Directrices para el tratamiento de conjuntos de datos en los archivos

Pilar Campos Martínez

materiales

acal
asociación de archiveros de castilla y león

Directrices para el tratamiento de conjuntos de datos en los archivos

Pilar Campos Martínez

Traducción del catalán a cargo de Francisco Fernández Cuesta

acal
asociación de archiveros de castilla y león

COLECCIÓN materiales 1

Título original:
Directrius per al tractament de conjunts de dades als arxius

Originalmente publicado por Associació de professionals de l'arxivística i la gestió de documents de Catalunya (AAC), Barcelona.
Publicado en español con el acuerdo de AAC

Primera edición en español 2023
© Asociación de Archiveros de Castilla y León (ACAL)
Armuña, 14. 37439 San Cristóbal de la Cuesta. Salamanca
Teléfono 650 336 756
www.acal.es

Primera edición en catalán 2022
© Associació de professionals de l'arxivística i la gestió de documents de Catalunya (AAC)
Rocafort, 242 bis, 1r. 08029 Barcelona
Teléfono 934 198 955
www.arxivers.com

© Pilar Campos Martínez, 2023
© por la traducción Francisco Fernández Cuesta, 2023

Diseño y maqueta:
Helvética edición y diseño

ISBN: 978-84-126136-1-2
Depósito Legal: S 61-2023

Índice

5 Introducción
7 Prólogo a la edición en castellano
9 Presentación
10 Conceptos previos
13 Conjuntos de datos en los archivos
13 Introducción
13 Origen y usos
15 Formatos
17 Clasificación y descripción
18 Interoperabilidad
21 Calidad de los conjuntos de datos
21 Introducción
21 Comprobaciones de la calidad
24 Validez
28 Exactitud
30 Integridad
36 Consistencia
39 Uniformidad
42 Calidad del contenido de los datos
45 Conclusiones
46 Bibliografía

Introducción

Este trabajo que tengo el honor de presentar es fruto de la cooperación entre dos asociaciones profesionales que, más allá de su territorialidad, comparten intereses y valores. Una visión de la archivística como herramienta al servicio de la sociedad y, por tanto, un archivística en continuo aprendizaje y evolución para adaptarse a las necesidades y expectativas de aquella.

De momento, esta cooperación se ha materializado exclusivamente en la traducción, edición y difusión de un texto breve, pero de sumo interés y actualidad profesional como estas *Directrices para el tratamiento de conjuntos de datos en los archivos* elaboradas por Pilar Campos y publicadas en su versión original en catalán dentro de la colección Textos que edita la Associació de Professionals de l'Arxivística i la Gestió de Documents de Catalunya (AAC). No será la única iniciativa conjunta, ni mucho menos: en el momento de escribir estas líneas, las juntas directivas de ambas asociaciones están trabajando en la redacción de un convenio de colaboración que siente las bases para nuevas vías de cooperación. Gracias a este convenio podremos asumir de forma conjunta los retos que supone la defensa de los intereses profesionales, permitiendo además que las personas asociadas se beneficien de los servicios de ambas asociaciones.

Por muchas más iniciativas como ésta.

Paloma Vaquero Lorenzo
Presidenta de la Asociación de Archiveros
de Castilla y León (Acal)

mm

Prólogo a la edición en castellano

Los datos, entendidos como unidades mínimas de información que se pueden comunicar, interpretar y procesar, son el petróleo del siglo XXI. Más allá del manido mantra de los gurús tecnológicos y del *management* –y, con ellos, de políticos y directivos de las principales empresas–, lo cierto es que las posibilidades de explotación de información de las que disponen las organizaciones digitales a partir del tratamiento de grandes volúmenes de datos y los cambios que las tecnologías están provocando en la forma en que se documentan las acciones, transacciones y eventos de las organizaciones, nos llevan a pensar que los archivos del siglo XXI son (o serán) también archivos de datos.

Ante este panorama, Pilar Campos Martínez, Técnico superior de archivos en el Institut Municipal d'Hisenda del Ayuntamiento de Barcelona, ha elaborado unas *Directrices* de las que valernos de cara a cualquier transferencia de conjuntos de datos al archivo. No se trata de un discurso teórico sobre la gobernanza del dato, sino de un documento de carácter eminentemente práctico, que apunta los requisitos para la captura de conjuntos de datos estructurados en sistemas de archivo, nos indica en qué cuestiones hemos de fijarnos y cómo podemos resolver los previsibles problemas.

Agradezco a Acal y a AAC que me invitaran a traducir el texto y a su autora, Pilar Campos, por su ayuda en la tarea y, sobre todo, por haber redactado un documento de tanta utilidad de una forma tan sencilla como brillante.

Francisco Fernández Cuesta
Traductor y responsable del Área de Estudios de Acal

mm

Presentación

La generación de datos en las instituciones en la gestión diaria y el análisis y la explotación de los datos para la toma de decisiones hoy en día es una realidad. Además, en el caso de las administraciones públicas y las empresas, es una obligación para la rendición de cuentas con la ciudadanía, con los clientes o con los inversores. Este incremento en la creación y recogida de datos hace que los archivos tengan que prever y planificar cómo hay que hacer la ingesta de estos conjuntos de datos. La dependencia que a menudo existe de sistemas de información, así como la diversidad de formatos y productores, dificulta su captura y el control de la calidad.

Documentar los conjuntos de datos o *datasets* es una preocupación cada vez mayor en la comunidad científica e industrial que debe explotar los datos a los que tiene acceso. Hay varias iniciativas para la creación de un estándar sobre qué información debería acompañar a los conjuntos de datos (Gebru *et al*., 2020). Y es que estos metadatos podrían ser difíciles de registrar en los archivos si en el momento de la ingesta no se han tenido en cuenta unos criterios determinados.

El objetivo de esta publicación es establecer algunos conceptos básicos sobre la gestión de conjuntos de datos en los archivos, analizar los posibles errores en la ingesta de datos y proponer una herramienta para la comprobación de la calidad de los datos previa a la transferencia al archivo.

Conceptos previos

En este texto se utilizan los siguientes términos con el significado que se indica. En algunos casos, con el fin de evitar errores, se han separado términos que son prácticamente sinónimos de *dato* pero que tienen un matiz diferenciado.

base de datos (ingl. *database*). 1. Colección de datos con una estructura para capturar, almacenar, difundir o recuperar datos por múltiples usuarios. 2. Colección de datos interrelacionados organizados según un esquema de bases de datos para dar servicio a una o más aplicaciones. (Butterfield *et al.*, 2016).

datos (ingl. *data*): Representación interpretable de información de manera adecuada para la comunicación, la interpretación o el procesamiento. (ISO/IEC, 1993).

dato (ingl. *data item*): Ítem que se considera indivisible en un contexto concreto. Es una unidad de datos para la que la definición, la identificación, la representación y los valores permitidos se especifican mediante un conjunto de atributos (metadatos). (ISO/IEC, 1994).

