

Guía práctica para la publicación de datos tabulares en archivos CSV

CSV

Contenido elaborado por Carlos de la Fuente García, experto en datos abiertos.

Este estudio ha sido desarrollado en el marco de la Iniciativa Aporta, desarrollada por el Ministerio de Asuntos Económicos y Transformación Digital, a través de la Entidad Pública Empresarial Red.es.

El uso de este documento implica la expresa y plena aceptación de las condiciones generales de reutilización referidas en el aviso legal que se muestra en: <http://datos.gob.es/es/aviso-legal>

Agradecimientos:

Por sus contribuciones, sugerencias y su disponibilidad en la elaboración de esta guía a:

- *Alicia Martínez Domingo (Experta en datos abiertos).*
- *Jorge López Pérez (Data Scientist)*
- *Jorge Cimentada Báez (Research Scientist)*
- *Comunidad Databeers Málaga:*
 - *María Sánchez Gonzalez (Universidad de Málaga).*
 - *David Bueno Vallejo (Ayuntamiento de Málaga)*
 - *Nuria Portillo Poblador (Universitat Politècnica de València)*
 - *Roberto Magro Pedroviejo (Ayuntamiento del Alcobendas)*

Índice

01	Introducción.....	04
02	El formato de datos tabulares.....	06
03	Diccionarios de datos.....	09
04	Pautas generales para publicar archivos CSV	14
P1:	Línea única de cabecera	15
P2:	Registro único por fila	18
P3:	Nombrado de columnas	21
P4:	Estructuras de datos verticales vs horizontales	23
P5:	Tratamiento de valores desconocidos	25
P6:	Subtotales, totales o agrupamientos	28
P7:	Tipificación de datos	31
P8:	Estandarización de valores de los campos	33
P9:	Campos codificados	35
P10:	Campos de texto	37
P11:	Campos numéricos	39
P12:	Campos con fechas	43
P13:	Campos con números de teléfono	45
P14:	Campos con direcciones postales	47
P15:	Campos con coordenadas geográficas	49
05	Pautas para exportar/importar datos tabulares desde herramientas de hojas de cálculo a CSV.....	51
06	Otras consideraciones sobre archivos CSV	53
07	Toolbox para archivos CSV.....	54
08	Dónde encontrar buenos ejemplos de datos en formato CSV	59
09	Enlaces de interés.....	66
Anexo I	Taxonomías y listas de códigos de uso común.....	67

01 Introducción

Hoy en día disponemos cada vez de más fuentes de datos a nuestro alcance. Sin embargo, paradójicamente, aun cuando los datos son más asequibles que nunca, las posibilidades de reutilizarlos son bastante limitadas. Los potenciales usuarios de esos datos tienen que hacer frente muchas veces a múltiples barreras que dificultan su acceso y su uso: baja calidad de datos, metadatos escasamente descriptivos y estandarizados, imprecisión de licencias o el uso inadecuado de formatos.

Estas dificultades para reutilizar los datos se mencionan una y otra vez en varios estudios de referencia como el [Open Data Barometer](#) o el [Global Open Data Index](#), y son debidas, en buena parte, al convencimiento inicial de los productores de los datos de que lo importante era publicar la mayor cantidad de información cuanto antes sin importar su calidad.

Como consecuencia, los catálogos de datos publican decenas de miles de datasets con deficiencias de calidad que solo pueden ser identificadas después de comenzar el proceso de reutilización, generando una carga de depuración y preparación en muchos casos inasumible para el usuario de datos abiertos. Este hecho produce frustración y pérdida de interés en el sector reutilizador afectando a la credibilidad de las instituciones publicadoras y a rebajar considerablemente las expectativas de retorno y generación de valor a partir de la reutilización de datos abiertos.

Por todo ello, y teniendo en cuenta el estado actual de madurez de las iniciativas de datos abiertos, resulta oportuno fortalecer la mejora de la calidad de los datos que se publican. Se inicia con esta guía la publicación de un compendio de pautas recopiladas con el objetivo de orientar a los publicadores en el uso adecuado de formatos y medios de acceso a datos abiertos. En esta ocasión, el foco es el formato CSV que es el más frecuentemente utilizado en la publicación de datos abiertos.

■ ¿Por qué CSV?

- La forma tabular de los datos es la más habitual en las transferencias e intercambios de información y se produce de múltiples maneras: archivos de datos con columnas delimitadas o campos de longitud fija, hojas de cálculo, tablas HTML, o descargas de tablas de datos SQL, entre otras.
- Se trata del formato más popular y utilizado en el contexto de la reutilización de Datos Abiertos. La mayoría de los recursos disponibles en los catálogos de Datos Abiertos, se encuentran en formato CSV.
- El portal de datos europeo dispone de más de 120 mil conjuntos de datos en formato CSV, siendo el formato que más abunda en este [catálogo de Datos Abiertos](#).
- Por su parte, el catálogo nacional datos.gob.es cuenta con casi 14 mil datasets en formato CSV, siendo igualmente, el formato mayoritario.
- Es conciso, fácil de interpretar tanto para personas como para máquinas y adecuado para la estructura natural de la mayoría de los datos. Se caracteriza por [contener datos dispuestos en forma de tablas](#), donde los campos están individualizados por un carácter separador y los registros por saltos de línea.
- No se requiere ningún software específico para abrir archivos en formato CSV, tan solo es suficiente usar cualquier editor de texto disponible en todos los sistemas operativos.

Pero...

- La simplicidad de CSV tiene como contrapartida que el formato no incluye ningún mecanismo para indicar el tipo de datos que hay en una columna o si los valores de ésta deben ser expresados obligatoriamente. Por lo tanto, es propenso a errores como valores ausentes o a la mezcla de diferentes tipos de datos dentro de una columna.
- La solución a estos problemas radica en la aplicación de buenas prácticas en la fase de preparación de conjuntos de datos, la articulación de medidas de control de calidad y la vinculación del archivo de datos tabulares con esquemas que expresen el modelo de datos por medio de metadatos.

02 Formato de Datos tabulares

- Los conjuntos de **datos tabulares** bien organizados se ajustan a un esquema predefinido, son fáciles de manipular, modelar y visualizar, y tienen una estructura específica basada en las siguientes reglas:

- Cada variable es una **columna**.



- Cada observación o registro es una **fila**.



- Cada intersección de fila y columna es una **celda**.



- Cada conjunto de observaciones es una **tabla**.



- Ejemplo:** Características de coches clásicos¹

marca	año	cilindros	consumo	potencia	aceleracion
chevrolet chevelle malibu	1970	8	18	130	12
buick skylark 320	1970	8	15	165	11,5
plymouth satellite	1970	8	18	150	11
amc rebel sst	1970	8	16	150	12
ford torino	1970	8	17	140	10,5

(1) Adaptación del "Auto MGP Dataset". Disponible en: <https://archive.ics.uci.edu/ml/datasets/auto+mpg>

Aunque no hay un estándar oficial para el formato de “valores separados por comas” (CSV, por sus siglas en inglés), el *Internet Engineering Task Force* (IETF) publica el documento de referencia [RFC4180](#).

■ Características fundamentales:

- Cada archivo debe contener una sola tabla de datos.
- Cada registro o fila es una línea.
- Todos los registros contienen el mismo número de campos o columnas, al menos una.
- Opcionalmente, puede haber una primera línea de cabecera que contiene exclusivamente los nombres de los campos.
- Las celdas de una misma columna proporcionan valores para la misma propiedad de las observaciones descritas en cada fila.
- Todos los valores de una misma columna deben ser del mismo tipo de datos (texto, enteros, decimales, fecha, etc.)
- Cada campo está separado del siguiente por un carácter singular: por ejemplo, una coma [“,”], un punto y coma [“;”], un carácter *pipe* [“|”] o un carácter tabulador [TAB].
- Cuando los campos están separados por un carácter tabulador [TAB], el formato de archivo es [TSV](#).
- Alternativamente, los campos pueden tener una longitud fija de caracteres.
- Los valores de los campos que incluyen comillas, comas o retornos de carro deben ir entre comillas.
- Los archivos en formato CSV deben utilizar la codificación de caracteres UTF-8.
- Respecto a los nombres de los archivos, es recomendable utilizar minúsculas con los caracteres de la a-z, los dígitos 0-9 y carácter guion bajo (‘_’) en lugar de espacios en blanco, para asegurar el procesamiento correcto de nombres de archivo tanto en servidores como en aplicaciones cliente.
- Cualquier información que no sea valores de datos, como metadatos, descripciones, comentarios o unidades de medida, deben indicarse de forma anexa al archivo de datos en forma de diccionario de datos.

