 PROPUESTAS PARA LA REPRESENTACIÓN DE DATOS NUMÉRICOS 29
3.3.3 PROPUESTAS PARA LA REPRESENTACIÓN DE UNIDADES DE MEDIDA Y MONEDAS 30
3.3.4 PROPUESTAS PARA LA REPRESENTACIÓN DE DATOS TEXTUALES 30
3.4 HERRAMIENTAS DE VALIDACIÓN 31
4. PROPUESTAS PARA ASEGURAR LA CALIDAD DE LOS METADATOS 32
4.1 FICHAS DE METADATOS EN PORTALES DE DATOS ABIERTOS DE MUNICIPIOS ESPAÑOLES 32
4.2 PROPUESTA DE FICHA METODOLÓGICA TIPO 33
5. ANONIMIZACIÓN DE DATOS 39
5.1 PROCESO DE ANONIMIZACIÓN DE CONJUNTOS DE DATOS 40
5.2 TÉCNICAS DE ANONIMIZACIÓN APLICABLES A CONJUNTOS DE DATOS 41
5.3 HERRAMIENTAS DE ANONIMIZACIÓN 43
6. CLÁUSULAS TIPO Y CHECKLISTS PARA ASEGURAR LA CALIDAD EN LOS DATOS GENERADOS A PARTIR DE CONTRATOS PÚBLICOS 44
6.1 CHECKLIST PARA LA COMPROBACIÓN DE CRITERIOS DE CALIDAD APLICABLES A DISTRIBUCIONES DE CONJUNTOS DE DATOS 45

1. CALIDAD EN LA PUBLICACIÓN DE DATOS ABIERTOS

De acuerdo con la guía internacional *Data Management Body of Knowledge*[1], publicada por la *International Data Management Association* (DAMA) y considerada por numerosas empresas y organizaciones como texto de referencia para la gestión de datos dentro de dichas organizaciones, el tratamiento de los datos debe centrarse en varios elementos, imprescindibles para extraer todo su potencial. Destacan los siguientes:

- **Los datos son un activo:** los datos tienen un valor y unas propiedades únicas. Su tratamiento conlleva un coste y tiene unos riesgos asociados. En este sentido -sobre todo cuando se trata de datos abiertos- la gestión de los datos debe tener en cuenta todo el ciclo de vida de los mismos; desde su adquisición y/o generación hasta su eliminación, sin olvidar la necesaria protección de los datos en todas las etapas de este ciclo de vida. Los datos no son un objetivo en sí mismos, su finalidad es servir a las propias instituciones y a los reutilizadores externos, ya sean empresas, investigadores o la ciudadanía.
- **Compromiso de la institución:** una buena gestión de datos requiere un enfoque claro sobre el uso que se les vaya a dar, tanto dentro como fuera de la organización. Para ello se hace imprescindible desarrollar una completa política de datos.
- **Visión y liderazgo para el cambio:** el programa de gestión de datos requiere de una fuerte inversión, tanto de tiempo como de esfuerzo. Es esencial tener clara la finalidad de su publicación, se necesita un enfoque proactivo. Esto requiere comunicación continua con las partes interesadas. Las razones, los objetivos y las expectativas de la gestión de datos deben quedar muy claros para todos los implicados en el proceso de publicación de los datos desde el principio. Todos los que participen en dicho proceso deben contar con la formación adecuada.
- **Colaboración:** la gestión de datos exige varias habilidades diferentes y requiere la participación de toda la organización. Todas las partes interesadas deben estar incluidas en el programa de gestión de datos. El equipo de trabajo debe estar bien equilibrado y contar con las habilidades necesarias.
- **Calidad de los datos:** el objetivo principal de la gestión de datos es proporcionar datos a todas las partes interesadas con una calidad que les permita llevar a cabo su actividad. La calidad de datos debe definirse, medirse y aplicarse mediante procesos y directrices claros. Entre las dimensiones de la calidad de datos figuran la corrección, la integridad y la coherencia de los mismos. La gestión de la calidad de los datos implica un coste y este debe ajustarse a las necesidades del proyecto.
- **Gobernanza de datos:** la gestión de datos requiere una gobernanza que garantice la calidad de los datos. Es necesaria una responsabilidad clara sobre todos los activos y procesos implicados en dicha gestión. La gobernanza deja claro quién debe definir y custodiar la calidad de estos activos, y quién debe rendir cuentas en caso de problemas (con la calidad) de los datos. La gobernanza debe marcar los procesos en torno a los datos para gestionar la calidad a lo largo de todo el ciclo de vida.
- **Enfoque general:** un programa de gestión de datos suele durar mucho tiempo. La agilidad es clave y no se debe hacer una inversión de varios años sin esperar beneficios. Por eso es necesario generar valor y comprobar, de manera regular, en qué se traduce éste.
- **Metadatos:** los metadatos son imprescindibles para gestionar los propios datos. Se utilizan para definir y clasificar los datos, describir su origen, especificar la calidad y orientar sobre dónde y cómo utilizar los datos, así como para facilitar su descubrimiento y reutilización. Los metadatos son muy importantes para medir el éxito del programa de gestión de datos.
- **Transparencia:** se debe mostrar claramente y de manera pública la hoja de ruta, los costes del proceso y los éxitos del programa (ahorro de costes, mejoras de calidad, generación de valor, etc.) a lo largo del proceso de gestión de datos. Es un elemento imprescindible para ganar la confianza de todas las partes interesadas, tanto generadores como publicadores y reutilizadores. La transparencia en los procesos de gestión de datos informa a los usuarios sobre las características y el estado de los propios datos y de los productos desarrollados a partir de su uso. Muestra de dónde vienen, su calidad, sus clasificaciones y limitaciones. Aporta toda la información importante para su reutilización. Los metadatos son también un requisito clave para ello, y la transparencia determina que estos datos se puedan publicar.

[1] https://www.dama.org/cpages/body-of-knowledge

- **Principios y normas:** la gestión de datos debe basarse en principios y normas establecidas previamente. Éstas formalizan las reglas que se acuerden en toda la institución. Los principios y las normas deben ser claros, concretos, concisos y coherentes para aportar valor a la organización.

Entre los elementos mencionados en la guía de DAMA destaca la **calidad de los datos** como principio esencial para gestionar con éxito cualquier proyecto basado en datos. *"Gestionar los datos es gestionar la calidad de los datos"*, afirman en la guía. Establecer de qué se habla cuando se habla de calidad de datos es imprescindible para asegurar el proyecto.

1.1 ¿QUÉ ES LA CALIDAD DE DATOS?

La mayor parte de las guías y recomendaciones existentes sobre publicación de datos abiertos hacen mención no sólo a la publicación de datos abiertos por defecto, sino que también destacan la necesidad de asegurar que los datos sean únicos y compartidos, georreferenciados y de calidad (categorías recogidas, por ejemplo, por la Guía de Datos Abiertos de la FEMP[2]).

Teniendo en cuenta el creciente papel que juegan los datos (abiertos o no) en la toma de decisiones a todos los niveles (estratégico, táctico y operacional), es muy importante que se pueda asegurar que los datos utilizados tengan una calidad suficiente y acorde con el tipo de decisión que se toma y con el uso que se les va a dar. Unos datos abiertos de mala calidad pueden impactar de forma negativa no sólo en la efectividad y acierto en dichas decisiones sino, en términos generales, en la confianza por parte de los miembros de la propia institución, por parte de los reutilizadores o de la ciudadanía.

En estudios e informes sobre los primeros pasos a dar para el desarrollo de un proyecto centrado en datos abiertos por parte de las administraciones públicas, se ofrecen definiciones básicas sobre qué es calidad de los datos y se proponen sistemas para su evaluación. En el Manual Práctico para Manejar la Calidad de los Datos Abiertos[3], publicado por datos.gob.es dentro de la Iniciativa Aporta, se indica que si bien no hay una única definición que permita hablar de calidad de los datos abiertos, sí existen referencias que pueden ayudar a fijar un concepto propio de calidad de datos, siempre acorde al proyecto que se esté desarrollando. Estas son:

- Los principios originales de los datos abiertos gubernamentales, definidos en la reunión de Sebastopol[4].
- Los principios de la Carta Internacional de los Datos Abiertos[5].
- Las dimensiones de la calidad de los datos abiertos, establecidas por la iniciativa *Open Data Support* de la Comisión Europea[6].
- Los atributos de calidad inherentes a los productos de datos, identificados por la norma ISO/IEC 25012[7].

Durante varios años, en este contexto, se han popularizado clasificaciones como el Esquema de desarrollo de 5 estrellas para Datos Abiertos[8], propuesto por Tim Berners-Lee en 2010, donde se relaciona un número de estrellas del 1 al 5 en orden ascendente en relación con la forma de publicación y la calidad de cada conjunto de datos publicados; considerándose de 1 estrella los conjuntos de datos que están alojados en la Web, de 2 estrellas los que se encuentran en formatos estructurados, de 3 estrellas los que se encuentran en formatos estructurados que además son abiertos, de 4 estrellas los datos en formatos adaptados a la Web (como RDF), y de 5 estrellas los datos enlazados entre sí:

[2] http://femp.femp.es/files/3580-1617-fichero/Gu%C3%ADa%20Datos%20Abiertos.pdf
[3] https://datos.gob.es/sites/default/files/doc/file/manual_practico_para_mejorar_la_calidad_de_los_datos_abiertos_1.pdf
[4] https://public.resource.org/8_principles.html
[5] https://opendatacharter.net/principles-es/
[6] https://www.opendatasupport.eu/
[7] https://www.aenor.com/normas-y-libros/buscador-de-normas/ISO?c=035736
[8] https://5stardata.info/es/

★ Publicar los datos en la Web, en cualquier formato, bajo una licencia abierta.

★★ Publicar los datos en la Web como datos estructurados (ej: Excel).

★★★ Publicar los datos en la Web usando formatos no propietarios (ej: CSV).

★★★★ Publicar datos en formatos adaptados a la Web (ej: RDF).

★★★★★ Publicar los datos enlazados (URL) con otros datos para proveer contexto (*linked data*).

Otra evaluación de calidad de datos que también se ha propuesto en el estado del arte es la de 6 dimensiones, *The Six Dimensions of EHDI Data Quality Assessment*[9], que valora la integridad, la singularidad, la oportunidad, la validez, la precisión y la coherencia de los conjuntos de datos. También se puede hablar del Modelo de 7 estrellas[10] para publicar y reutilizar conjuntos de datos enlazados. Por último, cabe destacar la Guía de Datos Abiertos de la FEMP[11], donde se define el dato de calidad como *"aquel que permite la interoperabilidad con otros, mediante el uso de modelos de datos y formatos comunes, y el uso de vocabularios consensuados"*, estando en línea con el modelo de 7 estrellas.

Aunque siguen siendo propuestas válidas, que conviene tener en cuenta de manera general, estas clasificaciones y categorizaciones sobre "calidad" se deben enmarcar en el contexto de las fases iniciales de publicación de datos abiertos, donde se pretendía fundamentalmente fomentar su publicación y disponibilidad por parte de las administraciones públicas, como complemento a la legislación emergente sobre reutilización de información del sector público. Este informe se va a centrar, de manera más completa, en **los pasos a seguir para asegurar la calidad de los datos durante la fase de consolidación de una estrategia de datos abiertos**, por lo que se partirá del trabajo previo realizado en estas fases, teniendo en cuenta otros marcos teórico/prácticos relacionados con la calidad de datos que están siendo ampliamente utilizados, tanto en contextos públicos como privados.

De manera general, se suele determinar que unos datos son de calidad si son *"adecuados para el propósito con el que se quieren utilizar"* (en inglés, *fitness for purpose*). Se trata de responder a la pregunta "¿es este conjunto de datos suficientemente bueno para aquello para lo que lo quiero utilizar?". Por tanto, se podría decir que el nivel de calidad variará dependiendo del propósito para el que se ha preparado el conjunto de datos. Mientras que en el contexto de los datos generados y utilizados dentro de una organización, esta definición puede ser fácil de entender y de aplicar, en el contexto de un portal de datos abiertos de un ayuntamiento, donde los datos pueden ser reutilizados por un amplio número de actores con muy diversos propósitos (ciudadanía, investigadores, empresas, la propia institución, etc.) asegurar la calidad de los datos es un trabajo mucho más complejo.

Por este motivo, es relevante entrar en más detalle sobre las técnicas que se pueden utilizar para asegurar la calidad de los datos -más allá de aquellas centradas exclusivamente en la limpieza de los datos o los formatos en los que los datos se proporcionan-, y considerar en este contexto otras dimensiones que suelen definirse en otros marcos de aseguramiento de la calidad de datos, como el propuesto por DAMA[12], que presta atención a los siguientes elementos: **completitud, unicidad, consistencia, pertinencia en el tiempo (*timeliness*), validez y exactitud (*accuracy*)**. Estas dimensiones se describen en más detalle en la sección 1.3.

1.2 LA CALIDAD DE DATOS EN EL CONTEXTO GLOBAL DE LA GOBERNANZA DE DATOS

La gobernanza de datos es una práctica en la que intervienen distintos elementos -personas, procesos y herramientas-, que confluyen en el desarrollo de cualquier proyecto. Sin embargo, para que la estructura de

[9] https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf
[10] https://link.springer.com/chapter/10.1007/978-3-319-11955-7_24
[11] http://femp.femp.es/files/3580-1617-fichero/Gu%C3%ADa%20Datos%20Abiertos.pdf
[12] https://damadach.org/dama-dmbok-functional-framework/

gobierno del dato pueda funcionar hay un elemento esencial, que prima sobre los demás: la calidad de los datos. Esto es lo que permite la continuidad del proyecto, garantizando la consistencia de los datos y la confianza de los reutilizadores. Como se indicaba en la sección anterior, estos últimos se hacen más complejos en el marco de un portal de datos abiertos de un ayuntamiento, ya que pueden estar vinculados a orígenes y propósitos muy distintos, pudiendo ser departamentos del propio ayuntamiento, empresas, investigadores, o la ciudadanía.

Existen distintos marcos de gobernanza de datos y todos ellos destacan la calidad de datos como uno de los aspectos clave a tener en cuenta a lo largo del proceso de gestión. Uno de estos marcos de trabajo es el propuesto por DAMA[13], como se ha comentado anteriormente, que sustenta los principios fundamentales de la gobernanza de los datos dentro de una organización sobre los siguientes aspectos:

- Los datos deben **definirse y describirse adecuadamente**, para permitir su uso completo, en relación con sus propósitos dentro de la organización.
- Los datos deben gestionarse a través de un administrador (propietario) **responsable de su uso y mantenimiento**.
- Los datos deben cumplir con los términos de **calidad** apropiados para su uso y propósitos dentro de la organización.
- Los datos deben ser **accesibles**, proporcionando procedimientos apropiados para garantizar su protección contra pérdidas, daños y uso indebido.
- Los datos deben estar **disponibles para ser compartidos** con todas las unidades organizativas autorizadas para acceder a ellos y utilizarlos.

En general, estos principios pueden considerarse muy alineados con los propósitos básicos de la gestión de datos abiertos por parte de una administración pública, con la excepción de que no sólo se trata de que sean utilizados dentro de la organización, sino también fuera de ella, en ocasiones con usos no previstos inicialmente.

La FIGURA 1 proporciona una visión de alto nivel de algunos de los aspectos clave a tener en cuenta en el entorno de la gobernanza de datos, según se recoge en el marco propuesto por DAMA, donde se incluyen tanto aspectos organizativos como técnicos. En el contexto de este informe, los aspectos más importantes son los siguientes:

- El eje de **calidad de los datos**.
- El eje de la **gestión de metadatos**, que incluye también aspectos relacionados con su calidad.
- El eje de los **datos maestros y datos de referencia**.
- El eje de la **seguridad de datos**, relevante en lo relacionado con los datos sensibles que deben ser anonimizados adecuadamente.

[13] https://dama.org/

FIGURA 1: *Gobierno de Datos. Esquema de conocimiento. DAMA Framework*

La elección de estos cuatro ejes no resta importancia a los demás elementos identificados por DAMA para asegurar la calidad de los datos. Por ejemplo, el eje central sobre gobierno de datos es clave para determinar la capa organizativa, de gestión y de recursos humanos, necesaria a su vez para realizar la gestión de los datos. Este eje central identifica, entre otros perfiles, a los responsables de asegurar la calidad de los datos y metadatos. Ejes técnicos como los de arquitectura, almacenamiento y *data warehousing* son necesarios para asegurar que se pueden llevar a cabo los pasos técnicos necesarios en el contexto del aseguramiento de la calidad.