***dataset* o conjunto de datos**: Conjunto estructurado de datos creados con un propósito específico. Los conjuntos de datos se pueden almacenar, administrar y publicar en una variedad de formatos y tecnologías. (The National Archives, 2011).

La diferencia con la base de datos es que la primera incluye lógicas internas de ejecución, explotación, búsqueda e interrelación de datos que no hay en el *dataset*; a menudo, los *datasets* son tablas exportadas de la base de datos.

diccionario de datos (ingl. *data dictionary*): Repositorio centralizado de información sobre datos como, por ejemplo, el significado, las relaciones con otros datos, el origen, el uso y el formato. Proporciona soporte a la gestión, a los administradores de bases de datos, a los analistas de sistemas, a los programadores de aplicación y a la planificación, el control, la evaluación, el almacenamiento y el uso de los datos. (IBM, 2001).

datos enlazados (ingl. *linked data*): Solución tecnológica que ofrece un conjunto de buenas prácticas para la publicación de datos y su interconexión, haciendo posible la interoperabilidad y la reutilización de la información disponible en la web. (Bizer *et al*., 2009) (Quílez, 2017).

mm

Conjuntos de datos en los archivos

Introducción

En los archivos se ha trabajado históricamente para intentar normalizar y codificar la documentación que había en lenguaje natural y hacer agrupaciones con descriptores, clasificaciones y puntos de acceso que facilitasen la recuperación de la información. Así, las guías, los inventarios y los catálogos se han convertido en las herramientas con las que se ha trabajado con el objetivo de organizar los fondos y darles acceso. Esta estructura pasa del mundo analógico al mundo digital mediante la creación de bases de datos y, en términos de explotación de datos, la estructuración de las tablas y de las consultas para extraer estos *datasets* de los fondos. Una vez están en formato de datos estructurados, se puede valorar la posibilidad de ofrecer las descripciones y las herramientas archivísticas como un nuevo producto de archivo o posicionarlos en portales de datos abiertos, como ya se hace con las tablas de los cuadros de clasificación.

Del mismo modo que los archivos evolucionan, también lo hacen las oficinas y los procesos de negocio, que cada vez trabajan más con aplicaciones y con datos que, cuando ha transcurrido cierto tiempo, puede que sea necesario pasar al archivo.

Origen y usos

Los orígenes de los conjuntos de datos pueden ser múltiples, tantos como lo son los orígenes de los documentos. En los sistemas de gestión de documentos

la procedencia habitual es la actividad de la propia organización que genera los datos como fruto de sus funciones. Por lo tanto, la captura de estos conjuntos por parte del archivo debería hacerse de forma natural, bien como transferencia para la conservación permanente o relacionada con la continuidad del negocio (The National Archives, 2011), aunque esta situación no resulta todavía habitual.

Actualmente, a los archivos electrónicos llegan, además de documentos textuales (habitualmente en formato PDF o editables, procedentes de procesadores de textos), ficheros estructurados en formato XML o JSON. Estos ficheros pueden ser el resultado de formalizar un registro de la base de datos o bien un documento en un formato concreto o un expediente. Por ejemplo, el resultado de una consulta a través del portal de transparencia o la evidencia de la notificación.

Una situación cada vez más habitual es la transferencia periódica de extracciones de bases de datos (por ejemplo, los registros de entrada anuales); o, en el contexto de la migración de una aplicación o de actualizaciones, la transferencia al archivo de los conjuntos de datos obsoletos que no se considera relevante tener en el nuevo entorno por el hecho de ser de tramitaciones finalizadas.

Otros orígenes pueden ser donaciones de fondos personales o privados, que cada vez serán más digitales y que pueden tener asociadas bases de datos propias o de los programas que utilizaban, como, por ejemplo, el fondo histórico de publicaciones en una red social o la contabilidad de una empresa sobre un determinado programa.

El uso de los conjuntos de datos es diverso y, al ser más fácilmente explotable que los fondos documentales analógicos o en lenguaje natural, no se circunscribe exclusivamente a la investigación histórica.

Por ejemplo, los profesionales del análisis de datos los tratan directamente con el fin de identificar patrones, extrapolaciones y apoyar la toma de decisiones. Además, por lo que se refiere a la ciudadanía, es una herramienta directa de rendición de cuentas por las administraciones.

El archivo debe poder facilitar estos datos y, además, asegurar su integridad y autenticidad, y debe poder explicar su significado y contexto de creación con el fin de evitar errores en la interpretación y, por tanto, en las conclusiones que muestren.

Formatos

Según el origen de los datos y las necesidades del creador de los *datasets*, los datos se pueden encontrar en diversos formatos. Establecer la política de formatos, decidir si se hace la migración y analizar los riesgos de esta migración es responsabilidad del archivo.

Uno de los formatos que se utiliza más ampliamente para representar datos es la hoja de cálculo organizada en filas y columnas (con extensiones como *xls*, *xlsx*, *xltx*, *ods*, *ots* o *csv*). Las columnas pueden estar separadas por tabuladores, comas u otros caracteres y las filas habitualmente están separadas por caracteres de línea nueva.

En cuanto a los modelos de datos abiertos, normalmente se utilizan documentos en formatos abiertos y etiquetados. Son habituales los formatos que se indican en la tabla de la página siguiente.

El Gobierno español, con el desarrollo de la Norma Técnica de Interoperabilidad de Reutilización de recursos de la información (NTI), establece unas pautas para los formatos de publicación de los conjuntos de datos de las Administraciones públicas. Esta norma indica que los datos deben facilitarse en diversos formatos, especialmente en formatos abiertos y estándares de uso generalizado por la

FORMATO	DESCRIPCIÓN
XML	Formato abierto que permite representar los datos de manera estructurada y jerárquica mediante etiquetas.
JSON	Formato de intercambio de datos entre aplicaciones informáticas basado en la notación de objetos usada por el lenguaje de programación JavaScript.
RDF-XML	Formato que permite estructurar los datos en tripletas, en la forma sujeto-atributo-valor, que hace posible incorporar información semántica a los datos.
RSS	Formato del lenguaje XML que permite la distribución de contenidos de páginas web.
KML/KMZ	Formato que es una notación específica de XML para la representación de datos geográficos. Los ficheros KML o KMZ permiten representar varias geometrías (puntos, polígonos, modelos 3D, etc.) expresadas en latitud, longitud y, opcionalmente, altitud.
SHP	Formato vectorial de almacenamiento digital en el que se guarda la localización de elementos geográficos y los atributos asociados a los mismos.
GeoJSON	Formato abierto para representar diferentes estructuras de datos geográficos juntamente con atributos no geográficos basado en el formato JSON.

ciudadanía, y que sólo se podrán utilizar otros más restringidos en casos concretos que lo justifiquen. También indica que se deben escoger preferentemente formatos que permitan hacer representaciones semánticas de la información (RDF), con el fin de mejorar la comprensión de la información representada y su tratamiento automatizado.