02

CSV

■ Ejemplo de archivo CSV

- Tabla de datos:



marca	año	cilindros	consumo	potencia	aceleracion
chevrolet chevelle malibu	1970	8	18	130	12
buick skylark 320	1970	8	15	165	11,5
plymouth satellite	1970	8	18	150	11
amc rebel sst	1970	8	16	150	12
ford torino	1970	8	17	140	10,5

- Archivo en formato CSV :

```
marca,año,cilindros,consumo,potencia,aceleración
"chevrolet chevelle malibu",70,8,18,130,12
"buick skylark 320",70,8,15,165,11.5
"plymouth satellite",70,8,18,150,11
"amc rebel sst",70,8,16,150,12
"ford Torino",70,8,17,140,10.5
```

CSV

03 Diccionarios de Datos

- El diccionario de datos es un complemento esencial de cualquier conjunto de datos ya que aporta al usuario de los datos información suficiente para procesar y comprender su contenido sin ambigüedad.
- Su propósito es garantizar que la estructura del conjunto de datos se define en términos fácilmente entendibles por los usuarios.
- Cada conjunto de datos publicable debe incluir su diccionario de datos como un documento separado, accesible normalmente mediante una URL desde el punto de descarga del archivo de datos.
- El contenido de un Diccionario de Datos puede ser expresado de diferentes formas, incluso como un archivo de texto que describe el contenido de cada columna del dataset.
- Las características o anotaciones que se expresan en el Diccionario de Datos son las propiedades de las tablas, columnas, filas y celdas que componen el conjunto de datos.
- A continuación, se expone un ejemplo de diccionario de datos expresado como un archivo de texto que puede ser suministrado a través de un servidor Web. Por ejemplo, a través de la URL: <http://example.org/automoviles.csv-metadata.txt>.

■ Ejemplo de diccionario de datos expresado como un archivo de texto



Archivo de datos: <http://example.org/automoviles.csv>

Descripción: Tabla con datos de automóviles clásicos

Publicador: Autor del ejemplo

Columna 1:

Título: marca

Descripción: Este campo contiene información sobre la marca y modelo de cada vehículo.

Tipo de datos: string

Columna 2:

Título: año

Descripción: Este campo contiene información sobre el año de fabricación de cada vehículo.

Tipo de datos: date

Columna 3:

Título: cilindros

Descripción: Este campo contiene información sobre el número de cilindros de cada vehículo.

Tipo de datos: integer

Columna 4:

Título: consumo

Descripción: Este campo contiene información sobre el consumo medio de cada vehículo, medido en litros / 100 kms.

Tipo de datos: decimal

Columna 5:

Título: potencia

Descripción: Este campo contiene información sobre la potencia de cada vehículo, medida en CV.

Tipo de datos: decimal

Columna 6:

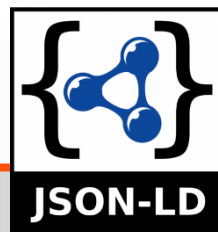
Título: aceleración

Descripción: Este campo contiene datos sobre la aceleración de cada vehículo, medida en m/seg2.

Tipo de datos: decimal

- Una buena práctica es que el Diccionario de Datos se exprese en un formato procesable, por ejemplo, [JSON](#) o [JSON-LD](#), mediante un vocabulario estandarizado que permita definir cada una de las características o anotaciones sobre cada uno de los elementos del archivo de datos.
- En algunas plataformas de Datos Abiertos, por ejemplo, las implementadas con [CKAN](#), se especifica el Diccionario de Datos como una sección asociada a cada recurso de un dataset.
- Según la [Norma Técnica de Interoperabilidad de Recursos de Información \(NTI\)](#), la forma de especificar el modelo de datos es utilizando la propiedad “*dct:relation*” en los metadatos de la distribución de los recursos del dataset.
- Para ajustar los valores de las propiedades es recomendable utilizar un lenguaje claro y conciso, dado que los usuarios finales de los datos no tienen porqué estar familiarizados con los datos.
- W3C recomienda un [modelo para datos tabulares](#) y propone un [vocabulario para la descripción de estas propiedades](#).
- El vocabulario de W3C es muy exhaustivo, no obstante, hay una serie de propiedades que es recomendable tener en cuenta para cualquier archivo tabular:
 - Para las tablas:
 - Título de tabla [**“dc:title”**]
 - Descripción [**“dc:description”**]
 - Publicador [**“dc:creator”**]
 - Ubicación del archivo que se describe [**“url”**]
 - Para las columnas:
 - Nombre de columna [**“name”**]
 - Título de columna [**“titles”**]
 - Descripción [**“dc:description”**]
 - Tipos de datos: [**“datatype”**]
- Además, es posible anotar por medio del uso de diferentes propiedades, entre otros metadatos, los siguientes: orden de columnas, valores esperados, valores requeridos, valores únicos, claves externas, listas de valores, idiomas de las cadenas, formatos, restricciones, validaciones, instrucciones para la transformación del CSV a otro formato.
- A continuación, se expone un ejemplo de diccionario de datos expresado como un esquema en formato JSON que puede ser suministrado a través de un servidor Web. Por ejemplo, a través de la URL: <http://example.org/automoviles.csv-metadata.json>.

■ Ejemplo de diccionario de datos utilizando el formato json-ld



```
{
  "@context": "http://www.w3.org/ns/csvw",
  "@type": "Table",
  "url": "http://example.org/automoviles.csv",
  "dc:description": "Tabla con datos de automóviles clásicos",
  "dc:creator": "Autor del ejemplo",
  "tableSchema": {
    "columns": [{
      "name": "identificador",
      "titles": "marca",
      "dc:description": "Este campo contiene información sobre la marca y modelo de cada vehículo",
      "datatype": "string" },
      {
        "name": "año",
        "titles": "año",
        "dc:description": "Este campo contiene información sobre el año de fabricación de cada vehículo",
        "datatype": {
          "base": "date",
          "format": "yyyy" },
      {
        "name": "cilindros",
        "titles": "cilindros",
        "dc:description": "Este campo contiene información sobre el número de cilindros de cada vehículo",
        "datatype": "integer" },
      {
        "name": "consumo",
        "titles": "consumo",
        "dc:description": "Este campo contiene información sobre el consumo medio de cada vehículo, medido en litros / 100 kms.",
        "datatype": "decimal" },
      {
        "name": "potencia",
        "titles": "potencia",
        "dc:description": "Este campo contiene información sobre la potencia de cada vehículo, medida en CV.",
        "datatype": "decimal" },
      {
        "name": "aceleracion",
        "titles": "aceleración",
        "dc:description": "Este campo contiene datos sobre la aceleración de cada vehículo medida en m/seg2",
        "datatype": "decimal"
      }
    ]
  }
}
```

03

- El diccionario de datos que se muestra como ejemplo está asociado al conjunto de datos que se muestra en esta guía.
- Entre otras propiedades, se está describiendo el nombre de cada una de las columnas utilizando la propiedad **"name"**.
- El uso de la propiedad **"title"** sirve para especificar la línea de cabecera de la tabla. La ausencia de esta propiedad indica que la tabla carece de línea de cabecera.
- Este encabezamiento de cada columna puede estar acompañado de un código de idioma de tal forma que la línea de cabecera se pueda expresar en diferentes idiomas.
- Se usa la propiedad **"datatype"** para describir los tipos de datos en los que se expresa cada valor de la columna correspondiente.
- La propiedad **"description"** permite incluir un texto descriptivo del contenido de cada columna.
- Herramientas como las que se describen en el apartado **"[Toolbox para archivos CSV](#)"** de esta guía, por ejemplo, CSVlint, permiten verificar la consistencia del conjunto de datos comparando el contenido del diccionario de datos y la estructura del archivo de datos.
- Entre otras comprobaciones se puede validar el número de columnas y su nombre, el tipo de valores permitidos, si éstos deben ser únicos o si existen vinculaciones de estos valores con otras tablas (claves externas).

CSV

04 Pautas para publicar archivo CSV

En el siguiente apartado de esta guía se recogen una serie de pautas relacionadas con los aspectos más comunes en la preparación de datos tabulares para su publicación como archivos en formato CSV.



P1-Línea única de cabecera



P2-Registro único por fila



P3-Nombrado de columnas



P4-Estructuras de datos verticales vs horizontales



P5-Tratamiento de valores desconocidos



P6-Subtotales, totales o agrupamiento



P7-Tipos de datos



P8-Estandarización de valores de los campos



P9-Campos codificados

...ABCD

P10-Campos de texto

...1234

P11-Campos numéricos

01/12/....