Otros ejes, como los de modelado y diseño de datos, así como el de integración e interoperabilidad, están muy relacionados con el trabajo que se realiza en el proyecto Ciudades Abiertas para la creación de modelos de datos comunes (ontologías y vocabularios) que permitan a distintas entidades publicar los datos de manera homogénea. Y finalmente, aspectos como la gestión de otros tipos de documentos y contenidos es también relevante en la estrategia global de datos abiertos, y será abordada en otro documento de esta serie de informes.

Para llevar a cabo una buena gestión de la calidad de los datos abiertos en el contexto de los marcos globales de gobernanza de datos conviene tener en cuenta el marco de trabajo descrito en el *Government Data Quality Framework*, del Reino Unido[14], que se basa en estándares, buenas prácticas y herramientas usadas a nivel internacional en entornos industriales, adaptadas también al entorno gubernamental. Este marco identifica los siguientes principios de calidad, orientados a que la organización consiga establecer una cultura donde la calidad sea un aspecto clave:

1. Comprométete con la calidad de los datos.
2. Conoce a tus usuarios y sus necesidades.
3. Monitoriza la calidad durante todo el ciclo de vida de los datos.

14 https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework

4. Comunica sobre la calidad de los datos de manera clara y efectiva.
5. Anticipa los cambios que puedan afectar a la calidad de los datos.

1.3 DIMENSIONES DE CALIDAD DE LOS DATOS

Como ya se ha comentado, se pueden tener en cuenta seis dimensiones básicas relacionadas con la calidad de los datos, según la estructura diseñada por DAMA. En esta sección se proporcionan más detalles sobre cada una de ellas:

Completitud. Describe cuántos datos están presentes en el conjunto de datos. Para que un conjunto de datos esté completo, todos los registros deben estar incluidos, y todos los datos importantes deben estar presentes en estos registros. Es decir, el conjunto de datos contiene todos los registros que debería contener y todos los valores esenciales en cada registro tienen algún valor. Esto no quiere decir que todos los datos sean correctos.

> Por ejemplo, el conjunto de datos del censo de locales comerciales se puede considerar completo si contiene todos los locales comerciales de la ciudad, independientemente de dónde estén situados (en la calle, en mercados, en centros comerciales, en infraestructuras de transporte) y además para todos ellos contiene la información de dirección, estado y actividad comercial que se realiza, si el local está abierto.

Unicidad. Describe el grado de no duplicación de registros, es decir, que cada registro del conjunto de datos contenga una única entidad, y que cada entidad esté únicamente incluida una vez como registro.

> Por ejemplo, en el caso de un local comercial que tiene dos puertas de entrada por dos calles distintas, no se crea un registro por cada una de estas puertas de entrada como si se tratase de dos locales comerciales, sino que se asocian ambas puertas de entrada al mismo registro.

Consistencia. Describe el grado en el que los valores de un conjunto de datos no contradicen otros valores que representen a la misma entidad dentro del mismo conjunto de datos o en otros conjuntos de datos en el catálogo.

> Por ejemplo, cuando en el censo de locales comerciales un local aparece asociado a una actividad empresarial no relacionada con la hostelería, pero en el conjunto de datos de terrazas ese mismo local aparece como un local de hostelería con su terraza correspondiente.
>
> Otro ejemplo se puede dar cuando la numeración de un portal en el callejero aparece de forma distinta en dos conjuntos de datos como consecuencia de una modificación urbanística (construcción de un edificio anexo que ocupa una parcelación diferente).

Pertinencia en el tiempo. Describe el grado en el que los datos reflejan de manera precisa el periodo que representan, y en el que los valores están actualizados. Se considera que los datos están suficientemente actualizados si el periodo de tiempo que pasa entre su recolección y su publicación es apropiado para el uso esperado de los mismos.

> Por ejemplo, si los datos del censo de locales comerciales se actualizan cada dos años, los datos podrán ser útiles para realizar estudios sobre la evolución de la actividad comercial de la ciudad, pero probablemente no sean tan útiles para que una empresa de geomarketing los pueda utilizar para detectar oportunidades de locales comerciales disponibles en una zona.

Otro ejemplo puede darse cuando se produce una cancelación o una reserva de un espacio público para la celebración de un evento y no se registra a tiempo para que otro ciudadano/empresa pueda utilizarlo. En este caso, no se consideraría pertinente en el tiempo (*timely*).

Validez. Describe el grado en el que los datos están disponibles en el rango y formato esperados.

Por ejemplo, la fecha de concesión de una licencia de actividad a un local comercial está representada en formato fecha, es posterior a la construcción del local comercial y anterior al día actual.

Exactitud. Describe el grado en el que los datos se corresponden con la realidad, indicando potenciales sesgos en su recolección o tratamiento posterior. Esta propiedad se puede medir por cada registro individual o de manera global para el conjunto de datos.

Por ejemplo, en el censo de locales comerciales pueden utilizarse dos técnicas distintas para la asociación de actividades comerciales a un local: el uso de los procedimientos administrativos de solicitud de licencia o encuestas a pie de calle analizando la actividad de cada local. La técnica de recogida de los datos debe ser clara y declararse de manera explícita para permitir a cualquier reutilizador entender los posibles sesgos en su adquisición.

Otro ejemplo puede darse cuando se dispone de las direcciones exactas de viviendas (portal, piso, letra, escalera, etc.), actualizadas y limpias para poder utilizar los datos a la hora de realizar actuaciones temporales de cualquier tipo: envío de documentación (censos electorales), revisiones del gas, envío de notificaciones de Hacienda, multas, etc.

Junto a estos elementos, que estructuran las claves de la calidad de datos, hay que tener en cuenta los principios fundamentales de las estadísticas oficiales. En el Reino Unido, por ejemplo, se identifican tres principios de calidad básicos para la estadística pública[15]: uso de **fuentes de datos adecuadas** (las estadísticas deben basarse en los datos más apropiados para cumplir con los usos previstos, se debe evaluar, minimizar y explicar el impacto de cualquier limitación de uso de los datos), aplicación de **métodos correctos** (los productores de estadísticas y datos deben utilizar los mejores métodos disponibles y estándares reconocidos, y ser abiertos sobre sus decisiones) y aseguramiento de la **calidad en todo el proceso** (los productores de estadísticas y datos deben explicar claramente cómo se aseguran de que las estadísticas y los datos sean precisos, fiables, coherentes y oportunos). En este caso, el concepto *fit for purpose* se centra en que los datos estadísticos sean útiles para los usos para los que han sido propuestos, que se utilicen datos y métodos apropiados y que no lleven a confusión, centrándose en la importancia de que la recolección, preparación, análisis y publicación de los datos estadísticos sea una actividad desarrollada siempre por profesionales cualificados.

En España, el Instituto Nacional de Estadística (INE), por su parte, basa sus índices de calidad en el Código de Buenas Prácticas de las Estadísticas Europeas (CBPEE)[16], cuyos fines son: establecer el estándar para el desarrollo, producción y difusión de las estadísticas basado en una definición de calidad de las estadísticas, común a todo el Sistema Estadístico Europeo (SEE), que afecte a todos los ámbitos, desde el entorno institucional a los procesos de producción estadística, y a los productos estadísticos; y asegurar la calidad y la credibilidad de los datos. Estos principios hacen referencia, entre otros aspectos, a la **independencia profesional**, la **protección**

[15] https://code.statisticsauthority.gov.uk/the-code/quality/
[16] https://www.ine.es/ine/codigobp/codigo_2017.pdf

de la confidencialidad, la **fiabilidad de los resultados**, su precisión, oportunidad, puntualidad, accesibilidad, claridad, comparabilidad y coherencia.

1.4 DIMENSIONES DE CALIDAD DE LOS METADATOS

Los archivos y fichas de metadatos y la información que ofrecen para los conjuntos de datos disponibles en los catálogos de los portales de datos abiertos institucionales son, también, elementos a tener en cuenta al hablar de la calidad de los datos. La mayor parte de los portales de datos abiertos en todo el mundo publica su catálogo de acuerdo con el formato DCAT-AP. Asimismo, pueden seguir recomendaciones adicionales, como las recomendaciones de datos.gob.es en el entorno nacional en España, así como los principios y herramientas de validación proporcionadas por las iniciativas SEMIC[17] y por data.europa.eu (como se describe más adelante en esta sección) en el entorno europeo.

En este sentido, los portales de datos abiertos ya consolidados deberían contar, a la hora de mejorar la calidad de la publicación de datos y la experiencia de los usuarios, con los siguientes elementos relacionados con la calidad de sus metadatos:

1.- Sistemas de monitoreo automatizados para detectar problemas con los datos abiertos:

- Señalar qué metadatos faltan (o no son válidos) de los conjuntos de datos ya subidos al portal.
- Informar a los editores cuando los datos publicados no se ajusten a los esquemas prescritos.
- Solicitar a los editores que actualicen los conjuntos de datos cuando estén desactualizados

2.- Mecanismos claros para que los usuarios informen cuando los datos son inaccesibles, inutilizables o contienen errores, y tener diseñado un proceso para actuar en esos casos.

En el contexto nacional, el informe "Cómo publicar datos abiertos de manera rápida y sencilla (con CKAN)"[18], identifica de manera clara y recomienda el uso de tres niveles de publicación de metadatos: **obligatorios**, **recomendados** y **opcionales**. Al analizar las fichas de metadatos de distintos portales de datos abiertos, es fácil encontrar valores para los que componen el primer nivel (los obligatorios): nombre, descripción, temática y organismo; y para los que son parte del segundo nivel (los recomendados): etiqueta y condiciones de uso o licencia. Sin embargo, los que conforman el tercer nivel en la publicación de metadatos (los opcionales), que incluyen la fecha de publicación, la fecha de última actualización, la frecuencia de actualización, el idioma, la cobertura geográfica o temporal, la vigencia, los recursos relacionados y la normativa, tan sólo se localizan en algunos conjuntos de datos, y no siempre de manera estructurada.

[17] https://ec.europa.eu/isa2/actions/improving-semantic-interoperability-european-egovernment-systems_en

[18] https://datos.gob.es/sites/default/files/doc/file/como_publicar_datos_abiertos_de_manera_rapida_y_sencilla_con_ckan_v2.pdf

FIGURA 2: *Tipología de metadatos propuesta en "Cómo publicar datos abiertos de manera rápida y sencilla (con CKAN)"*

De hecho, según se indica en el Manual práctico para mejorar la calidad de los datos abiertos[19], es habitual que *"los metadatos puedan mostrar que la fecha de última actualización es bastante reciente, sin embargo, cuando accedemos a los datos nos encontramos con que son mucho más antiguos de lo que se indica en esos metadatos y que no se han actualizado hace bastante tiempo"*. Este problema, que se debe normalmente a limitaciones o características de las herramientas utilizadas para la actualización de datos, perjudica a índices de calidad como la exactitud, la consistencia y la actualidad. Debe prestarse la debida atención a los metadatos, ya que siempre hacen referencia a las características de los datos a los que acompañan y no a los registros que los contienen.

No todos los conjuntos de datos requieren del mismo nivel de información ni del mismo tipo de frecuencia y forma de actualización. Es posible que su publicación vaya acompañada de una fecha de creación y una fecha de última actualización, obviando una frecuencia de actualización, no siempre necesaria. En el caso de conjuntos de datos en los que sí se hace necesaria una actualización más frecuente (tráfico en las ciudades, medición del ruido y/o de la calidad del aire, etc.), y según el análisis de los conjuntos de datos de los portales analizados, sí suele aparecer este elemento concreto.

1.4.1 HERRAMIENTAS DE VALIDACIÓN DE METADATOS

En esta línea, los agregadores de conjuntos de datos (comúnmente conocidos como federadores), como por ejemplo el de datos.gob.es, realizan comprobaciones sobre los metadatos de los catálogos de datos abiertos que federan, sirviendo como herramientas de validación iniciales para determinar la calidad de dichos metadatos.

Asimismo, el portal de datos europeo cuenta con una serie de principios e instrucciones adicionales para los proveedores de datos[20], así como con un sistema de ayuda, el *Metadata Quality Assurance (MQA)*[21], para que dichos proveedores de datos puedan asegurar la calidad de los metadatos. Este sistema comprueba los datos con periodicidad semanal y monitorea la calidad de los metadatos recolectados o almacenados manualmente con el

19 https://datos.gob.es/sites/default/files/doc/file/manual_practico_para_mejorar_la_calidad_de_los_datos_abiertos_1.pdf

20 https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy

21 https://data.europa.eu/mqa?locale=en

formulario de creación de metadatos EDP. La calidad de los metadatos se basa en la validación frente al estándar de metadatos DCAT y en la disponibilidad de distribuciones de un conjunto de datos.

La funcionalidad y la metodología de este sistema se estructura en torno a una serie de indicadores, detallados en la FIGURA 3, cuyos resultados se almacenan de acuerdo con el Vocabulario de Calidad de Datos (DQV, Data Quality Vocabulary). En este proceso se llevan a cabo sucesivas comprobaciones de las URL de acceso y descarga. Las comprobaciones periódicas de la accesibilidad de todas las distribuciones garantizan la calidad: las distribuciones se comprueban a través del acceso a las URLs indicadas a través del protocolo HTTP, y las validaciones de metadatos se realizan con el lenguaje de validación SHACL[22].

La puntuación total de un conjunto de datos, según el Panel de Control del MQA, se obtiene sumando los puntos de cada una de las dimensiones. La **facilidad de localización** podría obtener una calificación máxima de 100 puntos, la **accesibilidad** otros 100, la **interoperabilidad** 110, la **reusabilidad** 75, y la **contextualidad** 20. Así, un conjunto de datos con un máximo de 405 puntos sería considerado excelente (351-405), habiendo también calificación buena (221-350), suficiente (121-220) y mala (0-120). El proceso completo de clasificación, las distintas dimensiones que contempla la herramienta MQA para determinar la calidad de los conjuntos de datos -que derivan de los principios FAIR[23]- y los detalles de la calificación se pueden consultar en "Metodología de evaluación de la calidad de los metadatos"[24].

FIGURA 3: Panel de control del MQA

[22] https://data.europa.eu/mqa/shacl-validator-ui/
[23] https://www.go-fair.org/fair-principles/
[24] https://data.europa.eu/mqa/methodology?locale=es

2. ASPECTOS ORGANIZATIVOS Y PROCESOS PARA ASEGURAR LA CALIDAD DE LOS DATOS

En la sección anterior se han descrito algunos de los aspectos más relevantes a tener en cuenta al considerar la calidad de los datos abiertos, incluyendo sus metadatos asociados. Se han identificado marcos de respuesta generales que inciden en la importancia de trabajar en la calidad de los datos y de los metadatos, y que están siendo utilizados en entornos públicos y privados. También se han descrito las dimensiones principales a tener en cuenta al analizar la calidad de los conjuntos de datos, así como algunos principios y herramientas que pueden ser usados para analizar la calidad de los metadatos correspondientes.

Hay muchas formas de garantizar que los datos conduzcan a mejores decisiones: haciéndolos lo más abiertos posible, asegurando que los datos sean precisos, que estén lo más libres posible de sesgos problemáticos. Todos estos elementos contribuyen a mejorar la calidad de los datos y aseguran el acceso a los datos a quienes los necesitan, contribuyendo a obtener resultados positivos al tiempo que evitan impactos dañinos.

En esta sección se propone un protocolo que permita a los proveedores de datos abiertos (por ejemplo, un ayuntamiento, y más concretamente la unidad organizativa responsable del portal de datos abiertos) determinar la calidad de los datos y sus metadatos correspondientes. Como parte de este protocolo también se proponen los roles que deben ser tenidos en cuenta para formar un equipo de trabajo no sólo centrado en la publicación de los datos sino también en el aseguramiento de la calidad de los datos abiertos publicados en el portal de datos abiertos.

2.1 EQUIPO DE TRABAJO PARA EL ASEGURAMIENTO DE LA CALIDAD DE DATOS

Entre los primeros pasos a seguir se hace necesario identificar los roles que debe asumir el **equipo de trabajo** que se encargará de ejecutar el protocolo de aseguramiento de la calidad de datos. Un equipo de calidad de datos en el contexto de un portal de datos abiertos es el responsable de asegurar que los datos abiertos que se publican en el portal (incluyendo sus metadatos) cumplen con los requisitos de calidad impuestos por la organización, asegurados a través de análisis durante todas las fases del ciclo de vida de los datos. También están encargados de proponer el marco general de requisitos de calidad, y de proponer las tareas necesarias para la mejora continua de la calidad de todos los conjuntos de datos y de sus metadatos, estableciendo prioridades y estrategias para alcanzar estos objetivos.