Clasificación y descripción

La clasificación y la descripción de los conjuntos de datos no debe ser diferente a la que se aplica a otros documentos y soportes tradicionales. Los *datasets* se generan en un contexto determinado y tienen una función concreta, sea la rendición de cuentas para la transparencia, el resultado de la comunicación de la institución (como en el caso de la extracción de datos de las redes sociales de una institución), los padrones de la gestión de determinados impuestos, etc.

Sin embargo, a veces no es tan sencillo, porque son agrupaciones de datos que responden a diversas funciones o series documentales como, por ejemplo, extracciones de bases de datos que contienen información de diversos planes de estudios; las licencias de obras menores y mayores de una circunscripción, etc. En estos casos se puede valorar hacer la separación funcional de los datos y no mantener el conjunto, o considerarlos documentos recapitulativos, como es el caso de los libros de registro de entrada y salida de un organismo (que contienen la información de inicio o respuesta de muchos de los trámites administrativos). En este aspecto, estándares como el modelo conceptual Records in Context (RiC-CM), del Consejo Internacional de Archivos (ICA), se puede adaptar bien a las necesidades de flexibilidad de la descripción de *datasets*, así como de otros conjuntos documentales (Pastor Sánchez, Llanes Padrón, 2017).

La forma que puede adoptar la descripción de un *dataset* se realiza mediante los metadatos asociados al fichero según un determinado esquema, con descripciones con estructuras multinivel, con jerarquías más flexibles, o bien con la vinculación a la función y a los registros de autoridad. No debería ser diferente de las descripciones que se hacen de documentos digitales o electrónicos. Revisando diferentes archivos nacionales, se detecta que no utilizan sistemas de descripción diferentes para los datos que para el resto de los documentos de sus fondos (Niu, 2016).

La propia norma de descripción ISAD(G) (International Council on Archives, 2001) permite la descripción de *datasets*, y es la utilizada por la Sección de Datasets Gubernamentales (National Digital Archive of Datasets, NDAD) de los Archivos Nacionales del Reino Unidad (The National Archives) ya desde el año 2000. La propuesta de funcionamiento que propone el NDAD es la ampliación del cuadro de clasificación en dos niveles por debajo de serie documental: el nivel de «expediente», que sería la captura concreta del conjunto de datos, y un nivel múltiple denominado «ítem», en el que se describirían las diferentes tablas que componen el conjunto de datos si están desvinculados entre ellas partiendo de un modelo relacional (Shepherd, Smith, 2000).

Interoperabilidad

Hablando de clasificación de los *datasets*, se debe mencionar la Norma Técnica de Interoperabilidad de Reutilización de Recursos de la Información (Ministerio de Hacienda y Administraciones Públicas, 2013) que hemos citado anteriormente. Esta norma está muy vinculada a la posibilidad de interoperar con conjuntos de datos de varias administraciones y

el *linked open data*, y en el apartado cinco establece unas directrices de descripción de la información y fija su clasificación temática. En los casos en que la ingesta proviene de datos administrativos, esta clasificación debería haberla establecido en el diccionario de datos el departamento de origen que hace la transferencia de datos, pero es posible que en datos internos o no publicados no se haya fijado la clasificación temática tal como pide la norma NTI.

Los ámbitos temáticos que marca la NTI son: población; territorio; ciudad y servicios; administración, y economía y empresa, con una serie de subapartados que pueden ser fácilmente relacionables con el cuadro de clasificación de cada institución. Además, la norma indica que cualquier información reutilizable debe llevar asociados unos metadatos mínimos para describir los conjuntos de datos para poder hacer la reutilización, y que habría que usar los vocabularios y los esquemas de valores que se proponen mediante tecnologías de la web semántica como las ontologías.

Además, en cuanto a la propia descripción de los *datasets*, establece unos esquemas de metadatos mínimos para describirlos que pueden servir de base para ampliar el esquema necesario para establecer todo el contexto.

mm

Calidad de los conjuntos de datos

Introducción

En cuanto a la calidad de los datos, hay que decir que hay una extensa bibliografía, porque es una cuestión que afecta directamente a las aplicaciones y a los negocios que las consulten y utilizan. El concepto de calidad es amplio y tiene diferentes dimensiones, y cuando llega al archivo hay que tener en cuenta una serie de aspectos, mientras que otros dependen directamente de quién ha creado estos datos y no se pueden corregir. En el momento de la creación o el diseño de las bases de datos o en el momento de extraer los datos se pueden tener en cuenta las recomendaciones del Consorcio World Wide Web (Lóscio *et al*., 2016) y el vocabulario de calidad de los datos (DQV) desarrollado para tener en cuenta las métricas y medidas de cada dimensión (Hyvonen *et al*., 2020). Esta información debería estar introducida como metadatos y la definición debería formar parte de la política de la organización que genera estos conjuntos de datos, pero desafortunadamente no es habitual.

Comprobaciones de la calidad

Se considera que el 80 % del análisis de datos se centra en el proceso de limpieza y preparación de los datos (Oni, 2019) y que es en este estadio donde se detectan los errores y se recaban nuevos datos.

En los archivos, la comprobación de la calidad de los datos sería similar a como se hace con las transferencias de documentos no estructurados: se pide que lleguen los expedientes netos de copias y

borradores. En el caso de las ingestas automáticas al gestor documental, se indican metadatos y se hacen las comprobaciones de tipologías documentales o documentos obligatorios.

El concepto de «dato erróneo» es amplio. Los datos erróneos son datos inexactos, sea por falta de datos o por la inclusión de datos inapropiados (por ejemplo, datos en columnas incorrectas), datos duplicados o datos incorrectos (errores ortográficos, tipográficos, de formato, etc.). Detectar y reducir estos errores disminuye el margen de error de explotaciones posteriores y mejora la comprensión del contexto de creación de los datos.