P12-Campos con fechas



P13- Campos con números de teléfono



P14-Campos con direcciones postales




P15-Campos con coordenadas geográficas

P1

Línea única de cabecera opcional

- Las tablas de datos pueden contener, opcionalmente, una y solo una línea de cabecera para especificar los nombres de los campos.
- A tener en cuenta:
 - La existencia de múltiples líneas de cabecera, aunque pueden incrementar la interpretación de los datos para las personas por su expresividad y formato, dificultan el procesamiento para las máquinas, por tanto, cualquier información adicional sobre los datos debe incluirse en la descripción de los mismos utilizando los metadatos apropiados en el Diccionario de Datos.
 - **Los nombres de las columnas** que se incluyen en la línea de cabecera son un tipo de anotación o metadato que describe cada columna y **no forma parte de los datos**, es decir, no se debe considerar cuando se cuenta el número de filas de datos en una tabla.
 - Para nombrar las columnas se deben usar **celdas simples** y en ningún caso, celdas combinadas.
 - Hay que tener en cuenta que no existe un mecanismo para discernir automáticamente si el primer registro de un CSV es una línea de cabecera ya que ésta se codifica como cualquier otro registro. Por tanto, es buena práctica especificar la presencia o ausencia de línea de cabecera, a través del diccionario de datos incluyendo la propiedad **"title"**.
 - Otra forma de indicar la presencia o ausencia de la línea de cabecera es mediante un parámetro del tipo de contenido cuando el archivo de datos es transmitido vía HTTP, de la forma: **Content-Type: text/csv;header=absent**.

■ **Ejemplo 1:** No usar múltiples celdas de cabecera



Datos sobre la de ventas de coches (años 1998 – 1999)		
Unidades expresadas en miles		
marca	año	ventas_por_año
chevrolet chevelle malibu	1998	2,50
chevrolet chevelle malibu	1999	2,63
buick skylark 320	1998	3,40
buick skylark 320	1999	3,57




marca	año	ventas_por_año
chevrolet chevelle malibu	1998	2,50
chevrolet chevelle malibu	1999	2,63
buick skylark 320	1998	3,40
buick skylark 320	1999	3,57


La información “Datos sobre la de ventas de coches (años 1998 – 1999)” y “Unidades expresadas en miles”, se debe trasladar al diccionario de datos utilizando la propiedad “description”.

CSV

■ **Ejemplo 2:** No usar celdas combinadas



marca	contacto_concesionario	
	concesionario_mail	concesionario_telefono
chevrolet chevelle malibu	mail@concesionario_chevrolet.com	+34-1111111
buick skylark 320	mail@concesionario_buick.com	+34-2222222



marca	contacto_concesionario_mail	contacto_concesionario_telefono
chevrolet chevelle malibu	mail@concesionario_chevrolet.com	+34-1111111
buick skylark 320	mail@concesionario_buick.com	+34-2222222

CSV

P2 Registro único por fila

- Las tablas de datos contienen observaciones sobre los datos y cada observación o registro de datos debe ocupar una fila. En cada registro debe haber exactamente el mismo número de campos.
- A tener en cuenta:
 - Cada registro o fila está marcado por una secuencia de uno o más caracteres invisibles denominados carácter de control, concretamente el retorno de carro (CR, *carriage return*) y salto de línea (LF, *line feed*). Desafortunadamente, los sistemas operativos representan los finales de línea usando diferentes secuencias:
 - Todas las versiones de DOS / Microsoft Windows representan finales de línea como CR seguidos de LF, es decir, CRLF o lo que es lo mismo ("`\r\n`").
 - Los sistemas operativos UNIX y similares, incluido MacOS, representan las terminaciones de línea como LF o lo que es lo mismo ("`\n`").
 - El documento de referencia RFC4180 para datos en formato CSV, define que las filas deben ser terminadas con los caracteres de control CRLF. Por tanto, es importante saber que esta cuestión puede generar algún problema cuando se intercambian, importan o exportan, archivos CSV con origen en diferentes sistemas operativos.
 - La existencia de caracteres de retorno de carro [CR] dentro del valor de un campo, por ejemplo, campos que contienen múltiples líneas o comentarios, deben ir siempre entre comillas dobles.

- **Ejemplo 1:** Comportamiento de una importación de un archivo CSV que contiene espacios en blanco y/o retornos de carro en varias circunstancias.

- **Caso 1: Importación de un CSV con valores de campos que incluyen espacios en blanco entrecomillados y separados por ‘,’**

```
marca,año,cilindros
"chevrolet chevelle malibu",1970,8
"buick skylark 320",1970,8
"plymouth satellite",1970,8
```



marca	año	cilindros
chevrolet chevelle malibu	1970	8
buick skylark 320	1970	8
plymouth satellite	1970	8

- **Caso 2:** Importación de un CSV con valores de campos que incluyen espacios en blanco y un carácter CR no entrecomillados.

```
marca,año,cilindros
chevrolet chevelleCRmalibu,1970,8
buick skylark 320,1970,8
plymouth satellite,1970,8
```



marca	año	cilindros
chevrolet chevelle		
malibu	1970	8
buick skylark 320	1970	8
plymouth satellite	1970	8

En el ejemplo anterior se están incluyendo el carácter de control de retorno de carro (CR) a modo ilustrativo. Tanto el carácter ‘CR’ como ‘LF’ no son visibles.

- **Ejemplo 2:** Comportamiento de la exportación de una tabla con campos que incluyen espacios en blanco o saltos de línea dentro de sus valores.

El **CSV** que se genera a partir de la tabla debe incluir el valor de cada campo entre comillas, incluyendo los caracteres de control, en un registro único.

campo_1	campo_2	campo_3
aaaaa	Bbbb Bbb Bbbb Bb	cccc
Aa aa	Bb,bb	c
Aaa aaaaaa	Bbbbb bbb bb	Ccc cccc

campo_1,campo_2,campo_3CRLF
 "aaaaa", "BbbbCRLFBbbCRLFBbbbCRLFbb", "cccc"CRLF
 "Aa aa", " bb,bb", "c"CRLF
 "AaaCRLFaaaaaa", " Bbbbb bbb bb", " CccCRLFcccc"LF


Se están incluyendo los caracteres de control de fin de línea (LF) y retorno de carro (CR) a modo ilustrativo, dado que éstos no son visibles.

CSV

P3 Nombrado de columnas

- Los nombres de los campos o columnas de una tabla de datos deben ser entendibles por las personas.
- A tener en cuenta:
 - En numerosas ocasiones las columnas de las tablas de datos mantienen los nombres asignados por los sistemas de gestión de bases de datos, normalmente sujetos a convenciones de índole técnico y difícilmente comprensibles para las personas.
 - Algunas recomendaciones para el nombrado de campos:
 - No repetir nombres de campo.
 - Usar nombres cortos (del orden de 20 caracteres) pero siempre teniendo en cuenta que el ahorro de caracteres no debe inducir malas interpretaciones del nombre del campo.
 - Evitar el uso de abreviaturas.
 - Utilizar solo caracteres ASCII en minúsculas (a-z; 0-9)
 - No usar caracteres especiales (por ejemplo: äüöàèèê, etc.)
 - No incluir tildes ni signos de puntuación.
 - Usar guiones bajos "_" para separar palabras que componen los nombres de columnas en lugar de espacios en blanco.
 - Evitar el uso de códigos, y si fuese absolutamente necesario, debe estar totalmente explicado en el diccionario de datos que documenta el dataset.
 - Los nombres de campo deben coincidir con los especificados en el diccionario de datos.

- **Ejemplo:** nombrado comprensible de columnas



Identificador_M	Año	Cil.	Consumo_por_cada_100_k ms_de_recorrido_urbano	HP	m/seg'
chevrolet chevelle malibu	1970	8	18	130	12
buick skylark 320	1970	8	15	165	11,5
plymouth satellite	1970	8	18	150	11



marca	año	cilindros	consumo	potencia	aceleracion
chevrolet chevelle malibu	1970	8	18	130	12
buick skylark 320	1970	8	15	165	11,5
plymouth satellite	1970	8	18	150	11

- Si existe una entidad que engloba varias características separadas en campos diferentes, es conveniente comenzar nombrando los campos con esa entidad y luego con los atributos más específicos (de lo más general a lo más específico). Por ejemplo:

cliente_nombre	cliente_cargo	solicitante_tipo_documento	solicitante_numero_documento

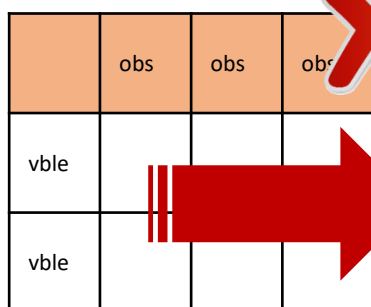
- Los campos que sean identificadores, pueden incluir el sufijo "_id" en el nombre del campo, salvo casos excepcionales donde un nombre alternativo sea más conveniente porque ofrece información sobre el sistema de identificación usado.
- En cuanto a los campos que contengan la descripción de ese identificador, se recomienda que incluyan el sufijo "_nombre", salvo que exista una forma más conveniente de nombrar el campo.