Este equipo de trabajo puede o no formar parte del mismo equipo de trabajo que se encarga del resto de tareas del ciclo de vida de los datos abiertos de la institución, según el tamaño de la institución, el volumen de datos que se deba tratar, cuestiones organizativas y de responsabilidades, etc. Este equipo de trabajo deberá contar con distintos perfiles, con el objeto de garantizar el éxito de la implementación de las líneas de aseguramiento de la calidad que se proponen.

En detalle, los perfiles del grupo de trabajo para el aseguramiento de la calidad de los datos abiertos publicados por una organización serán los siguientes:

- **Gestor del programa de calidad de datos**. Este rol es el encargado de especificar los requisitos de calidad de datos a aplicar en la organización en el contexto global de la gobernanza de datos. Estos requisitos pueden ir variando en el tiempo conforme la organización va madurando y mejorando en su calidad. Esta persona es responsable de llevar el día a día de las tareas de gestión de la calidad de los datos, asegurando que el resto del equipo trabaja conforme a la planificación, tanto en términos de tiempo como de esfuerzo dedicado, y que se respetan los estándares de calidad acordados.
- **Responsable de aspectos legales (privacidad de los datos).** Este rol puede ser asumido por el gestor del programa de calidad de datos, aunque en ocasiones se trata de una persona distinta, normalmente un abogado encargado de la gestión de los aspectos legales relacionados con los datos y su adaptación a la normativa vigente.

- **Analista de datos**. Este rol puede estar distribuido entre múltiples personas dentro de la organización, que pueden dedicar parte de su tiempo a estas tareas. Las tareas de este rol están centradas en la interpretación de cada conjunto de datos asignado y la realización de informes sobre los mismos (cómo se adquieren, dificultades relacionadas con la calidad, procesos que se llevan a cabo, necesidades del modelo de datos, etc.). Puede ser común tener analistas de datos especializados en distintos dominios (comercio, tráfico, hacienda, etc.), así como analistas de datos generalistas que conocen bien los procesos utilizados para capturar, procesar y publicar datos durante su ciclo de vida (por ejemplo, procesos ETL, anonimización de conjuntos de datos, etc.) .
- **Bibliotecario/documentalista de datos (*Data Steward/Curator*)**. Este rol es el encargado de asegurar, por cada conjunto de datos, que se cumplen los requisitos de calidad especificados tanto para los datos como para los metadatos, basándose en los informes generados por los analistas de datos, y conforme al plan generado por el gestor del programa de calidad de datos. En muchos casos, interviene en el tratamiento de los datos para asegurar su calidad y en la edición de metadatos.

El equipo de trabajo se responsabilizará de las tareas propias de organización y mantenimiento de la calidad de los datos, asegurando la sostenibilidad del portal de datos abiertos, como se indica en el informe *Recommendations for Open Data Portals: from set up to sustainability*[25].

2.2 PROTOCOLO DE ASEGURAMIENTO DE LA CALIDAD DE DATOS

Además de la definición de roles y responsabilidades en el equipo de gestión de la calidad de los datos, es importante contar con un protocolo de actuación para el equipo, que se denominará **protocolo de aseguramiento de la calidad de datos**. En este protocolo conviene tener en cuenta los dos grandes bloques de tareas sobre los que se debe organizar:

- La **administración y gestión de la calidad de los datos**, que forma parte del proceso global de gobernanza de los datos, como ya se ha indicado en la sección 1. El objetivo de este bloque de tareas es el de mejorar las pautas para que la organización establezca, implemente, mantenga y mejore su sistema de aseguramiento de la calidad de datos durante todo su ciclo de vida. Se centra en la definición de estándares para la representación de datos que sean aplicables a toda la organización, en la definición de procesos y procedimientos internos, así como en la definición de todos aquellos criterios que son comunes para el aseguramiento de la calidad de datos. Estas tareas deberían realizarse desde el inicio del proyecto de creación de un portal de datos abiertos, aunque también pueden abordarse en cualquier momento durante su ejecución, y pueden dar lugar a modificaciones de los estándares, procesos y procedimientos, o criterios aplicables en cualquier momento.
- La **inspección continua de conjuntos de datos específicos**, necesaria para validar los atributos y características de calidad de los mismos. El objetivo del proceso de inspección es definir un enfoque común y estandarizado para la inspección de datos cuando el proyecto del portal de datos abiertos ya está en marcha, tanto durante su fase inicial como durante su fase de consolidación, tal y como se apuntaba en las primeras líneas de este informe, en el apartado 1.1.

En el contexto del primer bloque, el de la administración y gestión de la calidad de los datos, en las fases iniciales de desarrollo de un portal de datos abiertos muchas de estas tareas se han guiado en estos años por lo definido por los sistemas de clasificación de portales de datos abiertos definidos en el Esquema de desarrollo de 5 estrellas para Datos Abiertos[26], de Tim Berners-Lee; la guía *The Six Dimensions of EHDI Data Quality Assessment*[27] o el Modelo de 7 estrellas[28] para publicar y reutilizar conjuntos de datos enlazados; así como con la Guía de Datos

[25] https://data.europa.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf
[26] https://5stardata.info/es/
[27] https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf
[28] https://link.springer.com/chapter/10.1007/978-3-319-11955-7_24

Abiertos de la FEMP[29], que habla del dato de calidad en su relación con la interoperabilidad, con el fin de facilitar el intercambio y la compartición de datos. Todos estos sistemas de clasificación, ya explicados brevemente en el apartado 1.1, hacen en general referencia a los primeros pasos en la puesta en marcha de un proyecto de datos abiertos.

Una vez que un portal de datos abiertos ya está en marcha y pasa a una fase de consolidación, que es en la que se centra fundamentalmente este informe, deben describirse de manera más detallada los procesos específicos de estandarización y certificación de la calidad de los datos y sus metadatos. En esta sección se describen las líneas generales de estos procesos, mientras que las secciones 3 y 4 proporcionan detalles más específicos en relación con tipos de datos específicos y sus metadatos asociados, y con las validaciones a tener en cuenta para cada uno de ellos.

2.2.1 PASO 1. DEFINICIÓN DE ESTÁNDARES DE DATOS

En la definición de este protocolo de actuación, se debe comenzar con la propuesta de los estándares de datos que serán aplicables a toda la organización, y en particular a la publicación de datos abiertos. Los estándares de datos se entienden como aquellos acuerdos en la representación y gestión de datos que son ampliamente reutilizables dentro de la organización (por ejemplo, dentro de todas las unidades organizativas del ayuntamiento), y que facilitan la publicación, acceso, compartición y utilización de datos con una mejor calidad en general (como se propone en la guía del Open Data Institute sobre Open Standards for Data[30]). Al mismo tiempo, identifican un conjunto de reglas consistentes para la gestión de tipos de datos específicos.[31] Los estándares de datos más comunes tienen relación con los formatos prioritarios a utilizar para gestionar y publicar los datos, con la forma de expresar fechas y tiempos, coordenadas espaciales, direcciones postales, etc. Es decir, se centran fundamentalmente en aquellos tipos de información que son bastante comunes en el catálogo de conjuntos de datos que gestiona y publica la organización, y proporcionan reglas claras para que cualquier persona en la organización que produzca o utilice datos sepa cómo tratar con ellos, así como aquellas organizaciones que puedan usar o proporcionar datos en el contexto de contratos públicos.

En general, gracias a la aplicación de estos estándares se puede conseguir: a) una reducción de los costes de mantenimiento de los datos; b) un ecosistema de datos mejorado; y c) una mejora en la interoperabilidad, que reduce el coste de compartir datos y puede aumentar la cantidad de personas y organizaciones que comparten datos utilizando los mismos estándares.

A pesar de tratarse de una definición de estándares aplicable al contexto de la organización que se encarga del portal de datos abiertos (un ayuntamiento), se considera una buena práctica que dicha definición y mantenimiento de estándares de datos esté alineada con los principios fundamentales, más generales, de los estándares abiertos[32], que se presentan a continuación:

1. **Cooperación:** cooperación respetuosa entre distintas organizaciones (en los principios se refiere a los organismos de normalización, pero esto es aplicable también a las distintas unidades organizativas de un ayuntamiento, o a varios ayuntamientos en caso de cooperación entre ellos). Cada una debe respetar la autonomía, integridad, procesos y reglas de propiedad intelectual de las demás.
2. **Adherencia a los cinco principios fundamentales del desarrollo de estándares:**
    - Claridad en los procesos. Las decisiones se toman con equidad y justicia entre todos los participantes. Ninguna de las partes domina ni guía el desarrollo de estándares. Los procesos de estandarización, revisión y actualización periódica son transparentes y están bien definidos.
    - Amplio consenso. Los procesos permiten considerar y abordar todos los puntos de vista.
    - Transparencia. Las actividades de estandarización son públicas, con una definición clara de su alcance y de las condiciones de participación, con registros accesibles de las decisiones tomadas

[29] http://femp.femp.es/files/3580-1617-fichero/Gu%C3%ADa%20Datos%20Abiertos.pdf
[30] https://standards.theodi.org/
[31] https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datastandards
[32] https://open-stand.org/about-us/principles/

y de los materiales utilizados para tomar decisiones. Asimismo, hay períodos de comentario público para cualquiera que quiera opinar sobre un estándar, especialmente en la fase final.
    - Equilibrio. Las actividades de normalización no están dominadas exclusivamente por ninguna persona, empresa o grupo de interés en particular (en el contexto de un ayuntamiento, esto sería aplicable a las distintas unidades organizativas del mismo).
    - Apertura. Los procesos de estandarización están abiertos a todas las partes potencialmente interesadas.
3. **Empoderamiento colectivo:** los estándares deben elegirse y definirse en función del mérito técnico, proporcionar interoperabilidad, escalabilidad, estabilidad y flexibilidad; permitir la competencia; servir como bloques de construcción para una mayor innovación; y contribuir a la creación de comunidades para el beneficio de todos.
4. **Disponibilidad:** las especificaciones de los estándares se ponen a disposición de todos para su implementación y despliegue.
5. **Adopción voluntaria:** los estándares se adoptan voluntariamente y el éxito lo determina su uso.

Es decir, idealmente en una organización como un ayuntamiento se deberían seguir procesos similares a los de un organismo de estandarización cuando se realice la definición de estos estándares de datos, con el objetivo de conseguir el mayor grado de aceptación y penetración de los estándares, así como para tener en cuenta la mayor cantidad de puntos de vista distintos de las distintas unidades organizativas del ayuntamiento.

En el contexto específico de un proyecto de datos abiertos de un ayuntamiento, se pueden tener en cuenta los siguientes tres grupos de estándares de datos:

- Estándares para la **definición de formatos comunes y reglas compartidas** en la publicación e intercambio de datos, incluyendo:
    - Formatos para almacenar y compartir los datos, de manera general. Por ejemplo, puede definirse como formatos prioritarios a seguir en la organización los que se proponen como ejemplos para conseguir dos y tres estrellas en la clasificación de Tim Berners Lee, tales como CSV, JSON, XML, GeoJSON, shapefiles, PC-Axis, etc., o los que se proponen como ejemplos para conseguir cuatro o cinco estrellas, tales como RDF.
    - Reglas para la definición de APIs para el acceso a datos. Por ejemplo, puede determinarse que las APIs deben seguir los principios REST, utilizar un esquema de nombrado de las URIs específico (por ejemplo, https://datos.*nombreciudad*.es/api/*dominio*/*recurso*) y permitir la autenticación y autorización de usuarios de una manera específica.
    - Reglas para tipos de datos específicos. Por ejemplo, el nombre de una organización o unidad organizativa es siempre un texto con un número de caracteres determinado, las fechas siempre se representan en zona horaria UTC, los números de calle para un portal se representan siempre con un texto y no con un número entero, etc.
- Estándares para **proporcionar orientación** y recomendaciones para compartir datos de mejor calidad, comprender su ciclo de vida completo y el flujo de información entre distintas entidades. Entre estos estándares se pueden considerar aquellas recomendaciones, modelos, protocolos y guías que permiten unificar aspectos como los siguientes:
    - Unidades de medida que se usan para adquirir y representar datos, tales como: grados Celsius para temperaturas, latitud y longitud (o coordenadas X e Y en una proyección como ETRS89) para posiciones geográficas, metros para definir alturas de elementos arquitectónicos o árboles, miligramos por metro cúbico para el monóxido de carbono, etc.
    - Protocolos o métodos para medir, capturar o compartir datos de manera consistente. Por ejemplo, aquellos relacionados con métodos estadísticos utilizados para el muestreo de poblaciones, métodos utilizados para las mediciones de calidad del aire/agua/etc., o del tráfico, y valores posibles para aquellas mediciones que están validadas y aquellas que no lo están, etc.
    - Reglas para generar identificadores para distintos tipos de recursos.
- Estándares para **compartir datos de acuerdo con ontologías, vocabularios o modelos de datos consensuados**. En estos casos, estos estándares definen los recursos que se van a publicar, así como todos sus atributos y relaciones y las posibles restricciones entre ellos. También definen listas de términos

controladas (normalmente representadas en forma de tesauros) que pueden ser utilizadas como valores de algunos atributos o relaciones (por ejemplo, para definir los tipos de eventos culturales que ofrece una ciudad). Este es uno de los trabajos que se ha venido realizando en el proyecto Ciudades Abiertas con varios de los conjuntos de datos propuestos como prioritarios por la FEMP (disponibles en https://vocab.ciudadesabiertas.es/).

Todos estos estándares, que se definen en el paso 1 y que se mantienen y evolucionan durante el desarrollo de un proyecto de datos abiertos, se suelen combinar para crear de esta manera un sistema amplio y consistente de reglas y recomendaciones, que pueda conducir a una mayor calidad en el tratamiento y publicación de los datos de la organización.

2.2.2 PASO 2. DEFINICIÓN DE PROCESOS Y PROCEDIMIENTOS INTERNOS PARA EL TRATAMIENTO Y PUBLICACIÓN DE DATOS ABIERTOS

En este paso se definen los procesos y procedimientos que todas las unidades organizativas de la organización deben seguir para poder proceder a la publicación de datos abiertos, siguiendo en todo momento el mayor número de estándares previamente definidos por la organización. En esta definición se tiene en cuenta el ciclo de vida habitual de los datos, que cuenta, como se indica en la FIGURA 4, con las siguientes fases: planificación; recolección, adquisición e ingesta; preparación, almacenamiento y mantenimiento; compartición y publicación; y archivado y destrucción. Es importante destacar que esta sucesión de tareas se realiza de manera iterativa, mediante un programa de trabajo que tendrá una frecuencia de actualización que dependerá de las características de la organización, y que será elaborado por el gestor del programa de calidad de datos y por el responsable de la gobernanza de datos de la organización.

FIGURA 4. Ciclo de vida de los datos y recomendaciones.

En la fase de **planificación**, se identifican los conjuntos de datos que se gestionarán durante el ciclo correspondiente, así como sus características principales, riesgos y oportunidades de la incorporación de nuevos conjuntos de datos o la modificación de conjuntos de datos existentes, etc. En el contexto específico del aseguramiento de la calidad de datos, no es necesario proporcionar unas guías ni una justificación específica de qué conjuntos de datos se seleccionan en cada momento y cómo se realiza la priorización, dado que esto es responsabilidad del gestor de la gobernanza de datos. Por tanto, en este protocolo sólo nos centramos en las tareas que dependen del aseguramiento de la calidad.

Tarea 1.1. Creación de ficha metodológica inicial. Por cada conjunto de datos seleccionado, se debe rellenar la parte de la ficha metodológica relacionada con el plan de gestión de los datos (descrita en la sección 4 de este documento). Esta tarea se realizará conjuntamente por los bibliotecarios de datos, los analistas de datos y los responsables de la adquisición y gestión de los mismos. Se identificarán aspectos como la procedencia de los datos, la política de almacenamiento y *backup*, modelos de datos originales, tamaño esperado, calidad esperada de los datos, formatos de publicación posibles, valores típicos en algunas columnas, frecuencia de actualización esperada, etc.

Tarea 1.2. Validación de las fichas metodológicas iniciales. El equipo de bibliotecarios de datos procederá a la validación de la información proporcionada en las fichas metodológicas iniciales, para determinar si las

descripciones proporcionadas son suficientes, para poder realizar recomendaciones a los gestores de datos, y para poder evaluar posibles riesgos en la obtención y publicación de los conjuntos de datos.