Los problemas de calidad de los datos están presentes en conjuntos, colecciones y bases de datos únicas. La integración de múltiples fuentes de datos es compleja, dado que las fuentes a menudo contienen datos redundantes o erróneos, sobre todo si hay grandes volúmenes de datos involucrados.

Con el fin de valorar la calidad de los conjuntos de datos se han desarrollado varios criterios que deben cumplir los conjuntos de datos (Olson, 2003):

- Validez: los datos son correctos si sirven para la finalidad para la que se han recogido. Garantizar la validez es complicado, especialmente en el uso de hojas de cálculo, por la dificultad de establecer controles en los datos que aportan los usuarios.
- Exactitud: grado de conformidad de una medida con un valor estándar.
- Integridad: grado en que se conocen todas las medidas que se requieren.
- Consistencia: grado en que un conjunto de medidas es equivalente en todos los sistemas. La inconsistencia tiene lugar cuando dos elementos de datos en el conjunto de datos se contradicen entre ellas.[1]

- Uniformidad: grado en el que se especifican las medidas de un conjunto de datos utilizando las mismas unidades de medida en todos los sistemas.

En el apartado siguiente se propone una lista de posibles errores de forma y contenido que se pueden comprobar rápidamente en el momento de la ingesta, sea manualmente con extracciones de los mismos sistemas o por medio de agentes terceros en un contexto de donación o recopilación de archivos sociales.

La lista recoge, en forma de tablas agrupadas según los criterios de calidad de los conjuntos de datos, la definición, la forma de detección, la forma de corrección y ejemplos o casos en los que se ha encontrado o se puede encontrar este error. En los casos en que es un error que afecta a varias propiedades, se ha incorporado sólo en una.

Para la elaboración de la lista se han recogido ejemplos de buenas prácticas y errores comunes de trabajo en *datasets* y revisión de bibliografía, especialmente basada en el artículo «The Quartz guide to bad data» (Graskopf, 2015), en el que hay una primera lista, y se ha organizado siguiendo los criterios anteriores.

[1] La herramienta OpenRefine's para agrupar textos puede ayudar en el proceso de comprobación de la consistencia y en la corrección, sugiriendo coincidencias entre valores inconsistentes en una columna (por ejemplo, relacionando Los Ángelos con Los Ángeles). Es necesario documentar los cambios de manera que se pueda garantizar el origen de los datos.

Validez

PROBLEMA	DEFINICIÓN	EJEMPLOS	ACCIÓN
El origen de los datos no está documentado	No se sabe el contexto de creación del *dataset*	Aunque el archivo pueda ser el receptor final de todo tipo de *datasets*, el archivo debe poder: - Identificar al productor: no es lo mismo si es la misma institución, un grupo seguidor de teorías de la conspiración, un grupo político, un lobby con intereses comerciales o una ONG. - Saber la forma de recogida: la fiabilidad también la marca el hecho de si el *dataset* se ha recogido mediante sensores, recogida automatizada, encuestas y entrevistas, transcripciones manuales, etc.; un sensor puede no estar bien calibrado, pero es más preciso que una medición manual. - Saber si la captura del *data set* se ha hecho en el mismo momento a mano o con una tableta táctil o un ordenador (*typed, tapped or scribbled*), para detectar los errores tipográficos o de uniformidad. - Conocer el contexto de captura: los datos geográficos tomados en tiempos de guerra pueden no ser precisos, los datos de un gobierno pueden tener sesgos ideológicos o no ser exhaustivos, etc.	Detección: antes de la transferencia, recogida de los datos de contexto. Solución: tener una política bien definida de transferencias en la que se obligue a informar de todos los datos de producción y captura de los *datasets*.

Las categorías han sido mal seleccionadas	No se han incluido todas las categorías posibles o bien las categorías están mal definidas	Campos verdadero/falso que no lo son realmente. Puede pasar en encuestas en las que se rechaza responder algunas preguntas o en las que no se sabe la respuesta de algunas preguntas, hechos que también son datos significativos. Ejemplo: el uso de la categoría «Otros» en listas cerradas; por ejemplo, en una lista de países, «otros» puede significar que es un país desconocido, que no es un país reconocido, que se trata de aguas internacionales... o, si se trata de personas, puede hacer referencia a personas refugiadas o expatriadas, que no saben la respuesta, o que no se identifican con ningún país de la lista, etc.	Detección: revisar las listas de las categorías. Solución: es un error en la captura de datos, sólo se puede documentar el sentido de las categorías definiéndolas bien para evitar los errores de interpretación.
Las categorías son demasiado amplias	Las categorías agrupan conjuntos demasiado amplios para que sean representativos	Se han agregado datos por provincias y para los conjuntos sería bueno tenerlos separados por comarcas o municipios, o se dan los datos anuales pero no disgregados en meses o días. Es posible que los datos estén así por motivos de confidencialidad (por ejemplo, no identificar al único residente nepalí de un municipio), aunque otras veces puede haberse hecho para reducir la extensión del *dataset*.	Detección: revisar las listas de las categorías. Solución: como una vez se han agregado los datos (lo que puede interesar de hacer por motivos de confidencialidad, pero se pierde el detalle) ya no se pueden volver a separar, hay que pedir a la fuente si tiene los conjuntos desagregados.

Valores anormalmente regulares	Se producen unos conjuntos de valores muy regulares sin motivo alguno	Ejemplo: la cantidad de metros eliminados de una serie durante muchos años es exactamente la misma y se eliminaron en el mismo momento. Es posible que sea una media de varios años, al tener previamente el dato agregado. No se debería dividir el dato en subconjuntos (otro ejemplo sería un dato anual dividido entre 12 meses y designado con el nombre de «media mensual»).	Detección: revisar los valores repetidos anormales. Solución: pedir los datos previos a la desagregación. Si no es posible, documentar los cálculos que se hayan hecho e informar de ello.
Existencia de valores sospechosos	Valores que suelen ser erróneos, sea por entrada manual o por errores del sistema de tratamiento de datos	Ejemplos de números: 65.535 255 2.147.483.647 4.294.967.295 555-3485 99999 (o cualquier otra secuencia larga de nueves) 00000 (o cualquier otra secuencia larga de ceros) Ejemplos de fechas: 1970-01-01T00:00:00Z 1969-12-31T23:59:59Z January 1st, 1900 January 1st, 1904	Detección: revisar los campos para buscar estos números. Solución: comprobar si son correctos y, en caso de que no lo sean, pedir a la fuente la corrección. Si no tiene los números correctos, hay que documentarlo como error.