concesionario_id	concesionario_nombre

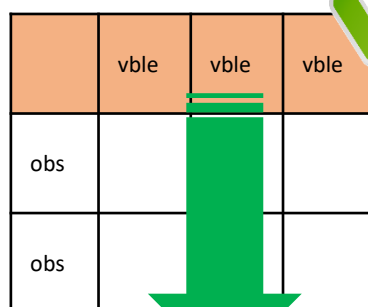
P4

Estructuras de datos verticales vs horizontales

- Cuando se diseñan estructuras de datos tabulares es recomendable evitar el crecimiento horizontal de valores.




	obs	obs	obs
vble			
vble			



	vble	vble	vble
obs			
obs			

- Siempre que sea posible es preferible situar las variables o atributos de los datos en las columnas de una tabla y añadir los valores correspondientes a las observaciones de los mismos en las filas.
- El crecimiento horizontal de una estructura de datos tabular puede dificultar su mantenimiento y la confección de visualizaciones.
- Por lo general, es más fácil identificar relaciones entre variables en columnas que entre filas y es más fácil hacer comparaciones entre grupos de observaciones, en filas, que entre grupos de columnas.
- No obstante, esta recomendación debe ajustarse según las necesidades de actualización de los datos:
 - Si es necesario registrar nuevas variables o atributos que no se habían registrado previamente, por ejemplo: una serie temporal, entonces es razonable el crecimiento horizontal de la estructura de datos. es decir, añadir nuevas columnas. Esto permitirá insertar observaciones para las nuevas variables manteniendo valores en blanco en las observaciones previas a la actualización para estas nuevas columnas, si es que no existe un valor asignable a esas observaciones. Al añadir nuevas observaciones, necesariamente tienen que introducirse nuevas filas.

■ **Ejemplo:** crecimiento horizontal vs vertical




marca	averias_radiador	averias_carburador	averias_suspension	averias_embrague
chevrolet chevelle malibu	0	7	1	0
buick skylark 320	1	2	2	2
plymouth satellite	0	4	4	1



marca	tipo_averia	cantidad_averias
chevrolet chevelle malibu	radiador	0
chevrolet chevelle malibu	carburador	7
chevrolet chevelle malibu	suspensión	1
chevrolet chevelle malibu	embrague	0
buick skylark 320	radiador	1
buick skylark 320	carburador	2
buick skylark 320	suspensión	2
buick skylark 320	embrague	2
plymouth satellite	radiador	0
plymouth satellite	carburador	4
plymouth satellite	suspensión	4
plymouth satellite	embrague	1

En el ejemplo se observa una forma de disponer los datos evitando el crecimiento horizontal de la estructura de datos agregando nuevas variables similares a las existentes. La trasposición a una estructura vertical mediante la creación de dos variables nuevas, “tipo_averia” y “cantidad_averias”, permite añadir fácilmente nuevas observaciones en forma de filas.

En cambio, cuando se publican series temporales, por ejemplo, el histórico de demanda de vehículos entre los años 1972-1977, es razonable el crecimiento horizontal de la estructura si surge la necesidad, por ejemplo, de completar la serie histórica de los años ‘70.




marca	1972	1973	1974	1975	1976	1977
chevrolet chevelle malibu	345	423	1234	1690	2345	2134
buick skylark 320	124	252	785	914	1353	896
plymouth satellite	57	71	165	315	1104	1561

P5 Tratamiento de valores desconocidos


- Los valores de los datos deben ser completos y deben estar expresados de forma precisa y coherente con el tipo de datos del campo para que puedan ser procesados en función de su valor real.
- Como norma general, hay que rellenar todas las celdas de una tabla y mantener un código común para los datos desconocidos.
- A tener en cuenta:
 - Los valores desconocidos, cuando se dejan sin explicar o simplemente están ausentes, suelen generar confusión, especialmente cuando la columna de datos es numérica. Por otro lado, generan resultados erróneos en tareas de ordenación.
 - Recomendaciones para evitar valores de datos desconocidos:
 - Si la celda en blanco representa un cero, entonces el valor debe ser 0.
 - Si la celda en blanco representa un valor "desconocido" o "no obtenido", entonces esta posibilidad debe explicarse en el diccionario de datos e indicarse con un código específico.
 - Si un valor en blanco tiene un significado, se debe valorar la opción de añadir una nueva columna para incluir la explicación del valor "en blanco" como un valor posible.
 - Una terminología aceptada para indicar valores desconocidos o ausentes es el valor o código específico NA o N/A.¹
 - El código que se utilice para indicar los valores desconocidos o ausentes, por ejemplo NA, debe especificarse en el diccionario de datos.

(1) Del inglés, *not available* (no disponible), *not applicable* (no corresponde en el caso) o *no answer* (sin respuesta; aunque este significado solo se usa en ciertas situaciones). <https://es.wikipedia.org/wiki/N/a>

■ **Ejemplo 1:** Ventas de coches por año (en miles).




marca	año	consumo	ventas
chevrolet chevelle malibu	1998	Alto	2,50
chevrolet chevelle malibu	1999	Bajo	2,63
chevrolet chevelle malibu	2000	Medio	
buick skylark 320	1998		3,40
buick skylark 320	1999	Medio	3,57
buick skylark 320	2000	Medio	N/A
plymouth satellite	1998		2,40
plymouth satellite	1999		2,52
plymouth satellite	2000	Alto	3,60




marca	año	consumo	ventas
chevrolet chevelle malibu	1998	Alto	2,50
chevrolet chevelle malibu	1999	Bajo	2,63
chevrolet chevelle malibu	2000	Medio	0
buick skylark 320	1998	NA	3,40
buick skylark 320	1999	Medio	3,57
buick skylark 320	2000	Medio	NA
plymouth satellite	1998	NA	2,40
plymouth satellite	1999	NA	2,52
plymouth satellite	2000	Alto	3,60

En el ejemplo se observa que el valor 0 en la columna “ventas” indica que para ese año las ventas de coches de ese modelo han sido 0. En cambio, cuando el dato de “ventas”, al igual que el de “consumo” se desconoce, se indica con NA. Todos los valores desconocidos en cualquier columna se indican con el mismo código: NA.

▪ **Ejemplo 2:** Ventas de coches por año (en miles).



marca	año	consumo	ventas
chevrolet chevelle malibu	1998	Alto	2,50
chevrolet chevelle malibu	1999	Bajo	2,63
chevrolet chevelle malibu	2000	Medio	3,75
buick skylark 320	1998	NA	
buick skylark 320	1999	Medio	3,57
buick skylark 320	2000	Medio	5,10
plymouth satellite	1998	NA	
plymouth satellite	1999	NA	2,52
plymouth satellite	2000	Alto	3,60



marca	año	consumo	ventas	Significado valor ausente ventas
chevrolet chevelle malibu	1998	Alto	2,50	
chevrolet chevelle malibu	1999	Bajo	2,63	
chevrolet chevelle malibu	2000	Medio	3,75	
buick skylark 320	1998	NA		< umbral mínimo (0.1)
buick skylark 320	1999	Medio	3,57	
buick skylark 320	2000	Medio	5,10	
plymouth satellite	1998	NA		< umbral mínimo (0.1)
plymouth satellite	1999	NA	2,52	
plymouth satellite	2000	Alto	3,60	

En el ejemplo se añade una nueva columna para explicar el significado del valor ausente en la columna “ventas”.

Hay circunstancias en las que no se pueden registrar determinadas medidas, porque los aparatos o sistemas que se utilizan para medir determinadas magnitudes solo registran valores a partir de un determinado umbral (por ejemplo, un sensor de contaminación ambiental). En esos casos, se explicará en el diccionario de datos y se indicará en la tabla con un código común.

P6

Subtotales, totales o agrupamientos

- 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.
- A tener en cuenta:
 - Un archivo que contiene resultados y/o operaciones realizadas con los datos no se puede considerar un archivo de datos en sentido estricto, sino un archivo de resultados de un determinado análisis de datos.
 - Cuando se incluyen filas o columnas con valores de datos agregados por ejemplo como resultado de una operación, resulta muy difícil y en ocasiones imposible recuperar el dato desagregado.
 - Un dataset debe ser consistente en el nivel de granularidad de los datos que contiene. Si el nivel de granularidad se establece según una determinada dimensión, por ejemplo: ventas mensuales, no se deben mezclar datos con otro nivel de granularidad, por ejemplo, ventas anuales.
 - Un nivel de granularidad superior siempre se puede obtener a partir de un nivel inferior, pero no a la inversa. Siguiendo el ejemplo, es posible obtener las ventas anuales a partir de los datos de ventas mensuales, pero no es posible recuperar los datos de ventas mensuales a partir de las ventas anuales.
 - Se debe evitar el agrupamiento de filas relacionadas con una entidad dejando ciertas celdas vacías repitiendo la entidad para todas las filas del agrupamiento. Este problema es común y puede ocasionar problemas cuando se modifica el orden original de las filas.