Tarea 1.3. Creación de un checklist. El equipo de bibliotecarios de datos, ayudado por analistas de datos, creará una checklist de calidad específica basada en los tipos de datos esperados para cada una de las columnas (atributos, valores), para la validación de datos de referencia usados (listas de códigos, tesauros, etc.), siempre en función del catálogo de estándares disponible en la organización. Esta checklist se le facilitará a los productores de datos para que puedan ser conscientes de cómo se procederá a la evaluación cuando se entregue el conjunto de datos para su publicación. El contenido de la checklist contendrá elementos identificados en la sección 3 de este documento.

En la fase de **recolección, adquisición e ingesta**, se preparan los datos para su posterior gestión en los sistemas de información de la organización así como para su posterior publicación, en fases posteriores, en el portal de datos abiertos. Esta fase la llevarán a cabo los analistas de datos correspondientes, que realizarán las operaciones necesarias para obtener los datos correspondientes, bien sea de sistemas de información ya existentes dentro de la organización (por ejemplo, una base de datos donde se almacena la información de locales comerciales y sobre la que hay que realizar un proceso ETL), de sistemas de información externos (por ejemplo, la base de datos nacional de subvenciones) o de sistemas que haya que poner en marcha (nuevas bases de datos, sensores que estén siendo instalados, datos procedentes de la ejecución de un contrato público específico, etc.). En todo momento, el equipo encargado de esta fase tendrá que seguir las recomendaciones realizadas en la tarea 1.3, a la que tienen acceso, para asegurar que los conjuntos de datos generados puedan ser validados adecuadamente.

La fase de **preparación, almacenamiento y mantenimiento** está conectada con la anterior y requerirá de una mayor o menor nivel de esfuerzo dependiendo de las características de los datos que se estén tratando. En general, y a pesar de ser muy importante para la evaluación posterior de la calidad de los datos que se publicarán, esta es una tarea en la que el equipo de aseguramiento de la calidad de los datos no se involucrará más allá de la intervención de los analistas de datos en alguno de los procesos. En todas las tareas de mantenimiento en las que se puedan producir cambios en los formatos de datos o en los valores de algunas de las columnas o atributos, se deberá mantener informado al equipo de aseguramiento de la calidad para que se pueda verificar de nuevo que los datos son conformes a lo especificado en las fichas metodológicas correspondientes. También se deberán actualizar las fichas metodológicas con cualquier modificación en el proceso de recolección, adquisición e ingesta, así como de preparación, almacenamiento y mantenimiento. Finalmente, en el caso de que los datos sean sensibles y tenga que ser anonimizados, se realizarán adicionalmente las tareas identificadas a continuación:

Tarea 3.1. Análisis de requisitos de confidencialidad y anonimización del conjunto de datos. En el caso de que en la primera fase se haya identificado que el conjunto de datos puede contener datos sensibles que deban ser tratados apropiadamente, el responsable de aspectos legales realizará un análisis de los mismos y propondrá una serie de técnicas a los analistas de datos para que pueda prepararlos para su publicación con las garantías correspondientes

Tarea 3.2. Preparación de los datos anonimizados para ser publicados. Los analistas de datos se encargarán del proceso de anonimización de los datos, en caso de que sea necesario, de acuerdo con las directrices indicadas por el responsable de aspectos legales en el paso anterior.

En la fase de **compartición y publicación** vuelve a participar de manera activa el equipo completo de aseguramiento de la calidad de los datos, dado que deben dar de alta el conjunto de datos en el catálogo de datos abiertos de la organización, debe realizar las validaciones de la calidad de los datos, y proponer revisiones en caso de que sean necesarias, y crear una versión final de la ficha metodológica, para su publicación junto con el resto de metadatos del conjunto de datos.

Tarea 4.1. Creación de entrada DCAT-AP en el catálogo de datos abiertos. El equipo de bibliotecarios de datos utilizará la información inicialmente capturada en la ficha metodológica de datos, así como la información adicional recogida durante el proceso, para la creación del registro (normalmente en DCAT-AP) correspondiente en el catálogo de datos, y validará este registro con las herramientas de validación correspondientes (federador de datos.gob.es, herramienta MQA de data.europa.eu).

Tarea 4.2. Revisión del conjunto de datos conforme a la ficha metodológica (tarea 1.2) y las checklists (tarea 1.3). Los analistas de datos deben entregar un informe sobre cómo el conjunto de datos es conforme a la ficha metodológica inicial, así como el análisis de las checklist que se crearon en la primera fase a partir de estas fichas metodológicas y la información adicional recopilada durante el proceso. Esta información será revisada por los bibliotecarios de datos y en última instancia puede ser revisada por el gestor del programa de calidad de datos y el responsable de aspectos legales, antes de proceder a la publicación definitiva.

Tarea 4.3. Publicación definitiva de datos y metadatos. Los bibliotecarios de datos crearán la versión definitiva de la ficha metodológica, que será añadida al registro de datos correspondiente, y publicarán el conjunto de datos de manera definitiva en el portal de datos abiertos.

En la fase de **archivado y destrucción**, el equipo de aseguramiento de la calidad (normalmente los bibliotecarios de datos) debe dar de baja el conjunto de datos del catálogo de datos e informar a todos los posibles usuarios de dicho conjunto de datos, si están registrados en el registro de reutilizadores, de la desaparición de este conjunto de datos, así como de las razones que han llevado a su desaparición (inclusión en un conjunto de datos de mayor alcance, eliminación del mantenimiento de los dispositivos que generaban los datos - por ejemplo, de unos sensores concretos que se determina que no son necesarios en el futuro -, etc.). Una buena práctica es realizar esta notificación con suficiente antelación, así como mantener en el catálogo una referencia a los datos archivados para así permitir posteriormente cualquier tipo de análisis post-mortem.

2.2.3 PASO 3. DEFINICIÓN DE CRITERIOS COMUNES PARA EL ASEGURAMIENTO DE LA CALIDAD DE DATOS

En este paso se definen todos aquellos criterios que serán comunes para el aseguramiento de la calidad de los datos, basado en los estándares acordados, y que se aplicarán durante las fases de revisión de calidad identificadas en el proceso correspondiente. Estos criterios se aplicarán tanto en el proceso de alta de nuevos conjuntos de datos como en los casos en los que haya modificaciones en los procesos de tratamiento de datos, inclusión de nuevos datos, etc., que se revisarán de manera puntual o periódica por el equipo de aseguramiento de la calidad. Estos criterios son los que se definen en las secciones 3 y 4 de este documento.

3. PROPUESTAS PARA ASEGURAR LA CALIDAD DE DATOS

La sección anterior se ha centrado en todos los aspectos organizativos necesarios para que la publicación de datos abiertos pueda realizarse con garantías de calidad, incluyendo la identificación de perfiles necesarios en los equipos de datos abiertos e identificando los procesos a seguir. Esta sección se centrará de manera más específica en las comprobaciones que deben realizarse en los distintos tipos de datos que pueden publicarse, y en las correspondientes acciones de mejora.

Antes de continuar, es importante destacar que no existe un único estándar para asegurar la calidad de los datos publicados. Por esta razón, asimismo, hay pocas herramientas que permitan la automatización de estas tareas (se hace mención a algunas de ellas en la sección 3.4), o que faciliten una creación más óptima de todos los metadatos (este último punto será tratado en la sección 4). Esta ha sido una de las razones, junto con la falta de procesos claros de aseguramiento de la calidad, por las que la calidad de los conjuntos de datos que se han venido publicando en muchos portales de datos abiertos no siempre ha sido tan alta como sería aconsejable. En muchas ocasiones, los equipos responsables de los datos abiertos reciben estos datos en formatos sin refinar, sin seguir unos estándares definidos por la organización (en muchas ocasiones porque estos no han sido definidos), desde distintas unidades organizativas con distintos puntos de vista y conocimientos sobre el tratamiento de datos abiertos. Es por ello que se ha dejado tradicionalmente en manos de los responsables de datos abiertos la responsabilidad de limpiar y refinar esos datos previo a su publicación, así como de la generación de los metadatos correspondientes y las guías metodológicas, en su caso. La falta de recursos y de herramientas para garantizar la calidad de los datos en tal diversidad de formatos hacen difícil mantener los estándares de calidad que se han discutido en las secciones previas.

La recomendaciones que se realizan en esta sección están basadas en la experiencia del equipo de redacción de este documento en la publicación y uso de datos en diversos tipos de formatos y para distintos tipos de aplicaciones, así como en diversos estándares y guías ya publicadas con recomendaciones específicas sobre determinados tipos de formatos y de datos, incluyendo la "Guía práctica para la publicación de datos tabulares en archivos CSV"[33] elaborada para datos.gob.es, y las guías para proveedores de datos de data.europa.eu[34] (especialmente sus anexos B y C). También es relevante el documento, más reciente, "Sympton tracking: standards and adaptors"[35], coordinado por el Open Data Institute y elaborado de manera colaborativa por varios colaboradores, donde se aborda la necesidad de contar con estándares y adaptadores para asegurar la calidad de los datos, y que surgió como respuesta a las dificultades en el uso de datos publicados en los primeros meses de la pandemia de COVID-19. Este último documento no contiene recomendaciones específicas, pero revela la necesidad de tener en cuenta guías claras a la hora de representar y publicar la información.

3.1 PROPUESTAS GENERALES POR TIPOS DE FORMATOS

En esta sección se incluyen propuestas generales asociadas a los distintos tipos de formato en los que los datos abiertos pueden ser publicados. Aunque existe una gran diversidad de formatos, las propuestas se centran en los más comunes en el ámbito de la categoría de dos y tres estrellas de Tim Berners Lee (CSV, XLS(X), JSON) y de cuatro y cinco estrellas (RDF). No se incluyen recomendaciones para formatos que están comenzando a dejar de ser populares, como XML, aunque aún se utilicen de manera amplia, u otros que son muy específicos (PC-Axis, Shapefiles, GeoJSON, KMZ, KML, TSV, GML, MARC-21, etc.).

Todas estas propuestas pueden servir como elementos dentro de la checklist que se proporciona a los productores de datos como resultado de la tarea 1.3 descrita en la sección 2.2.

[33] https://datos.gob.es/es/documentacion/guia-practica-para-la-publicacion-de-datos-tabulares-en-archivos-csv
[34] https://data.europa.eu/sites/default/files/edp_s1_gdl_data-supplier-guidelines_v9.pdf
[35] https://docs.google.com/document/d/1YBIyc8J3Tprj4_6kvU3mtDlCZSPm9UtGfefuJPn7Kzg/edit?ts=5f4e20cd

3.1.1 PROPUESTAS GENERALES PARA DATOS EN FORMATOS TABULARES: CSV Y XLS(X)

Existen diversas propuestas de carácter general que son relevantes para aquellas organizaciones que proporcionan datos en formatos como CSV y XLS(S). A continuación se muestra un listado de propuestas básicas, de carácter general, para la publicación de datos en este tipo de formatos tabulares, que son los más populares entre los portales de datos abiertos. Es importante destacar que la numeración no coincide con la utilizada en los documentos anteriormente referenciados, puesto que se ha decidido organizar la información de manera distinta en el contexto de este documento.

CSV/XLS(X)-1. Debe utilizarse el conjunto de caracteres de codificación UTF-8. El fichero CSV o XLS(X) debe utilizar, preferiblemente, un conjunto de caracteres de codificación UTF-8. En caso de que utilice otro conjunto de caracteres, esto debe quedar reflejado en la documentación para facilitar su uso correcto.

CSV/XLS(X)-2.- El fichero debe contener una única tabla de datos. Un archivo debe constar de una única tabla con todos los datos, incluso aunque no estén normalizados (atendiendo a las formas normales que son habituales en bases de datos relacionales). Si se necesitan varias tablas, cada una de ellas deberá ser un archivo independiente, en el caso de CSV, o varias hojas de cálculo dentro del archivo XLS(X). Asimismo, los metadatos - descripciones, comentarios, datos de última actualización, etc. - no deben incluirse en ninguna fila de un archivo CSV, sino que se incluirán en la guía metodológica del conjunto de datos, así como en los metadatos DCAT-AP correspondientes. En caso de que se necesite información adicional y cambios en cada distribución que no estén cubiertos por los anteriores casos, será preferible usar otro formato, como XML o JSON.

XLS(X)-3 (no aplicable a CSV). No incluir hojas vacías. No deben usarse hojas vacías en el formato XLS(X), porque pueden llevar a confusión.

CSV/XLS(X)-4. La primera fila (y únicamente la primera fila) debe contener el encabezado, indicando los nombres de las columnas. Como alternativa a la fila de encabezado, se puede utilizar una URI que apunta a la descripción del contenido del archivo CSV utilizando el formato CSV on the Web, que se especifica en JSON-LD. Ejemplo: "#@contexto: http://aaa/bbb/fileformat.ld". Especialmente en el caso de XLS(X) es común que estos documentos estén formateados para que puedan ser visualizados o impresos de una manera "agradable", no respetando esta regla. En este caso, al menos debería haber una distribución correspondiente en CSV, que sea fácilmente procesable.

CSV/XLS(X)-5. Los nombres de las columnas, en el encabezado, deben tener un formato de texto universal y ser entendibles por humanos. Se recomienda que los nombres de las columnas tengan la siguiente estructura:

- Sólo deben usarse letras minúsculas (a-z) y dígitos del 0 al 9.
- No se deben dejar espacios entre los nombres de las columnas.
- Las palabras separadas deben estar unidas con el carácter de subrayado (_).
- No se deben utilizar caracteres especiales (como äüöáéê, etc).

CSV-6 (no aplicable a XLS(S)). Uso de ";" como campo separador de caracteres. En CSV se pueden utilizar distintos tipos de separadores entre columnas y valores. Se recomienda el uso de ";".

CSV/XLS(X)-7. Mismo número de columnas en todas las filas. Todas las filas de un archivo CSV deben tener el mismo número de columnas, es decir, todas las filas deben tener un valor de campo por cada columna. Las columnas que estén vacías, porque se desconoce su valor, deben tener un contenido que esté descrito en un diccionario de datos y que indique claramente que el valor es desconocido, y posiblemente su causa (por ejemplo, es común usar el carácter ":" para este tipo de situaciones, aunque podría ser también "desconocido", "no disponible" o cualquier otro).

CSV/XLS(X)-8. Sólo un tipo de dato por columna. Los valores de datos en la misma columna deben ser siempre del mismo tipo de datos (ejemplo: texto, entero, decimal, fecha, hora, etc.). En el caso de una columna que contiene ambos valores, enteros y decimales, el tipo de datos debe ser decimal.

CSV/XLS(X)-9. Asignación de un ID único. Se recomienda que alguna columna (normalmente la primera) contenga un identificador del registro descrito en esa fila. En este caso, se debe cumplir, para futuras actualizaciones del archivo de datos: a) las filas de datos existentes deben mantener su ID; b) las filas de datos que no se mantengan en una actualización deben eliminarse del archivo de datos, incluido su ID, que no puede reutilizarse/reciclarse; c) las nuevas filas de datos deben contener un ID nuevo y único.

La Tabla 1 resume estas propuestas e identifica en cuáles de los documentos referenciados anteriormente aparece información relacionada:

TABLA 1. Correspondencia entre propuestas generales para ficheros en formatos CSV y XLS(X), y guías existentes sobre calidad de datos.

Propuesta	Guía datos.gob.es	Guía data.europa.eu
CSV/XLS(X)-1. Debe utilizarse el conjunto de caracteres de codificación UTF-8	--	CSV-1, XLS(X)-1
CSV/XLS(X)-2.- El fichero debe contener una única tabla de datos	--	CSV-2, XLS(X)-2
XLS(X)-3 . No incluir hojas vacías	--	XLS(X)-5
CSV/XLS(X)-4. La primera fila (y únicamente la primera fila) debe contener el encabezado	Pauta P1	CSV-3, CSV-15, XLS(X)-3
CSV/XLS(X)-5. Los nombres de las columnas, en el encabezado, deben tener un formato de texto universal y ser entendibles por humanos	Pauta P3	CSV-4, XLS(X)-4
CSV-6. Uso de ";" como campo separador de caracteres	--	CSV-5
CSV/XLS(X)-7. Mismo número de columnas en todas las filas	Pauta P5	CSV-7, XLS(X)-7
CSV/XLS(X)-8. Sólo un tipo de dato por columna	Pauta P7	CSV-8, XLS(X)-8
CSV/XLS(X)-9. Asignación de un ID único	--	CSV-14, XLS(X)-14

3.1.2 PROPUESTAS GENERALES PARA DATOS EN FORMATO JSON

JSON es un formato que se ha popularizado en los últimos años debido a la facilidad que ofrece para ser procesado con distintos tipos de librerías que se utilizan en el contexto del desarrollo Web, y por tanto debido a su utilización masiva en el desarrollo y publicación de APIs de acceso a datos. Muchas de las recomendaciones que se han proporcionado anteriormente para formatos tabulares son también aplicables a JSON, por lo que se citan a continuación con su numeración correspondiente para así facilitar la creación de checklists.