		Ejemplos de geolocalización: 0°00'00.0 DCPN+0°00 Código postal EEUU 12345 (Schenectady, New York) Código postal 01001 (Abrevadero, Vitoria, España)	
Los datos están mezclados con formatos y anotaciones	Repeticiones de los encabeza-mientos o definiciones y anotaciones en las tablas	Ejemplo: Una hoja de cálculo en la que la primera fila es el título de la tabla, la segunda es el nombre de las columnas y la tercera es la descripción o las anotaciones, y los datos empiecen a partir de la cuarta. También es posible encontrar tablas en las que cada cierto número de filas se repite el encabezamiento, sobre todo si se han utilizado para imprimirse.	Detección: revisar si hay filas extras en los encabezamientos. Solución: pedir preferentemente las tablas en formato XML o HTML que permitan ver una separación clara entre los datos y el formato.

Exactitud

PROBLEMA	DEFINICIÓN	EJEMPLOS	ACCIÓN
No se especifica el tiempo de recogida de los datos.	Los valores e referencia se han modificado a lo largo del tiempo y no se conoce el momento en que se toma el dato.	Ejemplo: debido a la inflación, el sueldo en euros del año 2002 no es equivalente al sueldo en euros del año 2021, aunque es la misma unidad de medida. Otro caso es cuando se indica que el sujeto cobra «el salario mínimo interprofesional», porque es un valor que va cambiando a lo largo de los años. Incluso, si lo dice, hay unidades que han cambiado el significado con el tiempo. Un dólar del año 2010 no tiene el valor de un dólar actual, o una tonelada corta no es una tonelada imperial ni una tonelada.[2]	Detección: mirar si los campos tienen medidas fijas (científicas) o variables a lo largo del tiempo. Si hay rangos de muchos años a lo largo del *dataset*, es más habitual que exista este problema. Solución: indicar el año de toma de los datos e informar de que hay campos que no son comparables directamente.
Diferencias de cuántos decimales admite.	En filas o columnas diferentes se admiten más	En algunos campos monetarios se han admitido hasta tres decimales y otros solo admiten dos o ninguno. A la hora de hacer cálculos (como el IVA repercutido), puede variar el importe final y dar errores que no son reales.	Detección: mirar las configuraciones del campo y hacer la comparación entre filas.

[2] Tonelada inglesa, imperial o larga: 1.016,0469088 kilogramos. Tonelada americana o tonelada corta: 907,185 kilogramos. Tonelada o tonelada métrica: 1.000 kilogramos.

	decimales o menos, algunos están redondeados y otros son más exactos.		Solución: pedir a la fuente el *dataset* limpio y más exacto o pasar todos los datos a un único formato.
Los nombres de los campos son ambiguos.	Los nombres de los campos no especifican el significado del dato y se pueden interpretar de maneras diferentes.	Un campo llamado «Residencia» puede significar el lugar donde se vive habitualmente, el lugar donde se pagan los impuestos, la ciudad, la comarca o el país.	Detección: hacer una lectura detallada de los nombres de los campos. Solución: pedir a la fuente la definición del diccionario de datos.

Integridad

PROBLEMA	DEFINICIÓN	EJEMPLOS	ACCIÓN
Valor en blanco o nulo.	Datos no introducidos.	Se dan datos anuales y hay un año está en blanco: hay que saber si es porque no se recogieron datos ese año o por algún otro motivo. En el caso de las encuestas, saber si el valor blanco es porque el entrevistado no sabe la respuesta o ha rehusado contestarla, o si es posible que no encuentren la respuesta a la lista controlada (por ejemplo, por la cuestión del género hombre/mujer en personas que se definen como no binarias).	Detección: buscar los valores en blanco y ver si tienen una explicación. Solución: preguntar a la fuente el significado y describirlo en el diccionario de datos.
Los datos en blanco se reemplazan por ceros.	Uso de valores arbitrarios para sustituir los valores en blanco o nulos.	Un valor falso para expresar la fecha se acostumbra a encontrar como «1970-01-01-01T00:00:00Z» o «1969-12-31T24:59:59ZZ», que es donde empieza el registro temporal Unix o Unix epoch for timestamps. Un falso 0 para la localización se puede representar como 0'00 Geli 00 Geli N+ 0º 00 Geli 00 Geli E o simplemente como 0ºN 0ºE, que es el punto del océano Atlántico justo al sur de Ghana conocido como Null Island.	Detección: buscar los valores 0, –1, «false» y «null» y comprobar si son coherentes. Buscar si aparece la fecha del registro temporal Unix o la geolocalización de Null Island. Solución: preguntar a la fuente el significado.

Faltan datos que deberían estar.	Falta de datos que por la temática o el concepto deberían estar.	Datos geográficos: todos los municipios, comunidades autónomas, países o estados deben estar representados, o debe haber alguna razón para que no estén. Si se trabaja con un *dataset* sobre las empresas del IBEX35, se debe comprobar que no falta ninguna.	Detección: tener el conocimiento del conjunto de datos. Solución: si hay carencias, pedir a la fuente el *dataset* completo o la explicación de los huecos para poder documentarlo.
El *dataset* tiene 65.536 o 1.048.576 registros.	El máximo de filas que admitían las versiones antiguas del Excel eran 65.536. Actualmente son 1.048.576.	Si se quiere hacer una transferencia de un *dataset* con este número exacto de registros es posible que se haya cortado en alguna migración a Excel y falten datos.	Detección: revisar el total de filas. Solución: pedir a la fuente el *dataset* completo.
El dataset tiene 255 columnas.	La aplicación Numbers de Apple solo admite 255 columnas y recorta las	Si se quiere hacer una transferencia de un *dataset* con este número exacto de columnas es posible que se hayan cortado en alguna migración o que se hayan abierto en algún momento con la aplicación Numbers de Apple.	Detección: revisar el total de filas Solución: pedir a la fuente el *dataset* completo.