CSV

- **Ejemplo 1:** Venta semestral de coches (en miles), con subtotales (mezcla de niveles de granularidad) y sin subtotales (mismo nivel de granularidad).




marca	año	ventas_semestrales
chevrolet chevelle malibu	1998	2,5
chevrolet chevelle malibu	1998	2,63
Subtotal anual	1999	5,13
buick skylark 320	1999	3,4
buick skylark 320	1999	3,57
Subtotal anual	1999	6,97
plymouth satellite	2000	2,4
plymouth satellite	2000	2,52
Subtotal anual	2000	4,92




marca	año	ventas_s1	ventas_s2
chevrolet chevelle malibu	1998	2,5	2,63
buick skylark 320	1999	3,4	3,57
plymouth satellite	2000	2,4	2,52

- **Ejemplo 2:** No usar agrupamientos en base al uso de celdas vacías.



marca	año	ventas_semestrales
chevrolet chevelle malibu	1998	2,5
	1999	2,63
	2000	3,13
buick skylark 320	1998	3,4
	1999	3,57
	2000	3,97



marca	año	ventas_semestrales
chevrolet chevelle malibu	1998	2,5
chevrolet chevelle malibu	1999	2,63
chevrolet chevelle malibu	2000	3,13
buick skylark 320	1998	3,4
buick skylark 320	1999	3,57
buick skylark 320	2000	3,97

Es importante tener en cuenta que la existencia de celdas vacías puede producir efectos no deseables ante posibles ordenaciones de los datos. En la tabla siguiente se observa el efecto que produce la ordenación de la tabla inicial según los valores del campo 'marca'.



marca	año	ventas_semestrales
buick skylark 320	1998	3,40
chevrolet chevelle malibu	1998	2,50
	1999	2,63
	2000	3,13
	1999	3,57
	2000	3,97

P7 Tipos de datos

- Los valores de una tabla de datos deben estar formateados acorde al [tipo de datos](#) de que se trate. Específicamente, los números siempre deben estar en celdas de formato/tipo "número", los campos de tipo textual deben estar en celdas de formato/tipo "texto" y los campos de tipo fecha deben estar en celdas de formato/tipo "fecha".

Dato	Tipo
Cadena de caracteres	string
Número	integer, decimal, float, double
Fecha	date, time, datetime, Year, Month, Day
binario	boolean

- A tener en cuenta:
 - Mantener el formato correcto de las celdas según el tipo de datos que contengan aumenta las probabilidades de que una exportación a otro formato se realice correctamente y logra que los datos sean más operables en la propia tabla de datos.

CSV

- **Ejemplo:** tipos de datos en tres columnas (string, integer, integer).



marca	consumo	potencia
chevrolet chevelle malibu	18	130 CV
buick skylark 320	15 litros	165
plymouth satellite	18	150 CV
amc rebel sst	16 l.	150 CV
ford torino	17	140



marca	consumo	potencia
chevrolet chevelle malibu	18	130
buick skylark 320	15	165
plymouth satellite	18	150
amc rebel sst	16	150
ford torino	17	140


Las buenas prácticas indican que las unidades de medida deben describirse en el diccionario de datos y no en el nombrado de los campos. En última instancia y si se carece de diccionario, es posible indicar la unidad de medida en el nombre del campo, por ejemplo: "consumo_litros" o "potencia_cv", siempre y cuando todos los valores de la columna tengan asociada la misma unidad de medida.

P8 Estandarización de valores de los campos


- El uso de valores estandarizados permite la correlación de datos entre conjuntos de datos, la comparación inter-administraciones (entre agencias y/o sectores), la interoperabilidad y el enlazado de datos. Para ello, los valores de determinados campos deben ser consistentes entre datasets.
- A tener en cuenta:
 - Solo es posible saber si una magnitud es grande o pequeña si se puede comparar con otra teniendo en cuenta las similitudes y diferencias, por ejemplo, entre conjuntos de datos originados por diferentes administraciones.
 - La [norma AENOR 137801:2015, Ciudades Inteligentes, Datos Abiertos](#), considera datos técnicamente correctos aquellos que, entre otras características:
 - Utilizan la misma codificación y normalización para el mismo tipo de dato publicado en diferentes datasets de un catálogo. Por ejemplo, las direcciones se publican siempre con la misma estructura, tipo, formatos en cualquier conjunto de datos y los elementos de georreferenciación utilizan el mismo sistema de coordenadas de referencia. ^[1]_[2]
 - La codificación y normalización utilizada se basa en algún estándar común reconocido y utilizado por otras organizaciones codificación. Por ejemplo: estándares aprobados por [EUROSTAT](#) o el [INE](#).
- Es recomendable:
 - Usar vocabularios de uso común para normalizar la estructura y valores de la información publicada en los conjuntos de datos.²
 - En el caso de no usar vocabularios de referencia, el valor que se asigne a un determinado atributo debe ser único y coherente en toda utilización de dicho valor a lo largo de la tabla. Es decir, si se opta por usar el valor "Barcelona", para referirse a esta ciudad, no se debe usar el valor "Ciudad de Barcelona".

(2) La Norma AENOR 137801:2015 incluye una relación de vocabularios de referencia disponibles en:
<http://vocab.linkeddata.es/datosabiertos/>

- **Ejemplo:** estandarización de la denominación y código de actividad económica.



marca	actividad_vendedor
chevrolet	Venta de coches
buick	Venta de vehículos
plymouth	Venta



marca	codigo_vendedor	actividad_vendedor
chevrolet	45.11	Venta de automóviles y vehículos de motor ligeros
buick	45.11	Venta de automóviles y vehículos de motor ligeros
plymouth	45.19	Venta de otros vehículos de motor

En este ejemplo, los valores del campo 'código_vendedor' son los correspondientes [a la nomenclatura estadística de actividades económicas de la Comunidad Europea](#) (NACE, Rev. 2) de EUROSTAT para la estandarización de las actividades económicas de los vendedores de vehículos.

CSV



P9 Campos codificados

- Referenciar correctamente esquemas de conceptos, es decir, listas de códigos (*code lists*) y taxonomías de términos como valores prescritos para propiedades o atributos definidos.
- A tener en cuenta:
 - Son muy útiles para realizar análisis de datos categorizados (por ejemplo: para extraer patrones, realizar filtrados, sumarios, ordenaciones, etc.), pero pueden ser difíciles de interpretar para personas no familiarizadas con tales codificaciones, si no están explicados con precisión en el propio conjunto de datos o por medio del Diccionario de Datos.
 - El [anexo I](#) de esta guía recoge una reseña de términos reutilizables que incluye taxonomías, clasificaciones y estándares armonizados a nivel nacional e internacional.
 - Si es posible, se debe añadir una columna incluyendo el valor estándar del atributo en cuestión. Por ejemplo. si el campo se refiere a Ayuntamientos, es aconsejable mantener dos columnas: "Código Ayuntamiento" y "Ayuntamiento". El Código será el referido en el ["Directorio Común de Unidades Orgánicas y Oficinas" \(DIR3\)](#) y el valor de la columna "Ayuntamiento", el nombre de esa institución.
 - En línea con lo anterior, no se debe incluir el nombre del campo dentro del valor de campo. Por ejemplo, si el nombre de la columna es "Ayuntamiento" el valor en cada fila no debe incluir las palabras "Ayuntamiento de".