JSON-1 (asimilable a CSV/XLS(X)-1). Debe utilizarse el conjunto de caracteres de codificación UTF-8.

JSON-2 (asimilable a CSV/XLS(X)-4). Es recomendable incluir un contexto JSON-LD para describir los nombres de propiedades. En JSON no existe el concepto de columna, como en CSV o XLS(S), pero sí el de propiedad, que se entiende como una combinación de nombre de propiedad y valor que toma esa propiedad en el objeto correspondiente. Si se usa un contexto JSON-LD, se pueden dar detalles adicionales de cada uno de

esos nombres de propiedades que se usan en el fichero (URIs, tipos de datos, etc.), lo que facilita su entendimiento y utilización.

JSON-3 (asimilable a CSV/XLS(X)-5). Los nombres de las propiedades deben tener un formato universal y ser entendibles por humanos. Se recomienda que los nombres de las propiedades utilicen una estructura como la propuesta en esa regla. Asimismo, no pueden utilizar caracteres como ":" dentro del nombre.

JSON-4 (asimilable a CSV/XLS(X)-9). Asignación de una ID única. Se recomienda incluir alguna propiedad que permita identificar el registro o recurso correspondiente. En el caso de utilizar JSON-LD, se utiliza el nombre de propiedad @id. De la misma manera que en el caso de los datos tabulares, se deben cumplir las mismas recomendaciones en futuras actualizaciones del archivo de datos: a) los recursos o registros ya existentes deben mantener su ID; b) aquellos que no se mantengan deben eliminarse del archivo de datos, incluido su ID, que no puede reutilizarse/reciclarse; c) los nuevos registros deben contener un ID nuevo y único.

De la misma manera, **no son aplicables** algunas recomendaciones que habían sido propuestas en el contexto de los datos tabulares en CSV y XLS(X). Por ejemplo, no tiene sentido hablar de que el fichero debe contener una única tabla de datos (recomendación CSV/XLS(X)-2), debido a que la estructura de árbol que se puede utilizar en un fichero JSON permite anidar datos que pertenecerían a distintos tipos de recursos. Por ejemplo, en un fichero JSON con el censo de locales comerciales se podría tener en el primer nivel información de locales comerciales, haciendo referencia en un segundo nivel de indentación a las terrazas de cada uno de ellos, en caso de ser aplicable. Asimismo, frente a la recomendación CSV/XLS(X)-7, que proponía poner siempre valores, incluso cuando son desconocidos, debido al formato tabular usado, es habitual que en JSON las propiedades para las que se desconoce un valor y que no tienen una codificación específica para determinarlo no se incluyan.

Finalmente, hay algunas recomendaciones adicionales que son relevantes (además de las que se presentarán para tipos específicos de datos en la sección 3.3):

JSON-5. Es recomendable que no haya mucha anidación en los datos. No es recomendable que el nivel de anidación en los ficheros JSON sea muy grande, pues esto significa que los ficheros serán en general muy grandes y representarán muchos grupos de datos que podrían pertenecer a varios ficheros, de manera general.

3.2 PROPUESTAS PARA CONJUNTOS DE DATOS ESPECÍFICOS

En esta sección se detallan algunas propuestas a tener en cuenta para asegurar una mejor calidad de algunos tipos de conjuntos de datos que suelen ser habituales en los portales de datos abiertos, como son por ejemplo los datos estadísticos y los datos con elementos geográficos, así como para situaciones en las que los conjuntos de datos son demasiado voluminosos para que resulte práctico publicarlos de manera completa (algo que ocurre, por ejemplo, con datos de observaciones de sensores).

3.2.1 PROPUESTAS PARA CONJUNTOS DE DATOS ESTADÍSTICOS

Los conjuntos de datos estadísticos publicados en los portales de datos abiertos de las administraciones públicas (por ejemplo, en el portal de datos abiertos de un ayuntamiento) son en general producidos por sus unidades de estadística. Por esta razón, se puede considerar que en la generación de estos conjuntos de datos se han seguido buenas prácticas en la adquisición y tratamiento de los datos, en especial en relación con posibles anomalías, con datos desconocidos, etc., así como buenas prácticas en los procesos de anonimización (que se describen en más detalle en la sección 5). Asimismo, se puede considerar que existirá, en general, una guía metodológica que describa todos estos procesos, pues también es una buena práctica de las unidades de estadística. En general, es bastante común que estos datos estén disponibles de manera nativa en el formato PC-Axis, que permite representaciones de datos multidimensionales, en SDMX o en el modelo JSON-stat. En todos estos casos, las recomendaciones generales sobre codificación, nombrado de columnas y propiedades, identificadores, etc., que se han presentado anteriormente siguen siendo aplicables para el contenido de los ficheros, además de las comprobaciones sintácticas correspondientes a cada uno de estos formatos.

Por esta razón, las recomendaciones y propuestas recogidas en esta sección se centrarán fundamentalmente en los casos en los que la publicación de estos conjuntos de datos, o de vistas sobre estos conjuntos de datos, se

realice en formatos tabulares. De hecho, para ser más precisos, las recomendaciones se centrarán en la exportación en formato CSV, dado que el formato XLS(X) en este tipo de datos suele centrarse en presentar algunos datos en tablas con formatos amigables, sin centrarse en su reutilización.

En este sentido, se pueden realizar las siguientes recomendaciones:

ESTAD-1. No se deben utilizar encabezados anidados. Esta propuesta es una especialización de la propuesta CSV/XLS(X)-4. Es muy habitual en este tipo de datos que se representen columnas anidadas, como la que se presenta en la Tabla 2. Esta forma de representación, tanto en un CSV como en un XLS(X), contradice la propuesta CSV/XLS(X)-4, y debería representarse con una tabla con una cabecera única del tipo *Distrito;Población española hombres;Población española mujeres;Población extranjera hombres;Población extranjera mujeres*.

TABLA 2. Ejemplo de columnas anidadas para un conjunto de datos estadístico sobre población

Distrito	Población española		Población extranjera	
	Hombres	Mujeres	Hombres	Mujeres
Distrito-1	2345	2564	457	..
Distrito-2	..	..	..	..

ESTAD-2. No se recomienda representar tablas bidimensionales (o de más dimensiones) en el formato CSV. Esta propuesta también se puede enlazar con la propuesta presentada en el caso anterior, donde se combinan dos tipos de información (sexo y nacionalidad). Para facilitar la reutilización de este conjunto de datos, es más recomendable crear una tabla con la siguiente cabecera: Distrito;Nacionalidad;Sexo;NumHabitantes. A continuación se ve un ejemplo de cómo serían los registros correspondientes a los dos datos indicados en la Tabla 2.

Distrito;Nacionalidad;Sexo;NumHabitantes

Distrito-1;España;Hombre;2345

Distrito-1;España;Mujer;2564

Distrito-1;Extranjero;Hombre;457

ESTAD-3. Las estructuras de datos deben ser siempre estables en el tiempo, no creciendo en número de columnas. Esta recomendación no es exclusiva para datos estadísticos, pues es aplicable también a datos de observaciones de sensores, datos generales que se van acumulando en el tiempo (por ejemplo, por meses), etc. Se corresponde con la pauta P4 identificada en la guía de datos.gob.es referenciada anteriormente. En el ejemplo de la Tabla 3, correspondiente al número de locales comerciales por categoría en cada distrito, cada vez que se añade un nuevo año en el conjunto de datos se crearía una nueva columna, lo que hace el procesamiento automático de este conjunto de datos muy complicado. Es preferible que estos datos "crezcan" en número de filas según se vayan añadiendo nuevos periodos temporales, tal y como se muestra en la Tabla 4.

TABLA 3. Ejemplo de conjunto de datos cuyo número de columnas cambia con el tiempo

Distrito	Categoría	Num locales 2018	Num locales 2019	Num locales 2020
Distrito-1	Alimentación	1500	1523	1524
Distrito-1	Lavandería	14	23	45

Distrito-1	...	..	..	..

TABLA 4. Ejemplo de conjunto de datos cuyo número de columnas es estable en el tiempo

Distrito	Categoría	Año	Num locales
Distrito-1	Alimentación	2018	1500
Distrito-1	Alimentación	2019	1523
Distrito-1	Alimentación	2020	1524
Distrito-1	Lavandería	2018	14
Distrito-1	Lavandería	2019	23
Distrito-1	Lavandería	2020	45
Distrito-1	...	..	..

ESTAD-4. No se deben incluir filas o columnas de totales o subtotales, a menos que sea absolutamente necesario, manteniendo el máximo nivel de desagregación de datos posible. Esto es especialmente importante en la distribución CSV de un fichero, dado que en ocasiones las versiones en XLS(S) contienen estos datos para permitir que puedan ser impresas o introducidas en informes, como se ha comentado en ocasiones anteriores, dificultando su reutilización. Esta propuesta también se recoge en la guía de datos.gob.es, como pauta P6.

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.

3.2.2 PROPUESTAS PARA CONJUNTOS DE DATOS GEOGRÁFICOS

Como diferencia relevante con respecto a los conjuntos de datos estadísticos, los conjuntos de datos publicados en los portales de datos abiertos de las administraciones públicas (por ejemplo, en el portal de datos abiertos de un ayuntamiento) que contienen referencias geográficas pueden haber sido producidos por diversas unidades organizativas. En ocasiones, estas unidades organizativas generan los datos en formatos adecuados para su tratamiento con herramientas de gestión de datos geográficos, como pueden ser los ficheros shapefile, servicios WMS y WFS, o ficheros GeoJSON, entre otros. En estos casos, como ya se comentó en la sección anterior para el caso de los datos estadísticos, las recomendaciones generales sobre codificación, nombrado de columnas y propiedades, identificadores, etc., que se han presentado anteriormente siguen siendo aplicables para el contenido de los ficheros, además de las comprobaciones sintácticas correspondientes a cada uno de estos formatos.

Esta sección, por tanto, se centrará en las recomendaciones y propuestas que se pueden realizar en relación con la publicación de conjuntos de datos que contengan referencias geográficas (fundamentalmente coordenadas geográficas), que son muy habituales en los portales de datos abiertos. Por ejemplo, conjuntos de datos que

contienen las coordenadas geográficas de un equipamiento de la ciudad (una estación de bicicletas de alquiler, una parada de autobús, una farola, una señal de tráfico, etc.), así como direcciones postales dentro de la ciudad. Se realizan recomendaciones para que los formatos CSV, XLS(X) y JSON, entre otros, contengan la información en un formato que sea de calidad y fácilmente reutilizable por reutilizadores. Se omite en esta recomendación cualquier referencia específica a INSPIRE como modelo de intercambio de datos entre administraciones, por centrarse fundamentalmente en la facilidad de uso por reutilizadores externos y no necesariamente conocedores de este modelo. Algunas de las recomendaciones que se realizan en esta sección se recogen en la guía de datos.gob.es, pauta P15.

GEO-1. Se recomienda incluir coordenadas en latitud y longitud, además de coordenadas en otros formatos habitualmente utilizados por la administración. Una de las recomendaciones habituales en la publicación de este tipo de referencias geográficas tiene que ver con facilitar la reutilización por usuarios no expertos en datos geográficos, que pueden querer integrar los datos en herramientas de visualización que sólo manejan datos cuando tienen coordenadas esféricas, como latitud y longitud. Por tanto, en caso de que los datos estén en otro sistema de coordenadas (por ejemplo, en ETRS89), se recomienda que el conjunto de datos también contenga latitud y longitud, y que la ficha metodológica contenga información sobre el sistema de coordenadas que se utiliza para las otras coordenadas. En general, basta con representar dichas coordenadas de latitud y longitud con un máximo de 6 decimales.

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.

GEO-3. Representar las direcciones postales con un doble objetivo de reutilización interna y externa. Esta recomendación indica que en el caso de tener que representar direcciones postales, es importante contar con columnas que permitan representar la dirección postal de una manera muy detallada, para que pueda ser utilizada internamente, siguiendo pautas habituales como las de representación de tipo de vía, nombre de vía, número, planta, puerta, código postal, etc., dado que de esta manera estos datos serán utilizables internamente, y también de manera más concisa como se propone en schema.org/PostalAddress (cadena de caracteres completa para la dirección, código postal, ciudad, país), para facilitar el uso de georreferenciadores inversos y posicionamiento en mapas.

3.2.3 PROPUESTAS PARA CONJUNTOS DE DATOS MUY GRANDES (NORMALMENTE DE OBSERVACIONES DE SENSORES)

Este último punto se centra en las propuestas aplicables a conjuntos de datos que son muy grandes y que normalmente van creciendo con el tiempo, debido a que continúan acumulando datos. Este tipo de conjuntos de datos son habituales cuando se trata con datos procedentes de observaciones realizadas por sensores (observaciones de tráfico, calidad del aire, calidad del agua, etc.), donde se crean en ocasiones nuevas filas continuamente (cada minuto o conjunto de minutos).

En general, en esta situación, además de tener en cuenta todas las recomendaciones que ya han sido propuestas de manera general y para algunos conjuntos de datos como los estadísticos y geográficos (dado que estos conjuntos de datos suelen tener elementos de ambos), así como las propuestas que se realizarán en la sección 3.3 sobre algunos tipos de datos específicos (mediciones y unidades de medida, valores numéricos, etc.), se puede realizar la siguiente propuesta:

BIG-1. Dividir el conjunto de datos en distintas distribuciones de manera que cada una de ellas sea suficientemente tratable con programas informáticos habituales. Por ejemplo, herramientas como Microsoft Excel tienen limitaciones importantes con respecto al número de filas que pueden tratar al abrir documentos (algo que es habitual para este tipo de conjuntos de datos teniendo en cuenta la recomendación ESTAD-3). Actualmente, la limitación es aproximadamente de un millón de filas, pero teniendo en cuenta las características de memoria de los ordenadores de muchos potenciales usuarios este número podría ser más

reducido. Algo similar ocurre con herramientas de tratamiento de datos como Open Refine. Es por ello que se recomienda que los conjuntos de datos sean de un tamaño manejable y que puedan irse dividiendo en distintas distribuciones, atendiendo a criterios como: "ofrecer todas las observaciones correspondientes a un mes, o a un trimestre, o a un año", siempre que dichos datos puedan ser fácilmente manejables.

3.3 PROPUESTAS ESPECÍFICAS SEGÚN EL TIPO DE DATOS DE UNA COLUMNA O PROPIEDAD

En esta sección se realizan propuestas específicas según el tipo de datos que puede tener una columna (en los formatos CSV o XSL(X)) o propiedad (en JSON), atendiendo a los tipos más habituales y que más complicaciones suelen acarrear en relación con el aseguramiento de su calidad.

3.3.1 PROPUESTAS PARA LA REPRESENTACIÓN DE FECHAS

La representación de datos relacionados con fechas es uno de los aspectos en los que se necesita de una mayor estandarización dentro de una organización, así como cuando los datos se ofrecen para su reutilziación por terceros. Algunas de las propuestas que se proponen en esta sección se describen también, parcialmente, en la guía de datos.gob.es, en la Pauta P12, así como en la guía de data.europa.eu, en las recomendaciones CSV-13 y XLS(X)-13.

FECHA-1. Los valores de datos de tipo fecha y fecha/hora deben describirse en formato ISO 8601. Este estándar es propuesto por la Norma Técnica de Interoperabilidad. Para valores de datos de tipo fecha y de tipo fecha/hora, este formato ISO debe utilizarse de la siguiente manera: YYYY-MM-DD, y YYYY-MM-DDTHH:MM:SS, respectivamente. El carácter de la hora "h" o"H" no debe incluirse en el valor de la fecha. Tampoco deben incluirse datos como los días de la semana, excepto si se decide incluir en una columna separada, aunque sea redundante, por facilitar el uso en algunos tipos de aplicaciones. Aunque el estándar también permite la representación de milisegundos, son escasos los conjuntos de datos en los que esto es relevante, y por ello no se incluyen en esta recomendación. Finalmente, se recomienda también usar la zona horaria UTC siempre que sea posible, añadiendo el carácter Z al final de la cadena de caracteres.

Algunos errores que son comunes en la representación de fechas son los siguientes:

- Fechas que combinan texto y año (por ejemplo, mayo-2021, 2021/05, 4-mayo-2021).
- Fechas donde se desconoce el formato específico (dd/mm/aa, mm/dd/aa).
- El uso de fechas como 01/01/aaaa por defecto, cuando sólo se tiene información parcial.