	filas que tienen más sin avisar al usuario.		
El *dataset* tiene unas fechas anormales.	Algunos programas tienen unas fechas por defecto y, si se detectan, es posible que se hayan incorporado automáticamente y que no respondan a una fecha real.	El Excel tiene como referencia para contar las fechas el 1 de enero de 1900, excepto si se abre con un Mac, que es el 1 de enero de 1904. Es posible que errores en cálculos de fechas acaben dejando la fecha por defecto y que sea alguna de ellas. Si aparece a menudo, o si es una fecha que resulta incoherente con el resto, se trataría de un error. Otras bases de datos y aplicaciones generan la fecha 1970-01-01T00:00:00Z o 1969-12-31T23:59:59Z, tomada como fecha muestra cuando un sistema intenta entrar un valor 0 o vacío a un campo de fecha obligatorio.	Detección: revisar los campos de fecha para comprobar si aparecen estos valores. Solución: pedir a la fuente el *dataset* sin errores o revisar los campos calculados para encontrar el error. Si no se dispone, eliminar estas fechas, documentando el motivo del año que está vacío.
Los espacios al final de la línea están	Los archivos con texto y datos, como	Windows, Mac y Linux han tenido un desacuerdo histórico sobre cuáles deberían ser los caracteres que marquen el salto de línea.	Detección: revisar la formación del CSV y del *dataset* para ver si los saltos de línea están bien

mal codificados.	es el caso de los *.csv*, utilizan caracteres invisibles para representar una nueva línea	Abrir un fichero con un programa que previamente se ha desatado con otro sistema operativo puede provocar que el Excel u otras aplicaciones fallen a la hora de identificar el salto de línea. Hasta el año 2018, el editor de texto tenía algunas incompatibilidades en la interpretación. (López, 2018)	interpretados. Solución: parametrizar correctamente los saltos de línea.
Falla la ley Benford[3].	Sistema de detección de posibles carencias de datos	Se puede utilizar este sistema para hacer una comprobación por si hay sospechas de que el *dataset* es incompleto o se ha modificado. En las direcciones es habitual que haya más registros con números bajos de calles que con números altos, ya que no todas las calles tienen la misma longitud, y pasa igual con los pisos. Del mismo modo, también hay más población con estudios primarios que superiores, etc.	Detección: revisar los campos numéricos que son incrementales. Solución: pedir a la fuente si hay alguna razón por la que no se cumple la ley Benford y, si no es así, pedir el *dataset* completo.
El texto se ha convertido en números.	No todos los numerales son números, pero es posible que	Los códigos postales identifican geográficamente y tienen una longitud determinada. Así, 01007 es el código postal de Armentia, Álava (España), no es el 1007, que es Buenos Aires (Argentina).	Detección: revisar las configuraciones de los campos y comprobar si son coherentes con el contenido.

[3] Ley Benford o ley del primer dígito: teoría según la cual los dígitos pequeños (1, 2, 3...) aparecen mucho más a menudo que los dígitos grandes (7, 8, 14...).

	en migraciones la definición de la celda se haya cogido por defecto y los haya transformado.	El Excel y otras hojas de cálculo asumen que los numerales son números y borran los ceros iniciales.	Solución: pedir a la fuente la corrección. Si no es posible, documentar el error.
Los números se han almacenado como texto.	La configura-ción de los campos se ha pasado a texto, algo habitual cuando se ha optimizado para presentar los datos en lugar de para hacerlo reutilizable.	En lugar de escribir un millón de euros con el número 1000000, la celda puede contener comas 1,000,000, espacios 1 000 000 o texto: «euros». El hecho de ser texto lo hace poco reutilizable y no comparable con otros datos.	Detección: revisar las configuraciones de los campos y comprobar si son coherentes con el contenido. Solución: pedir a la fuente la corrección y la limpieza de los datos. Una buena práctica es desechar los números sin formato e incluir después información de soporte en el nombre de la columna o en los metadatos.

La hoja de cálculo ha interpretado mal el formato de los campos.	Los campos se han abierto en una hoja de cálculo y se han desatado con el formato erróneo.	Un campo de fecha que se ha interpretado como una fórmula y divide o resta los números, un campo textual que se convierte en numérico... Hay que repasar el origen y las posibles migraciones. Ejemplo de 2020: Se renombraron por error 27 genes humanos porque se había abierto con Excel el *dataset* que los contenía y los confundió por fechas (Jacobson, 2020).	Detección: aunque es complicada, hacer una comprobación previa abriendo una copia del *dataset* en una hoja de cálculo y ver si interpreta correctamente los campos. Solución: pedir a la fuente una extracción directa de la fuente y utilizar formatos como CSV.

Consistencia

PROBLEMA	DEFINICIÓN	EJEMPLOS	ACCIÓN
Filas duplicadas.	Filas o partes de filas que aparecen duplicadas sin motivo alguno.	En *datasets* financieros es posible que aparezcan dos filas casi idénticas y que utilizan los mismos identificadores que la transacción original, pero una es la original y la otra es la corrección. Hay que tenerlo en cuenta para no eliminarla como duplicado, pero se debe informar para evitar cálculos erróneos.	Detección: buscar valores duplicados en varias filas. Solución: preguntar a la fuente el significado de los valores duplicados y documentarlo. Trabajar con identificadores únicos.
Ortografía inconsistente.	Errores tipográficos.	Nombres con errores de grabación, faltas de ortografía o abreviaturas. Ejemplo: Santa Coloma / Sta. Coloma / Sta Coloma / Santa Culoma / Sta Coloam... Son muy habituales en procesos automáticos de lectura y reconocimiento de caracteres, en procesos manuales de captura de datos (encuestas, entrevistas en la calle), en grabaciones de datos manuales y en transcripciones de textos.	Detección: buscar campos controlados (como los nombres de países, ciudades, etc.). Solución: documentar que la recogida es manual y, por lo tanto, puede contener errores en los campos no controlados. Hacer la corrección de los campos si se pueden confrontar con la fuente y documentarlo.

El orden de las palabras es inconsistente.	En expresiones o nombres, los datos están separados con diferentes criterios.	En *datasets* que tienen nombres y apellidos con orígenes étnicamente diversos, la unificación de «nombre de pila (*first name*)» «segundo nombre (*middle name*)» «apellido 1» y «apellido 2» no siempre es la correcta porque no se ha seguido el mismo criterio de concatenación. Por lo tanto, si hay un *dataset* con nombres y apellidos unidos, hay que comprobar que sea correcto.	Detección: buscar nombres que se conozcan dentro del *dataset*, nombres compuestos y de orígenes diferentes y comprobar si son correctos. Solución: pedir el *dataset* sin los campos concatenados o hacer una revisión a partir de tablas controladas de nombres.
Los formatos de fechas son inconsistentes.	Las fechas no siguen el mismo formato.	Si algunas fechas se escribieron siguiendo el sistema europeo, y otras por el sistema norteamericano, las fechas son inconsistentes. AñoMesDía Horas:Minutos DíaMesAño Horas:Minutos MesDiaAño Horas:Minutos:Segundos	Detección: comprobar la configuración de los campos y mirar si hay fechas imposibles en los diferentes formatos. Solución: preguntarlo a la fuente y hacer la conversión a un formato único.
Los formatos de los decimales	La separación entre los miles y los	Por ejemplo: 1,000 > 1.000, o 1,000 < 1.000? Si la pareja de cifras fue escrita por un americano, la primera es un millar y es mayor que la segunda, que es	Detección: comprobar la configuración de los campos y mirar si hay cifras imposibles.