CSV





- **Ejemplo:** uso de campos codificados



marca	consumo	potencia	cambio	aceleración
chevrolet chevelle malibu	18	130	M	12
buick skylark 320	15	165	A	11,5
plymouth satellite	18	150	A	11
amc rebel sst	16	150	M	12
ford torino	17	140	M	10,5



marca	consumo	potencia	cambio	cambio- descripción	aceleracion
chevrolet chevelle malibu	18	130	M	Manual	12
buick skylark 320	15	165	A	Automático	11,5
plymouth satellite	18	150	A	Automático	11
amc rebel sst	16	150	M	Manual	12
ford torino	17	140	M	Manual	10,5

CSV



P10 Campos de tipo texto

- Los campos de tipo texto deben incluir cadenas de texto sin espacios en blanco iniciales o finales. Si no es así, los caracteres (espacios o tabuladores) que existan adyacentes a los separadores de campo serán ignorados.
- A tener en cuenta:
 - Los campos de texto siempre se pueden delimitar con comillas dobles y las aplicaciones que realicen operaciones de lectura de datos analizarán y descartarán tales delimitadores.
 - Si el valor de un campo contiene alguna cadena de texto que presente alguno de los casos siguientes, debe tratarse de la forma indicada:
 - Si los espacios en blanco al principio o final de una cadena de texto son significativos y es necesario preservarlos, el campo debe estar delimitado con comillas dobles.
 - Cadenas que incluyen comas: el campo debe estar delimitado con comillas dobles.
 - Cadenas que incluyen texto entrecomillado: los caracteres de comillas dobles incrustados deben duplicarse y el campo debe delimitarse con comillas dobles.
 - Cadenas que incluyen saltos de línea: los campos deben estar entre comillas dobles.

CSV

- **Ejemplo 1:** comportamiento de la exportación de una tabla con valores que contienen espacios en blanco (iniciales o finales). Al exportar la tabla como CSV, los valores en blanco desaparecen.

marca	año	cilindros
BBBBBBchevrolet chevelle malibu	1970	8
BBBBBBbuick skylark 320BBBB	1970	8
plymouth satellite	1970	8

marca,año,cilindros
 "chevrolet chevelle malibu",1970,8
 "buick skylark 320",1970,8
 "plymouth satellite",1970,8

- **Ejemplo 2:** comportamiento de la importación de un archivo CSV con valores que contienen espacios en blanco (iniciales o finales) y caracteres entrecomillados.

campo1,campo2,campo3
 "valor del campo1"," valor del campo2 con espacios iniciales","valor del campo "3" que incluye un dato entre comillas"


campo1	campo2	campo3
valor del campo1	BBBBvalor del campo2 con espacios iniciales	valor del campo "3" que incluye un dato entre comillas

En estos ejemplos, se están indicado los espacios en blanco usando el carácter 'B' de modo ilustrativo.

P11 Campos de tipo numérico

- Los campos numéricos deben codificarse exclusivamente como tipos de datos numéricos (enteros o decimales). Si no es así, determinadas operaciones con los datos, como por ejemplo una ordenación, pueden generar problemas inesperados.
- A tener en cuenta:
 - De forma general:
 - No se deben usar separadores de millares. Tampoco espacios en blanco como separadores.
 - El separador decimal puede ser una ‘,’ o un ‘.', depende de la configuración regional de las aplicaciones de tratamiento de datos. En España, Francia o Alemania, entre otros países, se usa la ‘,’ y en el ámbito anglosajón el ‘.'. El uso de uno u otro carácter separador puede implicar el preprocesado de los datos para su reutilización con determinadas herramientas y lenguajes de programación.
 - Los valores negativos deben ir precedidos de un signo menos (-). No usar paréntesis para indicar valores negativos.
 - Si una columna contiene valores enteros y decimales, el tipo de dato debe ser decimal y por tanto, se incluirá el separador ‘,’ o ‘.’ y el número de cifras decimales que proceda, normalmente dos decimales.
 - Si una columna solo contiene valores enteros, se expresarán sin separador decimal.
 - No se debe mezclar texto con valores numéricos. Por ejemplo: no se debe usar 50€ o 27 km como valor en un campo numérico.
 - Recomendaciones en el caso de moneda:
 - Los valores numéricos deben expresarse sin decimales o con 2 decimales.
 - No debe variar el número de decimales usados para dar formato a toda la columna de valores. Si varía, se vulnera la característica de consistencia de los datos.
 - No incluir símbolo de moneda, ni separadores de millares (puntos o comas, según sea el caso).

- Recomendaciones en el caso de unidades de medida:
 - Se deben usar los espacios decimales que sean necesarios.
 - Es recomendable utilizar el diccionario de datos para expresar las unidades de medida asociadas a valores numéricos. Si se carece de diccionario, es posible indicar la unidad de medida en el nombre del campo, por ejemplo: "distancia_metros", siempre y cuando todos los valores de la columna tengan asociada la misma unidad de medida.
 - En el caso de que la unidad de medida tenga diferentes valores para una misma columna, los valores correspondientes a cada unidad de medida den indicarse en una columna separada seguida de la columna que contiene el valor numérico.
 - Por ejemplo:



marca	precio_venta	moneda
chevrolet chevelle malibu	23540,20	EUR
buick skylark 320	22189,00	EUR
plymouth satellite	28362,65	USD
amc rebel sst	29200,00	USD

- Recomendaciones en el caso de números que indican valores codificados:
 - Cuando existen ceros como números de cabecera de valores que si son significativos, deben ser tipificados como valores de texto para evitar que éstos sean truncados. Por ejemplo: Código con valor: "00000023456".

CSV

- **Ejemplo 1:** evitar el uso de separadores de miles. Uso de coma decimal. Consistencia en el número de cifras decimales. Evitar la mezcla de texto con valores numéricos.




marca	ingresos_ventas
chevrolet chevelle malibu	1.234.454,34 €
buick skylark 320	2.345.567,892 euros
plymouth satellite	344,678.23
amc rebel sst	267.331
ford torino	1.234678.67




marca	Ingresos_ventas
chevrolet chevelle malibu	1234454,34
buick skylark 320	2345567,89
plymouth satellite	344678,23
amc rebel sst	267331,00
ford torino	1234678,67

En este ejemplo, la moneda que se utiliza para todos los valores del campo “ingresos_ventas” es la misma y se describirá en el diccionario de datos. En su defecto, el nombre del campo podría ser “ingresos_ventas_euros”

- **Ejemplo 2:** usar el número de cifras decimales adecuado a cada tipo de datos numérico. Usar el signo menos (-) para valores negativos. Tipificar como texto valores con ceros significativos.



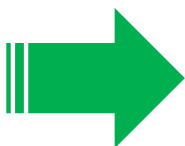
identificador_marca	punto_kilometrico	latitud	longitud
345600	2.3	43,2345678	(5,1234567)
0000345601	12,56	43,345	-5,2345678



identificador_marca	punto_kilometrico	latitud	longitud
0000345600	2,334	43,2345678	-5,1234567
0000345601	12,567	43,3456789	-5,2345678

- **Ejemplo 3:** comportamiento ante un proceso de ordenación en campos
 - **Caso 1:** resultado correcto tras un proceso de ordenación cuando el campo “ranking” es de tipo “numérico”.

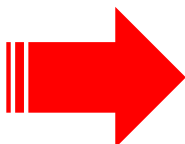
marca	ranking
chevrolet chevelle malibu	1
buick skylark 320	9
plymouth satellite	6
amc rebel sst	2
ford torino	4



marca	ranking
chevrolet chevelle malibu	1
amc rebel sst	2
ford torino	4
plymouth satellite	6
buick skylark 320	9

- **Caso 2:** resultado erróneo tras un proceso de ordenación cuando el campo “Ranking” es de tipo “texto”.

marca	ranking
chevrolet chevelle malibu	“123”
buick skylark 320	“9”
plymouth satellite	“324”
amc rebel sst	“11”
ford torino	“45”



marca	ranking
amc rebel sst	“11”
chevrolet chevelle malibu	“123”
plymouth satellite	“324”
ford torino	“45”
buick skylark 320	“9”

En este último ejemplo, se están incluyendo las comillas que abarcan a cada cadena de números con el objetivo de hacer más explícito el ejemplo, pero en una hoja de cálculo real no tienen porque ser visibles aun cuando el campo es de tipo “texto”.

Una buena práctica es comprobar en el sistema u hoja de cálculo donde se genera el archivo o por medio de una herramienta de validación como las que se detallan en el apartado [“Toolbox para archivos CSV”](#) cada tipo de datos antes de su publicación.

P12 Campos de tipo fecha

- Codificar las fechas usando un estándar es crítico para facilitar el análisis de series temporales de datos. El estándar de referencia en la NTI de reutilización es [ISO 8601](#), que codifica los valores de fecha con el formato YYYY-MM-DD, en su forma abreviada o YYYY-MM-DDTHH:MM:SS, en versión extendida.
- A tener en cuenta:
 - Recomendaciones en el caso de fechas:
 - Rellenar con ceros los valores de 0-9 para indicar MM, DD, HH, MM y SS, dado que cada valor debe tener un número fijo de dígitos.
 - Usar "-" como separador para fechas y ":" para horas.
 - Usar la letra "T" como carácter separador entre una fecha y una hora cuando se usa la versión extendida.
 - Es preferible usar fechas completas (YYYY-MM-DD) a indicar solo meses o años, siempre que esto sea posible a partir del origen de datos.
 - 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.
 - Si la fecha completa no se puede incluir, es preferible expresar año y mes en columnas separadas. Incluso es razonable incluir una columna para indicar el nombre del mes.