FECHA-2. Los periodos temporales deben incluirse como dos columnas/propiedades. Para incluir un período temporal o una duración, se deben usar dos columnas: una para la fecha o fecha/hora de inicio y otra para la fecha o fecha/hora de finalización. Ambos valores deben seguir las pautas indicadas en FECHA-1.

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)

3.3.2 PROPUESTAS PARA LA REPRESENTACIÓN DE DATOS NUMÉRICOS

Algunas de estas recomendaciones se describen también, parcialmente, en la guía de datos.gob.es, Pauta P11, así como en la guía de data.europa.eu, en las recomendaciones CSV-9, CSV-10, CSV-11, XLS(X)-9, XLS(X)-10 y XLS(X)-11.

NUM-1. Representar los decimales con un punto. El caracter de punto (.) debe utilizarse sólo para valores decimales, tal y como se usa en los formatos en inglés. Para columnas de tipo entero, todos los valores enteros deben ir sin punto decimal y sin decimales. Esto permite generar archivos de datos más compactos y permite una identificación visual del tipo de columna como un número entero. Para las columnas de tipo decimal, todos los valores numéricos deben incluir el punto decimal e idealmente el mismo número de decimales, aunque este aspecto no es obligatorio.

NUM-2. No se deben utilizar caracteres de formato de "miles". Los valores numéricos (enteros, decimales) sólo pueden tener los dígitos 0-9, el carácter "-" para números negativos y el punto (.) para los decimales. No pueden contener caracteres de formato/separadores, como "," o blanco para la posición "miles" o símbolos de moneda, como €, £, $.

NUM-3: No se deben incluir ceros a la izquierda. Los valores enteros o decimales no deben incluir ceros a la izquierda. En el caso de que los ceros a la izquierda sean obligatorios y se deban conservar (por ejemplo, en determinados códigos, números de DNI, etc.), el tipo de datos no debe ser numérico, sino de tipo texto, y por tanto los valores de los datos deben ir entre dobles comillas (""), como se indica en la sección 3.3.4.

3.3.3 PROPUESTAS PARA LA REPRESENTACIÓN DE UNIDADES DE MEDIDA Y MONEDAS

Algunas de estas recomendaciones se describen también, parcialmente, en la guía de datos.gob.es, Pauta P11, así como en la guía de data.europa.eu, en las recomendaciones CSV-12 y XLS(X)-12.

UNID-1. Las unidades de medida y monedas deben indicarse por separado, o en el nombre de las columnas. Las unidades de medida para valores numéricos (por ejemplo, km/h, mg/m3) o divisas (por ejemplo, euros) no deben mezclarse con los valores numéricos en la misma columna. En caso de que las unidades de medida o divisas sean las mismas para todos los valores de una columna, la unidad se puede unir al nombre de la columna, como "Cantidad_eur", para facilitar su comprensión, además de especificar la unidad en la guía metodológica. En caso de que las unidades de medida o divisas puedan tener diferentes valores para distintos registros, entonces la unidad de medida o divisa debe ponerse en una columna separada (preferiblemente a continuación de la columna de valor numérico). Ejemplos de unidades:

- "descripción"; "cantidad"; "moneda"
- "Artículo-1"; 123.45; "EUR"
- "Artículo-2"; 67.89; "USD"

3.3.4 PROPUESTAS PARA LA REPRESENTACIÓN DE DATOS TEXTUALES

Algunas de estas recomendaciones se describen también, parcialmente, en la guía de datos.gob.es, Pautas P2, P10 y P13, así como en la guía de data.europa.eu, en las recomendaciones CSV-6 y XLS(X)-6.

TXT-1. Uso de cadenas de caracteres con comillas dobles. Los valores en los campos de texto deben ir, de manera recomendada en el caso de CSVs y de manera obligatoria en el caso de ficheros JSON, marcados con comillas dobles (" "). Esto es especialmente útil, en el caso de los CSVs, en aquellos casos en los que los caracteres especiales (punto y coma, comillas, retornos de carro, etc) forman parte de los valores del campo. Por ejemplo, se pueden dar las siguientes situaciones:

- En un CSV, en caso de que los valores de un campo textual incluya alguna comilla doble ("), este carácter debe ser duplicado, es decir, debe aparecer dos veces (por ejemplo: "este texto cuenta con dos ""caracteres de comillas dobles"".").
- En un JSON, en el caso anterior, este carácter debe ser precedido del carácter \ (por ejemplo: "este texto cuenta con dos \"caracteres de comillas dobles\".").

- En un CSV, en caso de que los valores del campo necesiten incluir el carácter separador de campo (;) o un carácter de retorno de carro, entonces el valor del campo debe ir entre comillas dobles. (ejemplo: "Este es el texto que incluye el carácter de punto y coma ;".
- En un JSON, en caso de que los valores del campo necesiten incluir este carácter también, se debe duplicar (ejemplo: "Este es el texto que incluye el carácter de barra invertida \\").

3.4 HERRAMIENTAS DE VALIDACIÓN

En resumen, como se ha descrito a lo largo de esta sección existen muchas comprobaciones que deben hacerse para asegurar la calidad de los datos abiertos que se publican. Para algunas de estas comprobaciones, también se dispone de soporte tecnológico. Algunas de estas herramientas están identificadas en el artículo publicado por ODI, con título "*How hard is it to publish good-quality open data?*"[36]:

- CSVLint[37] es una herramienta desarrollada por ODI para ayudar a los analistas y bibliotecarios de datos a verificar que un archivo CSV es legible, realizando comprobaciones como la falta de cabeceras o nombres de columnas, la revisión de filas en el fichero cuyo número de columnas no se corresponde con el número de columnas de la cabecera, filas en blanco, caracteres extraños que podrían causar errores, uso de comillas dobles y simples que no estén balanceadas, valores posiblemente inconsistentes (por ejemplo, una columna que contiene una gran cantidad de valores numéricos pero donde algunas filas contienen textos), etc. También permite verificar un CSV de acuerdo con una descripción del esquema que debería seguir.
- Para facilitar el uso de CSVLint, la herramienta Octopub[38], también desarrollada por ODI, está disponible como una aplicación en GitHub que proporciona los mecanismos de validación de CSVLint a través de una interfaz Web de usuario, facilitando su uso.
- Otras herramientas de similares características son Lintol[39], que trabaja con lo que se denominan perfiles de datos (*data profiles*), o *goodtables*[40], que proporciona opciones de validación continua de datos para hojas de cálculo en formatos CSV, Excel y LibreOffice, entre otros.

En el ámbito nacional, Data Koality[41] es una herramienta web que permite analizar archivos CSV a partir de las recomendaciones de publicación de datos tabulares de datos.gob.es, anteriormente mencionada. Desarrollada por un equipo del Open Summer of Code 2021, es capaz de generar un informe y resultados cualitativos sobre el conjunto de datos analizado, identificando las pautas de publicación que no se cumplen y recomendando posibles mejoras.

Finalmente, en el caso de formatos como JSON, también se puede identificar la herramienta JSONLint[42]. Esta herramienta se centra únicamente en la verificación sintáctica de los ficheros JSON.

Asimismo, también existen guías con recursos para ayudar a encontrar y elegir estándares abiertos para los datos[43], así como guías con instrucciones paso a paso para mejorar el flujo de trabajo de publicación de datos (por ejemplo, *Frictionless Data Field Guide*[44]).

36 https://theodi.org/article/how-hard-is-it-to-publish-good-quality-open-data/#1538642561561-bdc33070-74ab
37 http://csvlint.io/
38 https://octopub.io/
39 https://lintol.io/
40 https://goodtables.io/
41 https://osoc-es.github.io/data-quality-madrid/
42 https://jsonlint.com/
43 https://standards.theodi.org/find-existing-standards/
44 https://frictionlessdata.io/tag/field-guide/

4. PROPUESTAS PARA ASEGURAR LA CALIDAD DE LOS METADATOS

En esta sección se realizan propuestas relacionadas con el aseguramiento de la calidad de los metadatos que se ofrecen en los portales de datos abiertos de las administraciones públicas (y más específicamente de los ayuntamientos). Se realiza en primer lugar un análisis de algunas fichas de metadatos y fichas metodológicas disponibles en varios portales de datos abiertos, y basado en este análisis se hace una propuesta de ficha metodológica tipo para cualquier conjunto de datos abiertos.

4.1 FICHAS DE METADATOS EN PORTALES DE DATOS ABIERTOS DE MUNICIPIOS ESPAÑOLES

Es habitual que tanto las fichas de metadatos como las fichas metodológicas asociadas a los conjuntos de datos de distintos portales de datos abiertos reciban nombres diferentes: ficha de metadatos, información adicional, información, etc. No presentan uniformidad, lo que hace compleja la elaboración de una comparativa sistemática, así como la propia búsqueda y comprensión de los datos por parte de los reutilizadores. Esta última parte es, además, una de las puntualizaciones de la Unión Europea al aplicar las especificaciones para que los registros de metadatos cumplan con las recomendaciones indicadas en el informe *About DCAT Application Profile for data portals in Europe*[45], donde se indica que los reutilizadores de datos *"pueden obtener una visión general de qué conjuntos de datos existen y qué administraciones públicas los mantienen"* si se aplican las recomendaciones del Vocabulario de Catálogos de Datos (en inglés, DCAT) para la publicación de metadatos en los portales de datos abiertos.

Existen casos concretos, como el del portal de datos del Ayuntamiento de Madrid, donde se puede encontrar publicada una ficha adicional con una estructura clara de los metadatos que se deben publicar para todos los conjuntos de datos. En esta ficha se indica qué metadatos debe incluir cada uno de los conjuntos de datos que componen el catálogo: nombre y/o título del conjunto de datos, descripción general, fecha de incorporación al catálogo, sector, etc. Esta guía se cumple de manera general, como se ha observado en el análisis que se ha realizado para la elaboración de este informe.

En otros portales de datos, sin embargo, no se ha localizado ningún documento de carácter general, aplicable a todo el catálogo, que defina cuáles deben ser los metadatos que se deben incluir para la descripción de los conjuntos de datos del catálogo, más allá de lo que se puede inferir por el necesario uso de DCAT-AP. Asimismo, en general, el nivel de detalle que se ofrece no es muy grande y en ocasiones no es suficientemente sistemático. Por ejemplo, en el portal de datos abiertos de Santiago de Compostela aparece información sobre los distintos idiomas en los que se publican los datos, que sería uno de los elementos del tercer nivel -opcional-, pero se agrupan otros que forman parte de niveles anteriores -obligatorio o recomendado-.

La Tabla 5 ofrece algunos ejemplos de conjuntos de datos con y sin ficha de metadatos asociadas, en las ciudades participantes en el proyecto Ciudades Abiertas, así como detalles sobre la información adicional relacionada con las fichas de metadatos de cada ciudad.

TABLA 5. Algunas fichas de metadatos de portales de datos abiertos.

Ciudad	Ficha de metadatos	Conjuntos de datos	Información	Información adicional Documentación asociada
A Coruña	NO	Demografía: habitantes por edad, distrito y sexo	Sin ficha	Normas uso del portal

45 https://joinup.ec.europa.eu/rdf_entity/http_e_f_fdata_ceuropa_ceu_fw21_fac376c94-74cf-4dd7-ade7-267d6a4ec4dc

Madrid	SI	Actuaciones limpieza urbana	Estructurada según la ficha	Portal de Transparencia
Santiago de Compostela	NO	Observacións de calidade de aire	Sin ficha	Documentación general sobre metadatos
Zaragoza	NO	Licencias urbanísticas	Sin ficha	Documentación general del portal

A modo de resumen, la información sobre metadatos es relativamente fácil de encontrar en los portales analizados, aunque tan sólo uno de ellos (el de Madrid) cuenta con una **ficha completa que indica cuáles deben ser los elementos analizados y los términos utilizados**, de los se debe encontrar información en cada conjunto de datos. Al seguir una estructura identificada desde el principio, la búsqueda de cualquier aspecto relacionado con los metadatos de un conjunto de datos es mucho más sencilla y clara.

4.2 PROPUESTA DE FICHA METODOLÓGICA TIPO

Se han analizado todos los conjuntos de datos del Ayuntamiento de Madrid. A partir de este análisis se han escogido 20 conjuntos de datos cuyas fichas metodológicas se han considerado las más completas, teniendo en cuenta también que hubiera suficiente variedad de unidades organizativas productoras de los datos y temáticas asociadas. Estas fichas se enumeran en la Tabla 6, junto con el nombre del conjunto de datos al que dan soporte y el sector al que éste pertenece. Se ha realizado un análisis de estas fichas metodológicas desde su dimensión como herramienta que busca proporcionar información, resumida y concreta, de cómo se analiza e interpretan los indicadores o las variables que conforman cada conjunto de datos.

Tabla 6. Tabla de fichas metodológicas analizadas.

	Conjunto de datos	**Ficha metodológica**	**Sector**	**Ciudad**
1	Aforo de peatones y bicicletas	Estructura fichero Aforo de peatones y bicicletas	Transporte	MADRID
2	Ocupación de la vía pública. Histórico de rodajes	Estructura fichero de datos. Histórico de ocupación de la vía pública	Transporte	MADRID
3	Consumo de energía en edificios municipales. Datos mensuales	Estructura de datos de consumo energía edificios municipales	Energía	MADRID
4	Inventario de instalaciones fotovoltaicas	Estructura del fichero de datos. Inventario de instalaciones fotovoltaicas	Energía	MADRID
5	Encomiendas de gestión a sociedades municipales	Estructura de fichero de datos. Encomiendas de gestión	Energía	MADRID
6	Contaminación acústica. Datos históricos diarios	Intérprete y estructura el archivo de datos	Medio Ambiente	MADRID
7	Contenedores de aceite vegetal usado	Estructura del fichero de contenedores de aceite vegetal usado	Medio Ambiente	MADRID
8	Denuncias y sanciones sobre limpieza urbana y zonas verdes	Estructura de datos del fichero de sanciones de limpieza 2020	Medio Ambiente	MADRID

9	Sedes. Institutos de Investigación	Estructura del conjunto de datos	Ciencia y Tecnología	MADRID
10	Inventario de Patrimonio Municipal del Suelo	Estructura del fichero de Patrimonio Municipal del Suelo	Urbanismo e Infraestructuras	MADRID
11	Licencias urbanísticas otorgadas, declaraciones responsables y comunicaciones previas	Estructura del fichero de datos estructuras urbanísticas otorgadas	Urbanismo e Infraestructuras	MADRID
12	Tráfico Calle 30. Histórico de incidencias y accidentes	Estructura de datos del fichero Calle 30 Histórico de incidencias y accidentes	Urbanismo e Infraestructuras	MADRID
13	Inspecciones de consumo	Estructura de datos de inspección de consumo	Comercio	MADRID
14	Mercamadrid: volumen y precio de productos comercializados	Documento de estructura. Mercamadrid volumen y precio	Comercio	MADRID
15	Padrón Municipal. Histórico	Conceptos fundamentales y estructura del fichero de datos de padrón municipal	Demografía	MADRID
16	Agencia para el Empleo. Perfiles de personas inscritas	Estructura de datos. Perfiles inscritos en agencia para el empleo	Empleo	MADRID
17	Madrid Salud. Inspecciones para el control oficial de alimentos en establecimientos alimentarios	Estructura de los datos es establecimientos alimentarios	Salud	MADRID
18	Cámaras de Videovigilancia en la Vía Pública de Madrid	Estructura de datos de cámaras de videovigilancia	Seguridad	MADRID
19	Unidades Integrales de Distrito de Policía Municipal	Estructura del conjunto de datos	Seguridad	MADRID
20	Viviendas en alquiler adjudicadas por la Empresa Municipal de la Vivienda y el Suelo (EMVS)	Estructura de datos del fichero Viviendas en alquiler adjudicadas	Vivienda	MADRID

El análisis de las fichas de metadatos ha resultado complejo, dado que, si bien existen guías con recomendaciones, no se ha descrito ningún modelo obligatorio (asociado) para la publicación de esta documentación. La información que contienen las fichas es muy dispar, el nombre para su localización es distinto y, en ocasiones, se publica información adicional que no es ficha metodológica y lleva a confusión. No todos los conjuntos de datos cuentan con ficha metodológica para la interpretación de los datos.

En resumen, aún habiendo una ficha metodológica tipo en este ayuntamiento, el análisis de este conjunto de fichas metodológicas demuestra que **existen bastantes diferencias entre distintos conjuntos de datos**. En algunas se incluye información completa sobre la unidad o departamento propietario de los datos mientras que en otras no, sobre los formatos en que se publican los datasets, las fechas de publicación y/o actualización, así como datos concretos de nombres de calles y números, códigos de actuación, formato de publicación de los datos o información adicional relativa al conjunto de datos al que acompañan.