son inconsistentes.	decimales utiliza de manera indistinta el punto (.) y la coma (,).	uno; en cambio, si fue escrita por un europeo, la primera es uno y es más pequeña que la segunda, que es un millar. Aun así, en general se habla más de la coma decimal que del punto.	Solución: preguntarlo a la fuente y hacer la conversión a un formato único.
El texto se ha descodificado con un código diferente al original.	Un texto tiene un conjunto de caracteres que no se pueden interpretar.	Textos en los que por causa del efecto *mojibake*[4] no es posible interpretar algunos caracteres. Los datos aparecen como si fueran basura o con estos símbolos: ⍰⍰⍰. Puede suceder en casos con alfabetos o fórmulas no habituales. Se recomienda utilizar por defecto la codificación UTF-8.	Detección: comprobar el contenido del *dataset*. Solución: en la mayoría de los casos el editor de textos detecta cuál es la codificación. La fuente debe indicar cuál es la codificación de sus datos.

[4] NdT: Distorsión de un texto como resultado no intencionado de la descodificación de sus caracteres, que provoca su sustitución por una codificación diferente que impide su comprensión (como letras con tildes, símbolos matemáticos o interrogaciones).

Uniformidad

PROBLEMA	DEFINICIÓN	EJEMPLOS	ACCIÓN
No se especifican las unidades.	Los campos numéricos sin valores absolutos no especifican la medida de referencia.	Una columna que se denomine «peso» o «coste» no aporta ninguna información sobre en qué unidad está medido, no se puede asumir si es en kilos, en libras, en euros o dólares. Los datos científicos habitualmente se expresan mediante el sistema métrico decimal y, del mismo modo, algunos precios extranjeros se pueden expresar en la moneda local, o una tonelada corta no es lo mismo que una tonelada imperial.	Detección: mirar los campos que necesitan especificar la unidad de medida. Solución: Preguntar a la fuente el significado.
Los símbolos monetarios son insuficientes.	El nombre y el símbolo de la moneda no es único a nivel mundial y hay que informar del país al que hace referencia.	Por ejemplo, 10$ son 10 dólares, pero hay varios países que utilizan el mismo nombre para su moneda, aunque tienen valores diferentes. En algunos casos, como los dólares australianos, se indica A$10 para diferenciarlos de los dólares de los demás países, pero es posible que si se han hecho simplificaciones pensando que sería un *dataset* de uso local, se hayan sacado.	Detección: mirar los campos monetarios y detectar si hay orígenes diferentes. Solución: preguntarlo a la fuente y completar las unidades con la información de la moneda a la que se hace referencia.

Los valores de las categorías son variables a lo largo del tiempo.	La categorización agrupa diferentes valores a lo largo del tiempo. Esto puede hacer que no sean comparables.	La mala categorización o una categorización que agrupe diferentes conceptos que son convenciones sociales, puede excluir datos artificialmente. Por ejemplo, en las estadísticas de crímenes el significado de «violencia de género» ha cambiado de definición a lo largo del tiempo. Esto puede hacer que a lo largo del tiempo un delito de agresión sexual esté en categorías diferentes o que haya periodos en los que simplemente no se haya contado.	Detección: revisar los campos con categorías que tienden a ser arbitrarias o cambiantes según la legalidad, especialmente con conjuntos de datos sobre etnias o policiales. Solución: preguntar a la fuente e informar de las categorías que agrupan, si es necesario, en cada momento. Documentar la fecha de la captura de datos y la categorización.

mm

Calidad del contenido de los datos

Existen diversas prácticas para revisar si el contenido de los datos responde a una realidad o si los datos pueden ser sesgados o erróneos, pero el objetivo de las directrices que se han proporcionado no es la auditoría de los datos en cuanto al contenido, del mismo modo que tampoco en los archivos textuales lo es establecer la verdad de lo que contienen los documentos. Los archivos deben asegurar la trazabilidad, la integridad del contenido y la autenticidad de los datos, y para ello es necesario recoger el contexto.

El archivo debe recoger, en la transferencia de los conjuntos de datos, el contexto amplio de captura y creación para que quien consulte los documentos lo pueda hacer con el máximo de conocimiento. Igualmente, debe establecer políticas claras respecto a qué y cómo se aceptará como datos, con pautas e información sobre los siguientes aspectos:

- Margen de error de los datos.
- Margen de aleatoriedad de la muestra: idiomas disponibles en las encuestas, lugares de publicación, sistemas de publicitado, etc.
- Ediciones posteriores: correcciones de errores del proceso de captura (por ejemplo, normalizaciones erróneas de apellidos, como «Peres» por «Pérez», que provocan confusiones, cambios masivos por anonimización de datos personales, etc.) y si se han documentado.

- Preguntas de las encuestas para obtener los datos: la manera de hacer la pregunta puede influir en la respuesta, y también quién la responde. El hecho de conservar las plantillas da una parte del contexto de captura.
- Formato original de los datos: si provienen de bases de datos, de documentos escaneados con reconocimiento óptico de letra manuscrita o mecanoscrita, o si provienen de ficheros PDF, etc.

mm

Conclusiones

Este documento ha presentado una serie de conceptos y una propuesta metodológica para la comprobación de la calidad formal de los conjuntos de datos en el momento de la ingesta en el archivo. En un mundo en el que la producción de datos es una realidad diaria hay que estar preparados para establecer en las organizaciones las políticas de aceptación de estos conjuntos de datos y traducir la metodología de archivo, una vez más, a un nuevo formato.

Los archivos se han adaptado a los documentos digitales ofimáticos y han preparado gestores documentales, estrategias de descripción con esquemas de metadatos y políticas de preservación. Sin embargo, hasta que se llegue a tener una visión completa del sistema, habrá que aprender a convivir con un entorno híbrido, con documentos en diferentes formatos y formas.

Las directrices que se proponen son un primer paso para que cada organización las adapte sabiendo cuál es su realidad, pero son, en todo caso, comprobaciones previas a la transferencia de datos. También habrá que desarrollar, en segundo término, cómo se puede hacer el tratamiento de los datos y las posibilidades de acceso que ofrecen los formatos de datos estructurados, adaptándose a las normas de descripción de los documentos digitales.

Bibliografía

BIZER, C.; HEATH, T.; BERNERS-LEE, T. (2009). «Linked data - The story so far». *International Journal on Semantic Web and Information Systems*. Vol. 5, núm. 3. <https://doi.org/10.4018/jswis.2009081901>.