- Recomendaciones en el caso de horas:
 - Es preferible usar el formato de 24 horas (HH:MM:SS) en vez del formato de 12 horas AM/PM (HH:MM:SS AM/PM).
 - Es preferible usar un campo único para incluir valores que representan fecha y hora combinadas usando el formato YYYY-MM-DDTHH:MM:SS.

■ **Ejemplo 1:** campos con fechas (forma abreviada)

marca	fecha_lanzamiento
chevrolet chevelle malibu	Enero de 1970
buick skylark 320	23/07/1970
plymouth satellite	1-Febrero-1970



marca	fecha_lanzamiento
chevrolet chevelle malibu	1970-01-31
buick skylark 320	1970-07-23
plymouth satellite	1970-02-01



■ **Ejemplo 2:** campos con fechas y horas (forma extendida)

Identificador_vehiculo	fecha_venta
0003407G236	12 de Enero de 1970, a las 3:10 de la tarde
0003507H003	23/07/1970 – 04:34PM
0003407G237	1-Febrero-1970-12-23-34



marca	fecha_venta
0003407G236	1970-01-31T15:10:00
0003407G236	1970-07-23T16:34:00
0003407G236	1970-02-01T12:23:34




P13 Campos con números de teléfono


- Cuando se incluyen valores con números de teléfono, lo más importante es asegurar la consistencia en el formato de esos números a lo largo de todos los valores de la columna. Es decir, se podría usar +34-6660000 ó (34)6660000 ó 34-666-00-00, pero siempre el mismo formato.
- A tener en cuenta:
 - Cuando sea necesario incluir más de un número de teléfono como valor de un campo, deberán agregarse nuevos campos.
 - Es recomendable incluir el código de país precediendo al número de teléfono.
 - Para números de teléfono que requieran la inclusión de un número interno, por ejemplo, una extensión, deberá considerarse la inclusión de otro campo de tipo texto, ya que los números de uso interno pueden incluir determinados caracteres. Por ejemplo: "*86", "#36", etc.

CSV

- **Ejemplo:** tabla con una columna consistente que contiene números de teléfono



marca	concesionario_número_teléfono
chevrolet chevelle malibu	+34-6760000
buick skylark 320	(34)6960001
plymouth satellite	34-676-00-03
amc rebel sst	346960004



marca	concesionario_número_teléfono
chevrolet chevelle malibu	+34-6760000
buick skylark 320	+34-6960001
plymouth satellite	+34-6760003
amc rebel sst	+34-6960004

CSV




P14 Campos con direcciones postales

- Usar una codificación precisa de las direcciones postales es fundamental para manejar dimensiones geográficas con los datos. Las direcciones postales bien estructuradas pueden ser geolocalizadas, generando las coordenadas de latitud y longitud, usando aplicaciones específicas para ello.
- A tener en cuenta:
 - Si bien existe una norma técnica para el [diseño de registros de los ficheros de intercambio de información referente al Padrón municipal](#) (INE, Callejero de Censo Electoral), es aconsejable codificar una dirección postal como una cadena de caracteres con los elementos necesarios para localizar un domicilio dentro de una localidad.
 - Estos elementos son:
 - tipo de vía (calle, avenida, plaza, ...)
 - nombre de la vía, número, bloque, planta, letra, etc.
 - localidad (nombre)
 - código postal (5 caracteres numéricos)
 - Siempre que sea posible, es recomendable usar además campos diferenciados para los valores de latitud y longitud correspondientes al punto geográfico de la dirección postal.


CSV



- **Ejemplo:** codificación de direcciones postales en campos separados.



Marca	Dirección del Concesionario
chevrolet chevelle malibu	Calle Automoción, 23, Bajo, Alicante
buick skylark 320	C/ Industria, 33, 33201, Asturias



marca	concesionario_dirección_tipo_via	concesionario_dirección_nombre_via	concesionario_dirección_localidad	concesionario_dirección_código_postal
chevrolet chevelle malibu	Calle	automoción, 23, Bajo	Alicante	03011
buick skylark 320	Avenida	industria, 33	Oviedo	33201

CSV




P15 Campos con coordenadas geográficas

- El formato más adaptable para representar coordenadas geográficas en mapas es la especificación de latitud y longitud en grados decimales, cuyos valores deben presentarse en columnas separadas, con cabeceras de columna que deben llamarse: Latitud y Longitud, respectivamente.
- A tener en cuenta:
 - Si es necesario especificar el nombre de una entidad de la cual se consignan las coordenadas, se usarán los sufijos “_latitud” y “_longitud”.
 - Se pueden usar conversores de coordenadas UTM o de grados sexagesimales a grados decimales. Por ejemplo: la [herramienta de conversión](#) de la Junta de Andalucía.
 - Para datos geográficos que no sean puntos, por ejemplo: líneas o polígonos, es recomendable seguir [especificación Well-known text representations of coordinate reference systems, \(WKT-CRS\)](#) del Open Geospatial Consortium.
 - Es aconsejable, siempre que se publiquen datos geográficos, en formatos como por ejemplo [SHP](#) o [KML](#), acompañar estos archivos de un archivo CSV donde también se incluyan las coordenadas geográficas para facilitar su reutilización.






- **Ejemplo:** coordenadas geográficas usando notación decimal.



concesionario	latitud	longitud
chevrolet	43º 14' 04"	5º07'24"
buick	43º20'44"	5º14'14"



concesionario	latitud	longitud
chevrolet	43,2345678	-5,1234567
buick	43,3456789	-5,2345678

CSV

05 Pautas para exportar/importar datos tabulares desde herramientas de hojas de cálculo a CSV

- Las herramientas para el procesamiento de hojas de cálculo, por ejemplo, Microsoft Excel o Libreoffice Calc, pueden ocasionar problemas si se exportan tablas de datos directamente porque pueden contener: celdas combinadas, formulas, macros, pestañas de expansión de datos, u otras características derivadas de la aplicación de funcionalidades propias de estas herramientas.
- A tener en cuenta:
 - Con el fin de exportar adecuadamente datos tabulares a partir de herramientas de procesamiento de hojas de cálculo, es necesario contemplar las siguientes pautas:
 - Dependiendo de los ajustes regionales, puede ser más conveniente usar como separador “;” que “,”. En Europa, la “,” no es un separador apropiado dado que se utiliza como separador decimal, por tanto, es preferible usar “;”. En cambio, en países como UK o USA, el separador apropiado es “,”, dado que el separador decimal que se utiliza es “.”. No obstante, para evitar ambigüedades se suele utilizar otro tipo de separador como el separador “pipe”, “|”.
 - No usar celdas combinadas.
 - No dejar celdas, filas y/o columnas en blanco entre los datos.
 - No usar campos calculados.

- Una práctica adecuada para verificar que los datos exportados están formateados correctamente en CSV consiste en abrir el archivo en un programa de texto como el bloc de notas o editor de texto. Si la exportación es correcta, se verán los datos exportados exactamente como están formateados por el software de exportación.
- Cuando Excel importa datos CSV elimina los ceros a la izquierda de los campos antes de mostrarlos, si los campos no están entre comillas. Esto puede ocasionar problemas cuando, por ejemplo, esos campos contienen claves numéricas. Igualmente, elimina siempre los espacios iniciales.
- Es recomendable exportar datos tabulares en formato CSV usando la [codificación estándar de caracteres UTF-8](#).
 - Las hojas de cálculo de Excel que contienen algunos símbolos especiales, como la letra "ñ", tildes, acentos, etc. o algún otro tipo de carácter especial, pueden generar problemas al realizar una exportación de a CSV y su posterior recuperación, porque el comando "Guardar como CSV" distorsiona cualquier carácter que no sea ASCII.
 - Si la versión de Excel utilizada no permite guardar directamente el archivo de datos codificado como UTF-8, es recomendable utilizar algún procedimiento de conversión de formatos que garantice esta codificación de caracteres. Además de Excel (a partir de la versión 16), herramientas útiles para realizar esta tarea son Google Spreadsheet y LibreOffice Calc.

CSV

06 Otras consideraciones sobre archivos CSV

- El tipo MIME más apropiado para denotar el tipo de contenido de los archivos CSV transmitidos por Internet se indica mediante el siguiente valor de la propiedad *Content-Type*:

Content-Type: text/csv

- También, aunque menos habitual, se puede expresar utilizando:

Content-Type: application/octet-stream

Content-Type: text/comma-separated-values

- Como se ha indicado, es recomendable codificar los archivos CSV utilizando UTF-8, si se utiliza otro tipo de *encoding* puede especificarse a través del parámetro *charset* en la cabecera Content-Type. Por ejemplo:

Content-Type: text/csv;charset=ISO-8859-1

- Si un archivo CSV no contiene línea de cabecera es posible indicarlo usando el parámetro *header* del tipo MIME. Por ejemplo:

Content-Type: text/csv;header=absent

- La implementación de estas propiedades ayudan a procesar los archivos CSV automáticamente y de forma consistente.