A partir del análisis realizado, se ha realizado una propuesta de una ficha metodológica tipo que podría ser utilizada por cualquier ciudad para la descripción de los conjuntos de datos que se publican en el portal de datos

abiertos. Esta ficha metodológica es la que se utilizará en el contexto del proceso de aseguramiento de la calidad que fue propuesto en la sección 2 de este informe. En esta ficha se incluyen las equivalencias en DCAT de algunos de estos elementos (lo que es útil para derivar el registro DCAT que se propone en la fase 1 de nuestro proceso de aseguramiento de calidad) y también se incluye una referencia de si ya está abordado por la ficha metodológica del Ayuntamiento de Madrid y si se incluye entre las categorías que se analizan en el Metadata Quality Assessment de data.europa.eu.

Tabla 7. Propuesta de ficha metodológica.

Elemento	**Descripción**	**Equivalente en DCAT**	**Madrid**	**Categoría MQA**
Descripción del conjunto de datos				
Identificador	Palabras clave que ayudan a la búsqueda y la facilidad de localización del conjunto de datos	dct:identifier	SI	Facilidad de localización
Descripción	Descripción del conjunto de datos	dct:description	SI	Facilidad de localización
Datos de contacto				
Organización y unidad de contacto	Este campo representa a la unidad organizativa que actúa como punto de contacto. Se recomienda que sea una URI usando DIR3, siempre que sea posible	dcat:contactPoint → vcard	NO	Accesibilidad
Nombre y cargo la persona de contacto	Este campo representa a la persona que actúa como punto de contacto dentro de la unidad organizativa	dcat:contactPoint → vcard	NO	Accesibilidad
Dirección postal, correo electrónico y número de teléfono	Este campo representa la información de contacto de la persona que actúa como punto de contacto dentro de la unidad organizativa	dcat:contactPoint → vcard	NO	Accesibilidad
Actualización y difusión de datos				
Última actualización/difusión de datos	Última fecha en la que se han actualizado los datos del conjunto de datos	dct:modified	SI	Contextualidad
Fecha de generación inicial de datos	Fecha en la que se creó el conjunto de datos	dct:issued	SI	Contextualidad

Frecuencia de actualización/difusión	Frecuencia con que se actualiza un conjunto de datos	dct:accrualPeriodicity	SI	Contextualidad
Formato de difusión: comunicados, newsletters, publicaciones, vídeos, etc.	Difusión de las actualizaciones de conjuntos de datos y formato de dichas actualizaciones	--	NO	Contextualidad

Sector/Temática (Clasificación NTI)				
Sector/Temática	Información que ayuda a los usuarios a explorar los conjuntos de datos atendiendo a su temática, de acuerdo con las temáticas identificadas en la NTI.	dcat:theme	SI	Facilidad de localización
Cobertura				
Cobertura temporal Periodo de referencia (año, mes, semana, etc.)	Información temporal que permite a los usuarios encontrar conjuntos de datos con una búsqueda acotada en el tiempo.	dct:spatial	NO	Facilidad de localización
Cobertura espacial	Información espacial que permite a los usuarios encontrar conjuntos de datos con una búsqueda acotada geográficamente.	dct:temporal	NO	Facilidad de localización
Parámetros de calidad				
Completitud	Describe cuántos datos están presentes en el conjunto de datos.	--	NO	Contextualidad
Unicidad	Describe el grado de no duplicación de registros.	--	NO	Contextualidad
Consistencia	Describe el grado en el que los valores de un conjunto de datos no contradicen otros valores que representen a la misma entidad dentro del mismo conjunto de datos.	--	NO	Contextualidad
Pertinencia en el tiempo	Describe el grado en el que los datos reflejan de manera precisa	--	NO	Contextualidad

	el periodo que representan, y en el que los valores están actualizados.			
Validez	Describe el grado en el que los datos están disponibles en el rango y formato esperados.	--	NO	Contextualidad
Exactitud	Describe el grado en el que los datos se corresponden con la realidad, indicando potenciales sesgos en su recolección o tratamiento posterior.	--	NO	Contextualidad
Procesos de aseguramiento de calidad seguidos	Describe los procesos utilizados para asegurar que se cumplen los seis puntos anteriores.	--	NO	Contextualidad
Procesos de anonimización realizados	Describe los procesos seguidos para asegurar la anonimización de los datos de cada conjunto de datos.	--	NO	Contextualidad

Descripción de los datos				
Enlace a diccionario de datos	Proporciona un enlace al diccionario de datos que describe en más detalle la estructura de los datos del conjunto de datos.	--	NO	Interoperabilidad
Sistemas de clasificación (SKOS, etc.) utilizados	Identifica los sistemas de clasificación utilizados dentro del conjunto de datos.	--	NO	Interoperabilidad
Vocabularios utilizados	Identifica los vocabularios utilizados dentro del conjunto de datos (especialmente relevante en el caso de las distribuciones en formato RDF)	dct:conformsTo	NO	Interoperabilidad
Marco normativo/legislativo				
Normas jurídicas que sustentan los conjuntos de datos	Referencias a los recursos que informan al usuario sobre los derechos de los que dispone cuando utilice el conjunto de datos.	--	NO	Contextualidad
Política de confidencialidad y tratamiento	Indicaciones de si el acceso a los datos es público o si hay alguna restricción dictada por la	--	NO	Reusabilidad

de datos confidenciales (restricciones de acceso)	legislación. Esta propiedad está relacionada con la de anonimización			
Licencia	Describe la licencia aplicable a los datos	dct:license	SI	Reusabilidad
Costes del proceso				
Costes	Coste aproximado asociado a la creación y mantenimiento de los datos	--	NO	--
Observaciones generales				
Observaciones	Cualquier tipo de observación adicional	rdfs:comment	NO	Contextualidad

5. ANONIMIZACIÓN DE DATOS

La anonimización de datos es una tarea que debe ser considerada dentro del proceso de publicación de datos abiertos (y por tanto también como parte de las tareas de aseguramiento de la calidad) en el caso de tratar con datos privados, sensibles o personales, como son aquellos que contienen información de carácter personal y que por tanto están protegidos por las leyes y regulaciones relacionadas con la protección de datos de carácter personal. Esto se indica, por ejemplo, en el Reglamento de Protección de Datos (RGDP) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016[46], que apunta a la necesidad de *"protección de las personas físicas en relación con el tratamiento de datos personales"* como derecho fundamental, aludiendo al artículo 8 de la Carta de los Derechos Fundamentales de la Unión Europea y al artículo 16 del Tratado de Funcionamiento de la Unión Europea (TFUE), que establecen que *"toda persona tiene derecho a la protección de los datos de carácter personal que le conciernan"*.

Sin embargo, el hecho de que un conjunto de datos contenga este tipo de datos no debe ser necesariamente un impedimento para que se pueda publicar y por tanto pueda ser reutilizados tanto para la propia administración pública como por reutilizadores externos, respetando en todo momento estas leyes y regulaciones. De acuerdo con la guía "Orientaciones y garantías en los procesos de anonimización de datos personales"[47], publicada por la Agencia Española de Protección de Datos, la finalidad del proceso de anonimización de datos es la de *"eliminar o reducir al mínimo los riesgos de re-identificación de los datos anonimizados, manteniendo la veracidad de los resultados del tratamiento de los mismos; es decir, además de evitar la identificación de las personas, los datos anonimizados deben garantizar que cualquier operación o tratamiento que pueda ser realizado con posterioridad a la anonimización no conlleva una distorsión de los datos reales"*. En otras palabras, al mismo tiempo que se elimina o reduce al mínimo la posibilidad de identificar a cualquier persona a partir de un conjunto de datos o de la combinación de varios conjuntos de datos, el uso o reutilización de los datos anonimizados no debe diferir del uso que se haría del conjunto de datos si no estuviesen anonimizados.

Con el fin de cumplir con la normativa europea, se hacen necesarias técnicas de anonimización de datos que permitan garantizar, desde las fases iniciales de publicación de un conjunto de datos abiertos, la privacidad y no identificación personal. Para ello hay que tener en cuenta la posibilidad de realizar una Evaluación de Impacto en Protección de Datos (EIPD)[48]. Una EIPD es una herramienta preventiva con la que se pueden identificar, evaluar y gestionar los riesgos a los que están expuestas sus actividades de tratamiento de datos con el objetivo de garantizar los derechos y libertades de las personas físicas. Conviene disponer de una metodología que considere los requisitos del RGPD, donde se establece que una EIPD deberá incluir como mínimo los siguientes elementos: una descripción sistemática de la actividad de tratamiento prevista, una evaluación de la necesidad y proporcionalidad del tratamiento de los datos respecto a su finalidad, una evaluación de los riesgos (indicando qué nivel de riesgo sería aceptable en función del proyecto), y las medidas previstas para afrontar esos riesgos (garantías y medidas de seguridad y mecanismos que garanticen la protección de datos personales).

Por tanto, el concepto de protección de datos no debe abordarse únicamente en la fase de publicación de los datos como datos abiertos, sino desde los primeros pasos de su gestión (algo que ya se hace en general con toda la información sensible), y mantenerse a lo largo de todo el proceso. Para poder establecer la anonimización hay que tener en cuenta los siguientes principios: **proactividad**, **privacidad por defecto**, **privacidad objetiva**, **plena funcionalidad**, **privacidad en el ciclo de vida de los datos** y **formación e información**.

- **Proactividad:** la anonimización debe establecerse desde el principio, de manera proactiva y no reactiva. La privacidad no puede garantizarse a posteriori. Es conveniente realizar una clasificación inicial de los datos y disponer de una escala o gradiente de sensibilidad desarrollando, por ejemplo, un esquema de clasificación con un sistema basado, como mínimo, en tres niveles de identificación de personas (microdatos, datos de identificación indirecta y datos sensibles), donde se asigne un valor cuantitativo a cada una de las variables de identificación. Esta escala puede ser clave en todo el proceso.

[46] https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R0679&from=ES
[47] https://www.aepd.es/sites/default/files/2019-12/guia-orientaciones-procedimientos-anonimizacion.pdf
[48] https://www.aepd.es/sites/default/files/2019-09/guia-evaluaciones-de-impacto-rgpd.pdf

- **Privacidad por defecto:** desde el principio se debe atender a la privacidad, teniendo en cuenta la granularidad que deben tener los datos anonimizados. Una escala como la descrita puede ser de ayuda para este punto.
- **Privacidad objetiva:** hay que establecer un nivel de riesgo aceptable sobre el que se trabajará, que debe ser conocido por quienes manejan los datos y por los posibles usuarios, informando del riesgo de uso de determinado conjunto de datos.
- **Plena funcionalidad:** hay que tener en cuenta la utilidad de los datos anonimizados, garantizando que no haya muchas modificaciones en su uso con respecto a las que habría con esos datos no anonimizados.
- **Privacidad en el ciclo de vida de los datos:** hay que garantizar la privacidad durante todo el ciclo de vida de los datos, partiendo de los datos sin anonimizar. Si es necesario, en el inicio del proceso de anonimización se pueden eliminar aquellas variables de identificación que no se consideren necesarias o que no fuera posible anonimizar.
- **Formación e información:** es clave, para que funcione bien el proceso, que todos los implicados cuenten con la formación necesaria para anonimizar los datos y con la información que haga falta para mantener los datos limpios durante todo su ciclo de vida.

Asimismo, es común añadir en el portal de datos abiertos una cláusula, entre los términos de reutilización, que limite el uso de los datos cuando se pretenda re-identificar a personas basado en los datos personales.

5.1 PROCESO DE ANONIMIZACIÓN DE CONJUNTOS DE DATOS

En general, para cada conjunto de datos que pueda ser susceptible de contener datos sensibles, se debe seguir el siguiente conjunto de pasos para asegurar su anonimización:

1. **Identificación de identificadores directos** entre todas las variables/columnas/atributos de los datos. En este paso se examinará la ficha metodológica del conjunto de datos, junto con el diccionario de datos, para identificar posibles identificadores directos, que deberán ser tratados adecuadamente posteriormente. Además del uso de la ficha metodológica y el diccionario de datos, en caso de tener acceso a los datos, se podrán identificar estos posibles identificadores mediante técnicas sencillas basadas en el facetado de las columnas (por ejemplo, con herramientas como OpenRefine): una columna donde todos los valores son distintos es susceptible de ser un identificador directo.
2. **Identificar identificadores o grupos de identificadores indirectos**. En este paso, se examinará de nuevo la ficha metodológica y el diccionario de datos para poder identificar grupos de identificadores que pueden dar lugar a la identificación indirecta de un registro (por ejemplo, combinaciones de marca y modelo de coche con la sección censal donde están registrados, en el caso de coches poco habituales).
3. **Análisis global del conjunto de datos**. En caso de tener acceso al conjunto de datos, se pueden realizar distintos tipos de análisis estadísticos de columnas individuales o de combinaciones de columnas para analizar si se pueden encontrar casos especiales que permitan identificar un individuo en un conjunto de datos (por ejemplo, en datos agregados - número de habitantes de un país de procedencia por sección censal, en situaciones en las que el número es muy bajo - por ejemplo por debajo de 3).
4. **Eliminación (o pseudoanonimización) de identificadores directos**. Dependiendo del uso esperado del conjunto de datos por parte de terceros, se deberá decidir si se procede con la eliminación de los mismos (por ejemplo, eliminar el atributo DNI) o se genera un pseudoidentificador (por ejemplo, aplicando un código hash). Se proporcionan más detalles en la sección 5.2.
5. **Agregación u ofuscación de identificadores indirectos.** En el caso de los identificadores indirectos, se debe analizar cómo pueden ser agregados u ofuscados (mediante técnicas de generalización, redondeo, perturbación, etc.) para que no puedan dar lugar a re-identificación.
6. **Reanalizar el conjunto de datos resultante para analizar cualquier riesgo adicional**. En caso de detectar algún riesgo adicional, volver al paso 1.

Todos estos detalles relacionados con la anonimización aplicada a un conjunto de datos deben ser claramente descritos en la ficha metodológica correspondiente.

5.2 TÉCNICAS DE ANONIMIZACIÓN APLICABLES A CONJUNTOS DE DATOS

La anonimización de datos se centra en eliminar o enmascarar **identificadores directos** de las personas, como pueden ser el nombre, el número del DNI, de la Seguridad Social, direcciones o cuentas bancarias, etc; así como las combinaciones poco habituales de datos (denominadas **combinaciones de identificadores secundarios**), que podrían facilitar la identificación personal, tales como la combinación de fecha de nacimiento junto al código postal o identificador de barrio. En general, este proceso busca eliminar u ofuscar datos o combinaciones de datos que puedan facilitar la identificación de un ciudadano o ciudadana. Por ejemplo, el resultado de la aplicación de una técnica de anonimización puede ser la transformación de una fecha de nacimiento exacta en un intervalo (por ejemplo, pasando la fecha exacta 11/12/1980 a un año **/**/1980), o la eliminación en la precisión de un dato de latitud y longitud (por ejemplo, reduciendo el número de decimales usados para estas coordenadas a dos dígitos, lo que genera un círculo donde se encuentra la posición exacta que ha sido anonimizada).

De manera general, hay que distinguir entre dos grandes bloques de técnicas de anonimización: la pseudoanonimización y la anonimización.

- En la **pseudoanonimización**, se reemplazan de manera arbitraria unos identificadores por otros. Esto permite que se puedan *"procesar datos personales de tal manera que los datos no sean atribuibles a un dato específico concreto sin el uso de información adicional"*. Por ejemplo, en un conjunto de datos que contiene todos los usos de una tarjeta de transporte público, el número de dicha tarjeta puede ser reemplazado por un código hash generado a partir del número original, de tal manera que dicho número original no pueda ser reconstruido con facilidad, reduciendo la posibilidad de enlazar el conjunto de datos original con otros conjuntos de datos (por ejemplo aquellos que contienen datos personales asociados a la tarjeta). Este código hash podría de hecho generarse a partir de una combinación de identificadores secundarios (número de tarjeta, fecha de nacimiento, nombre y apellidos). La pseudoanonimización es por tanto un método que reduce la enlazabilidad de un conjunto de datos con la identidad original de los datos, y puede ser una herramienta útil según el tipo de conjunto de datos y el uso esperado del mismo, pero no cuando se tiene que respetar el RGPD.
- La **anonimización** es en cambio un proceso completamente irreversible donde se eliminan por completo los datos personales. En el caso anterior, esto consistiría en eliminar completamente la columna referente al número de tarjeta de transporte público, aunque como resultado inicial se perdería la posibilidad de hacer un recuento del número de veces que ha sido utilizada cada tarjeta, o de los trayectos realizados, lo que puede disminuir su utilidad en algunos tipos de procesos de transformación y tratamiento de los datos. Es decir, este proceso permitiría seguir realizando estadísticas de uso de cada parada, pero no permitiría realizar análisis de trayectos habituales de partes de la población a lo largo del tiempo.