BUTTERFIELD, A.; EKEMBE NGONDI, G.; KERR, A. (ed.) (2016). *A Dictionary of Computer Science*. 7ª ed. Oxford University Press. <https://doi.org/10.1093/acref/9780199688975.001.0001>.

DEPARTAMENTO DE CULTURA. *Cuadro de clasificación de la documentación. Datos abiertos de Cataluña*. Portal de datos abiertos de la Generalidad de Cataluña. <https://analisi.transparenciacatalunya.cat/Sector-Public/Quadre-de-Classificacy-de-la-documentaci-/5t23-dy8y>. [Consulta: 28, julio, 2021].

DEPARTAMENTO DE CULTURA. *Qué son los datos abiertos. Gobierno abierto*. Generalidad de Cataluña. <https://governobert.gencat.cat/ca/dades_obertes/dades-obertes/que-son-les-dades-obertes/>. [Consulta: 28, julio, 2021].

GEBRU, T.; MORGENSTERN, J.; VECCHIONE, B.; VAUGHAN, J. W.; WALLACH, H.; DAUMÉ III, H.; CRAWFORD, K. (2020). «Datasheets for Datasets». *Archiv*. <https://arxiv.org/abs/1803.09010v7>.

GRASKOPF, C. (2015). «The Quartz guide to bad data». *Quartz* (15 de diciembre de 2015). <https://qz.com/572338/the-quartz-guide-to-bad-data/>.

HYVONEN, E.; ALBERTONI, R.; ISAAC, A. (2020). «Introducing the Data Quality Vocabulary (DQV)». *Semantic Web*. Vol. 12, núm. 1. <https://doi.org/10.3233/SW-200382>.

IBM. (2001). *Dictionary of computer science, engineering, and technology*. Vol. 38, núm. 10. <https://doi.org/10.5860/choice.38-5325>.

INTERNATIONAL COUNCIL ON ARCHIVES (2001). ISAD(G): Norma Internacional General de Descripción Archivística = General International Standard Archival Description; adoptada por el Comité de Normas de Descripción, Estocolmo, Suecia, 19-22 de septiembre de 1999. Versión catalana al cuidado de Josep Matas y Jaume Rufí Pagès; con el asesoramiento de Ángeles Bernal *et al*. 2ª ed. Vol. 2ª, p. 112. Barcelona: de

Archiveros de Cataluña; Departamento de Cultura de la Generalidad de Cataluña. <http://cataleg.ub.edu/record=b1492091TMVS1*cat>.

ISO/IEC (1993). *ISO/IEC 2382-1:1993 – Information Technology – Vocabulary – Parte 1: Fundamental Terms*. [Consulta: 24, febrero, 2021].

ISO/IEC (1994). *ISO - ISO/IEC 11179-3:1994 – Information technology – Specification and standardization of data elements. – Part 3: Basic attributes of data elements*. <https://www.iso.org/standard/19184.html>. [Consulta: 24, febrero, 2021].

JACOBSON, D. (2020). «Human genes renamed as Microsoft Excel reads them as dates – BioNews». *BioNews* (10 de agosto de 2020). <https://www.bionews.org.uk/page_151346>.

LÓPEZ, M. (2018). «Introducing extended line endings support in Notepad». *Windows Command Line* (8 de mayo de 2018). Microsoft. <https://devblogs.microsoft.com/commandline/extended-eol-in-notepad/>. [Fecha de consulta: 26, febrero, 2021].

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N.; GREINER, A.; ISAAC, A.; IGLESIAS, C.; LAUFER, C. (2016). «Data on the Web Best Practices». *W3C* (enero de 2016). <https://www.w3.org/TR/dwbp/>.

MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS (2013). Resolución de 19 de febrero de 2013, de la Secretaría de Estado de Administraciones Públicas, por la que se aprueba la Norma Técnica de Interoperabilidad de Reutilización de recursos de la información. *Boletín Oficial del Estado*. Núm. 54 (4 de marzo de 2013), p. 17045 a 17071.

NIU, J. (2016). «Organisation and description of datasets». *Archives and Manuscripts*. Núm. *44*, p. 73-85. <https://doi.org/10.1080/01576895.2016.1179585>.

OLSON, J. E. (2003). «Data Quality: The Accuracy Dimension». Dentro: *Data Quality: The Accuracy Dimension*. <https://doi.org/10.1016/B978-1-55860-891-7.X5000-8>.

ONI, S.; CHEN, Z.; HOBAN, S.; *et al.* (2019). «A comparative study of data cleaning tools». *International Journal of Data Warehousing and Mining*. Vol. 15, núm. 4, p. 48-65. <https://doi.org/10.4018/IJDWM.2019100103>.

PASTOR SÁNCHEZ, J. A.; LLANES PADRÓN, D. (2017). «*Records in Contexts* y la publicación de conjuntos de datos archivísticos interoperables». Dentro: MELO SIMÕES, M. da G.; MANUEL BORGES, M. (ed.). *Tendências atuais e perspetivas futuras em organização do conhecimento*, p. 587-599. Centro de Estudos

Interdisciplinares do Século XX - CEIS20. <https://dialnet.unirioja.es/descarga/libro/719647.pdf>.

QUÍLEZ, J. (2017). «Integración de archivos, museos y bibliotecas con tecnologías de la web semántica: de la modelización conceptual a la iniciativa linked open data». *Lligall*. Núm. 40 (2017), p.16-57. <https://www.raco.cat/index.php/lligall/article/view/340577/431359>.

SHEPHERD, E.; Smith, C. (2000). «The application of ISAD(G) to the description of archival datasets». *Journal of the Society of Archivists*. Vol. 21, núm. 1, p. 55-86. <https://doi.org/10.1080/00379810050006911>.

THE NATIONAL ARCHIVES (2011). *Managing the Continuity of Datasets*, p. 1-35. <http://www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm;orwriteto>.

Pilar Campos Martínez

Archivera del Instituto Municipal de Hacienda de Barcelona y vocal de comunicaciones de la Asociación de Archiveros-Gestores de Documentos de Catalunya.
Vocal de proyección internacional con participación en la Conferencia Internacional de Archivos y Proyecto Interpares Trust AI.

Licenciada en Documentación y Máster de Gestión Documental e Información en Empresas.
Ha trabajado en la Universidad de Barcelona diseñando e implantando el sistema de gestión documental y herramientas de administración electrónica.
Ha sido consultora de gestión documental y formadora en el ámbito público y privado.