Algunas de las técnicas específicas que se pueden utilizar para estos procesos de anonimización son las siguientes:

- **Generalización de los datos**. La generalización consiste en el uso de datos generales en lugar de datos específicos que permitan conocer una identidad concreta. Ya se han dado previamente algunos ejemplos relacionados con fechas (convirtiendo fechas concretas en rangos) y con posiciones geográficas (reduciendo la precisión de los valores de latitud y longitud). También puede aplicarse a otros atributos como edades (generalizando en franjas de edad en lugar de ofrecer edades concretas). Y también pueden aplicarse otras transformaciones como redondeos de cifras.
- **Ofuscación de los datos**. En estos casos, se puede sustituir un dato por otras combinaciones de números, dígitos y caracteres especiales (por ejemplo, sustituyendo un código postal por una combinación de números y asteriscos, o un DNI por una combinación de números y asteriscos).
- **Supresión de los datos**. En estos casos, se puede sustituir un dato concreto completamente, reemplazándolo con un asterisco (o cualquier otro caracter o conjunto de caracteres). En este caso se pierde la información correspondiente a este valor, algo que sólo sería relevante para un reutilizador en el caso de que la presencia de este valor fuera útil para el tratamiento posterior.
- **Perturbación de los datos**. En estos casos se modifican los valores de algún atributo o conjunto de atributos de manera sistemática, de tal manera que las cifras no sean suficientemente precisas para

revelar datos individuales. Por ejemplo, pueden intercambiarse datos entre celdas de distintos registros, de tal manera que medidas de agregación como la media, desviación típica, etc., se seguirán manteniendo iguales, aunque no sean aplicables a los individuos concretos. Este tipo de técnica se aplica habitualmente en censos de población.

- **Muestreo de los datos**. Se eliminan algunos registros, siempre que la muestra de datos sea suficientemente grande, de tal manera que las propiedades estadísticas del conjunto de datos original y de la muestra son similares. Esta técnica sólo es adecuada para algunos tratamientos estadísticos de los datos, y no es recomendable para la publicación de datos detallados como datos abiertos.
- **Disociación de datos**. En este caso, se separan los datos personales de los no personales, en distintos conjuntos de datos, por lo que se podría publicar el conjunto de datos que no contiene datos personales con garantías.
- **Eliminación de "outliers"**. En el caso de datos continuos, por ejemplo, puede ser útil determinar los valores mínimo y máximo del rango de valores, para así poder eliminar lo que se conocen como "outliers", que pueden ser más fácilmente identificados. Por ejemplo, cuando se publica el salario de personas, aquellas con salarios muy diferentes a los demás podrían ser identificadas.

En general, en los datos anonimizados se persigue que se cumplan algunas características concretas. Por ejemplo, una característica habitual es la que se denomina **k-anonimización**. Se dice que un conjunto de datos tiene la propiedad de k-anonimato (o es k-anónimo) si la información de todos y cada uno de los registros en ese conjunto es idéntica al menos con otras k-1 personas que también aparecen en dicho conjunto. Para comprender mejor esta propiedad se incluye en la Tabla 8 un breve ejemplo de datos personales de un censo de habitantes ficticio, y un posible conjunto de datos anonimizados a partir de esta tabla que cumple esta propiedad (Tabla 9).

Tabla 8. Ejemplo ficticio de un censo de habitantes.

DNI	Género	Edad	Nivel estudios	Religión	Sección censal
53106334X	Hombre	29	3	Católica	11-012
53106335Y	Hombre	32	3	Católica	11-012
53106336Z	Mujer	35	3	N/A	10-013
53106337A	Hombre	33	3	Musulmana	11-010
53106339C	Mujer	33	2	Budista	10-012
53106340D	Mujer	40	1	Católica	10-011
53106341E	Hombre	26	3	Musulmana	11-007
53106342F	Hombre	29	4	Agnóstico	11-008
53106343F	Hombre	37	2	Musulmana	11-007

Tabla 9. Ejemplo anonimizado de un censo de habitantes.

DNI	Género	Edad	Nivel estudios	Religión	Distrito
*	Hombre	21-30	3	*	11
*	Hombre	31-40	3	*	11

*	Mujer	31-40	3	*	10
*	Hombre	31-40	3	*	11
*	Mujer	31-40	2	*	10
*	Mujer	31-40	1	*	10
*	Hombre	21-30	3	*	11
*	Hombre	21-30	4	*	11
*	Hombre	31-40	2	*	11

En este proceso de anonimización se han realizado supresiones de los datos de dos columnas (DNI y religión), y se han realizado dos generalizaciones (en rangos de edades y de sección censal a distrito). Este conjunto de datos es k-anónimo con k=2 para la combinación de atributos Género, Edad y Distrito, puesto que para cualquier combinación de valores disponible en el conjunto de datos anonimizado existen al menos otros k=2 registros que también contienen esos valores.

Es importante determinar que este tipo de anonimización no excluye poder extraer algo de información adicional. Por ejemplo, si se sabe que Luis es un hombre de 23 años que vive en el distrito 11, y el conjunto de datos se sabe que es completo (no es el caso en este ejemplo) entonces se puede saber que su nivel de estudios será 3 o 4. Este tipo de anonimización se ha demostrado útil únicamente cuando el número de dimensiones es pequeño. Si es muy grande, la capacidad de identificación de individuos es grande.

Existen otras propiedades (y técnicas asociadas) más avanzadas, como ***l-diversity*** o ***t-closeness***, que no se describen en este informe por considerarse fuera de su alcance. También otras que se están utilizando en la actualidad como las de **privacidad diferencial**, en las que se describen características de grupos de datos dentro de los conjuntos de datos, sin dar información sobre los individuos. En los algoritmos de anonimización en los que se asegura este tipo de anonimización se puede asegurar que para cualquier búsqueda que se realice en el conjunto de datos agregado no se puede determinar si un individuo fue utilizado o no para computar ese conjunto de datos agregado.

5.3 HERRAMIENTAS DE ANONIMIZACIÓN

Existen múltiples herramientas de anonimización que pueden ser utilizadas para el procesamiento de datos por parte de un experto en anonimización. A continuación se hace referencia a una de las herramientas que se están proponiendo en el contexto de la nube de datos abiertos de investigación europea (European Open Science Cloud - EOSC) para este propósito, Amnesia[49], y que está ganando mucha aceptación entre la comunidad científica, por lo que podría ser aplicable en el contexto de los datos abiertos gubernamentales.

Amnesia es una herramienta de escritorio que se puede utilizar para la anonimización de datos, ofreciendo diversas versiones de anonimización. Permite personalizar la solución de anonimización de acuerdo con los posibles usos posteriores que se prevean de los datos anonimizados, ofreciendo herramientas gráficas para el análisis y la anonimización de datos. De esta manera, las decisiones sobre la anonimización realizada pueden hacerse públicamente disponibles y auditarse mejor.

En relación a las técnicas de generalización, Amnesia se basa fundamentalmente en el uso de jerarquías de generalización. Para ello, se debe facilitar información suficiente para que el algoritmo correspondiente pueda reducir los datos, estableciendo una generalización que impide una identificación concreta. Por ejemplo, se puede

[49] https://amnesia.openaire.eu/

hacer con secciones censales y distritos (donde las secciones censales podrían ser sustituidas por los distritos a los que pertenecen en caso de que fuera necesario), códigos postales y distritos, donde los códigos 28033, 28043, 28027, 28017 y 28032 serían transformados en caso de ser necesario al distrito Ciudad Lineal, etc.

6. CLÁUSULAS TIPO Y CHECKLISTS PARA ASEGURAR LA CALIDAD EN LOS DATOS GENERADOS A PARTIR DE CONTRATOS PÚBLICOS

En esta sección recopilamos un conjunto de cláusulas tipo que se recomienda añadir (con las adaptaciones que puedan ser consideradas necesarias en cada caso) a los conjuntos de cláusulas de los pliegos de prescripciones técnicas asociados a aquellos procesos de contratación pública en los que se generen datos que puedan ser susceptibles de ser publicados en los portales de datos abiertos de las ciudades. Este tipo de cláusulas también podrían ser generalizables a otras administraciones públicas, además de los ayuntamientos.

Estas cláusulas se pueden unir a las ya identificadas en guías como la de la Federación Española de Municipios y Provincias, donde la cláusula de datos abiertos aparece descrita como sigue (basada en la utilizada por Zaragoza):

> *"La empresa adjudicataria deberá proporcionar, a lo largo de toda la vigencia del contrato, la información relativa a la prestación de servicios públicos o al ejercicio de potestades administrativas delegadas que el Ayuntamiento considere que ha de ser publicada de acuerdo con lo dispuesto en la Ordenanza Municipal sobre Transparencia y Libre Acceso a la Información. La información que se facilite deberá cumplir con los criterios de calidad establecidos en la normativa municipal".*

En esta cláusula no sólo se describe la importancia de publicar los datos con respecto a lo que dicta la ordenanza que regula la apertura de datos, sino que también se habla de criterios de calidad.

Asimismo, con mayor nivel de detalle, se pueden encontrar información adicional, también sobre Zaragoza, en: https://www.w3.org/community/opendataspain/wiki/images/a/a6/Zaragoza_clausulaOpenData_concesi%C3%B3ndeservicios.pdf

> ***Contratos que no incluyan la concesión de servicios públicos de titularidad municipal o la delegación de potestades administrativas***
>
> *1. En virtud de lo dispuesto en la Ordenanza sobre transparencia y libre acceso a la información, el Ayuntamiento de Zaragoza podrá publicar o poner a disposición de quien la solicite toda la información relativa a la presente licitación, con la única excepción de la información técnica aportada por las empresas licitadoras que quede cubierta por el secreto comercial.*
>
> *2. El Ayuntamiento de Zaragoza podrá solicitar a la empresa adjudicataria cualquier información relativa al objeto del contrato y la las circunstancias de su ejecución cuando esta sea de interés para los ciudadanos, debiendo la empresa facilitarla en un formato apropiado y en el plazo máximo de una semana, salvo que por su volumen o complejidad se justificara su ampliación. Si la empresa considera que es de aplicación alguna de las limitaciones a la publicidad previstas en la Ordenanza sobre transparencia y libre acceso a la información podrá alegarlo ante el órgano competente en materia de publicidad de la información, que resolverá en plazo de tres días hábiles. Esta obligación subsistirá durante los dos años posteriores a la finalización de las obligaciones principales del contrato.*
>
> ***[y para contratos que incluyan la concesión de servicios públicos de titularidad municipal o la delegación de potestades administrativas, una cláusula adicional:]***

3. La empresa adjudicataria deberá proporcionar, a lo largo de todo el periodo de ejecución del contrato, la información relativa a la prestación de servicios públicos o al ejercicio de potestades administrativas delegadas que el Ayuntamiento considere que ha de ser publicada de acuerdo con lo dispuesto en la Ordenanza sobre transparencia y libre acceso a la información. La información que se facilite deberá cumplir con los criterios de calidad establecidos en la normativa municipal. Según se disponga en el pliego de prescripciones técnicas, la publicación en Internet podrá ser realizada por la empresa adjudicataria, por la web municipal o por ambas, estableciéndose en dicho pliego las especificaciones técnicas para la publicación y, en particular, el formato o formatos a utilizar.

En esta sección se incluyen algunas propuestas de textos/cláusulas que se considera que pueden incluirse en todos aquellos pliegos de prescripciones técnicas de contratos públicos en los que los adjudicatarios deban proporcionar datos a la administración pública que realiza la licitación, que posteriormente puedan ser publicados como datos abiertos. Estos textos/cláusulas están centrados en los aspectos relacionados con la calidad de los datos que se proporcionen y podrán ser adaptados según se considere necesario en cada caso (identificados como obligatorios, evaluados como parte de la evaluación de las ofertas recibidas, etc.).

En relación con los criterios de calidad aplicables a los datos que sean generados y aportados para su posible publicación en el portal de datos abiertos de la institución, deben cumplirse los siguientes criterios:

- El adjudicatario debe seguir el proceso de aseguramiento de la calidad descrito en <<URL>> (contenido de la sección 2 de este documento). Específicamente, debe proporcionar en su oferta una descripción inicial de la ficha o fichas metodológicas para los conjuntos de datos que se vayan a generar.
- Con especial atención, el adjudicatario debe describir los criterios de calidad identificados en la ficha metodológica (completitud, unicidad, consistencia, pertinencia en el tiempo, validez y exactitud), atendiendo a la manera en la que se realizará la adquisición y tratamiento de los datos durante su ciclo de vida, así como describir los procesos de aseguramiento de calidad a seguir.
- En el caso de que sea aplicable, el adjudicatario debe describir los procesos de anonimización de datos que se realizarán para asegurar un tratamiento adecuado de los datos sensibles antes de su publicación como datos abiertos.
- Para cada una de las distribuciones (formatos) en los que se generen los datos, se debe rellenar la siguiente lista de comprobación, indicando si se cumplirá o no con los criterios identificados y aplicables en cada caso: <<ver sección 6.1>>.
- El adjudicatario indicará asimismo qué herramientas se utilizarán para la validación de los criterios de calidad anteriormente identificados, en caso de que sea aplicable su utilización.

6.1 CHECKLIST PARA LA COMPROBACIÓN DE CRITERIOS DE CALIDAD APLICABLES A DISTRIBUCIONES DE CONJUNTOS DE DATOS

Comprobación	¿Aplicable? (S/N)	Método de validación
CSV/XLS(X)-1. Debe utilizarse el conjunto de caracteres de codificación UTF-8		
CSV/XLS(X)-2.- El fichero debe contener una única tabla de datos		
XLS(X)-3 . No incluir hojas vacías		

CSV/XLS(X)-4. La primera fila (y únicamente la primera fila) debe contener el encabezado		
CSV/XLS(X)-5. Los nombres de las columnas, en el encabezado, deben tener un formato de texto universal y ser entendibles por humanos		
CSV-6. Uso de ";" como campo separador de caracteres		
CSV/XLS(X)-7. Mismo número de columnas en todas las filas		
CSV/XLS(X)-8. Sólo un tipo de dato por columna		
CSV/XLS(X)-9. Asignación de un ID único		
JSON-1. Debe utilizarse el conjunto de caracteres de codificación UTF-8.		
JSON-2. Es recomendable incluir un contexto JSON-LD para describir los nombres de propiedades		
JSON-3. Los nombres de las propiedades deben tener un formato universal y ser entendibles por humanos		
JSON-4. Asignación de una ID única		
JSON-5. Es recomendable que no haya mucha anidación en los datos		
ESTAD-1. No se deben utilizar encabezados anidados		
ESTAD-2. No se recomienda representar tablas bidimensionales (o de más dimensiones) en el formato CSV		
ESTAD-3. Las estructuras de datos deben ser siempre estables en el tiempo, no creciendo en número de columnas		
ESTAD-4. No se deben incluir filas o columnas de totales o subtotales, a menos que sea absolutamente necesario, manteniendo el máximo nivel de desagregación de datos posible		
ESTAD-5. Siempre que sea posible, utilizar valores en los campos que estén estandarizados		

GEO-1. Se recomienda incluir coordenadas en latitud y longitud, además de coordenadas en otros formatos habitualmente utilizados por la administración		
GEO-2. En el caso de representar polígonos u otras formas geométricas, se recomienda usar el formato WKT		
GEO-3. Representar las direcciones postales con un doble objetivo de reutilización interna y externa		
BIG-1. Dividir el conjunto de datos en distintas distribuciones de manera que cada una de ellas sea suficientemente tratable con programas informáticos habituales		
FECHA-1. Los valores de datos de tipo fecha y fecha/hora deben describirse en formato ISO 8601		
FECHA-2. Los periodos temporales deben incluirse como dos columnas/propiedades		
FECHA-3. Si no se tiene información completa sobre una fecha, no debe representarse como tal, sino en diversas columnas		
NUM-1. Representar los decimales con un punto.		
NUM-2. No se deben utilizar caracteres de formato de "miles".		
NUM-3: No se deben incluir ceros a la izquierda.		
UNID-1. Las unidades de medida y monedas deben indicarse por separado, o en el nombre de las columnas.		
TXT-1. Uso de cadenas de caracteres con comillas dobles.		