07 Toolbox para archivos CSV

- Existen herramientas para el tratamiento de archivos CSV que aportan diferentes funcionalidades.
- Son herramientas gratuitas que normalmente se ofrecen como servicio, es decir, están disponibles online y no requieren la instalación de software o en su caso, ésta es mínima.
- Por lo general, son capaces de validar estructuras de datos o lo que es lo mismo, la consistencia entre el número de campos de cabecera y los existentes en cada una de las filas de datos además de la consistencia de los tipos y valores de los datos de cada celda.
- Destacan por su facilidad de uso y funcionalidad:

	Servicio / Suite de herramientas	Validación de estructura	Limpieza De datos	Conversión de formatos
<u>CSVLint</u>	On line	✓	✗	✗
<u>Goodtables</u>	On line	✓	✗	✗
<u>Data Curator</u>	Suite	✓	✓	✓
<u>CSVkit</u>	Suite	✗	✓	✓

Herramienta: CSVlint

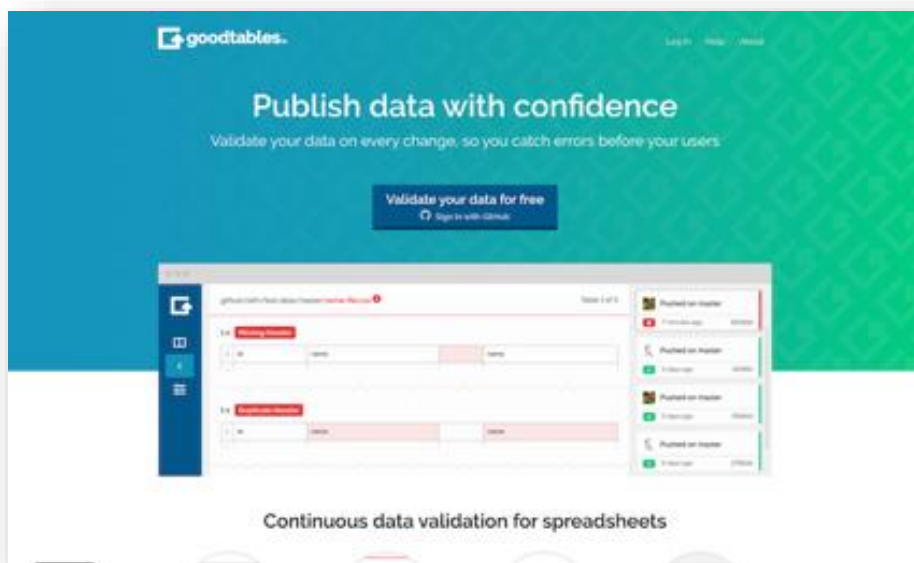
- **CSVlint** es una herramienta disponible online gestionada por [Open Data Institute](https://open-data-institute.org/) para verificar que un archivo CSV es perfectamente legible y está bien formado, es decir, valida si contiene las columnas y los tipos de valores que debería. Permite validar archivos y esquemas de tablas de datos en CSV.
- El análisis se realiza sobre archivos subidos directamente a **CSVlint** o disponibles online.
- El análisis devuelve información sobre errores, necesariamente corregibles para usar los datos, advertencias, cuya subsanación ayuda a los usuarios de los datos y mensajes informativos sobre determinados consejos y sugerencias para facilitar el uso de los datos.
- La herramienta genera un distintivo que es posible embeber en el sitio web del propietario de archivo CSV usando el correspondiente código HTML.
- Algunos errores de codificación de caracteres son corregidos automáticamente, generando una nueva versión estandarizada del archivo CSV original.
- **CSVlint** guarda un registro de las validaciones y esquemas recientemente utilizados que puede ser útil para identificar errores comunes.



Disponible en: <https://csvlint.io/>

Herramienta: Goodtables

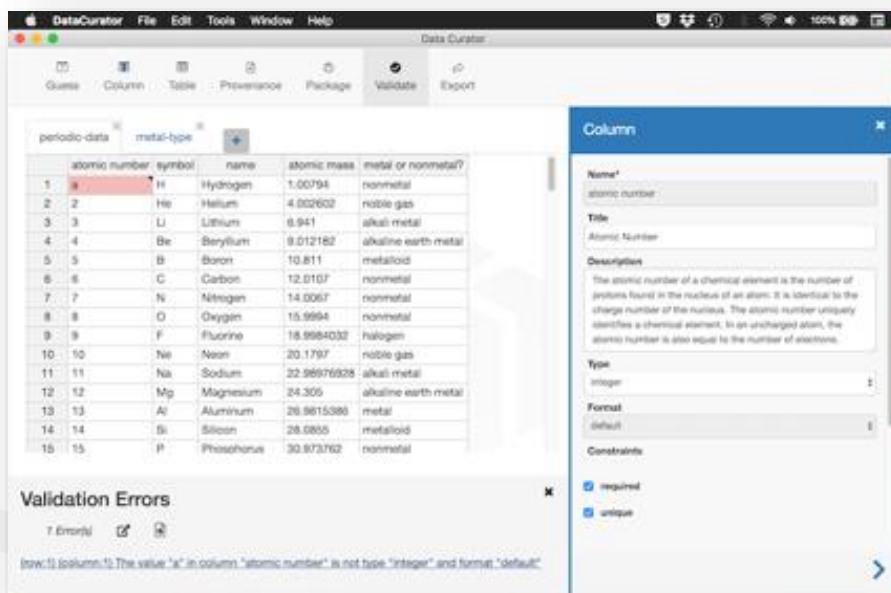
- **Goodtables** es un servicio de validación continua para datos tabulares de [Open Knowledge Foundation](#) (OKI) que permite la verificación de errores y problemas comunes en archivos de datos tabulares.
- Como servicio de validación permite que Goodtables pueda ser integrado en diferentes flujos de publicación de Datos Abiertos.
- Valida tanto datos tabulares en diferentes formatos: CSV, Microsoft Excel o LibreOffice Calc, como esquemas en formato JSON
- Permite, además, la validación directa de datos tabulares albergados sobre Github o Amazon S3.
- Soporta *Tabular Data Package* que es un formato simple para publicar y compartir datos tabulares promovido desde el proyecto [FrictionLess Data](#) de OKI que combina datos archivados como CSV, esquemas de tablas y metadatos según la especificación [Data Package](#).
- **Goodtables** está disponible como librería de Python de tal forma que pueda ser invocada para cargar y validar tablas de datos.



Disponible en: <https://goodtables.io/>

Herramienta: Data Curator

- **Data Curator** es una herramienta de escritorio implementada por [Open Data Institute](#) que permite la edición, validación y publicación de archivos de datos tabulares reutilizables como Datos Abiertos.
- Con esta herramienta es posible generar datos tabulares (CSV, TSV, entre otros), partiendo de cero o a partir de plantillas de estructuras de datos y esquemas.
- Automáticamente corrige problemas comunes encontrados en archivos CSV y Excel.
- Es posible, crear de forma automática, esquemas que describen los campos de datos e incluir reglas específicas de validación (por ejemplo, valores únicos, obligatorios, de longitud mínima o máxima, o sujetos a expresiones regulares), al igual que permite describir la procedencia de los datos.
- La herramienta valida el archivo de datos contra el esquema definido y genera archivos de valores separados en los diversos dialectos CSV (comas, puntos y comas, tabuladores o campos de ancho fijo).
- Permite encapsular datos y esquema para exportar utilizando la especificación Data Package.
- Además, los archivos de datos descritos y validados pueden ser publicados directamente sobre portales de datos CKAN.



Disponible en: <https://github.com/ODIQueensland/data-curator>

Herramienta: Csvkit

- **CSVkit** son un conjunto de herramientas para usar desde línea de comandos en entornos Linux / MacOS que permite convertir y trabajar con archivos CSV.
- Entre otras funcionalidades útiles, **CSVKit** permite: convertir archivos Excel o JSON a CSV y viceversa; realizar diferentes operaciones a nivel de columna, fila o celda; generar sumarios estadísticos y realizar consultas SQL sobre los datos.
- Además realiza determinados análisis de los datos e infiere algunas de sus características como la ausencia / presencia de cabecera o tipos de datos.



Disponible en: <https://csvkit.readthedocs.io/en/latest/>

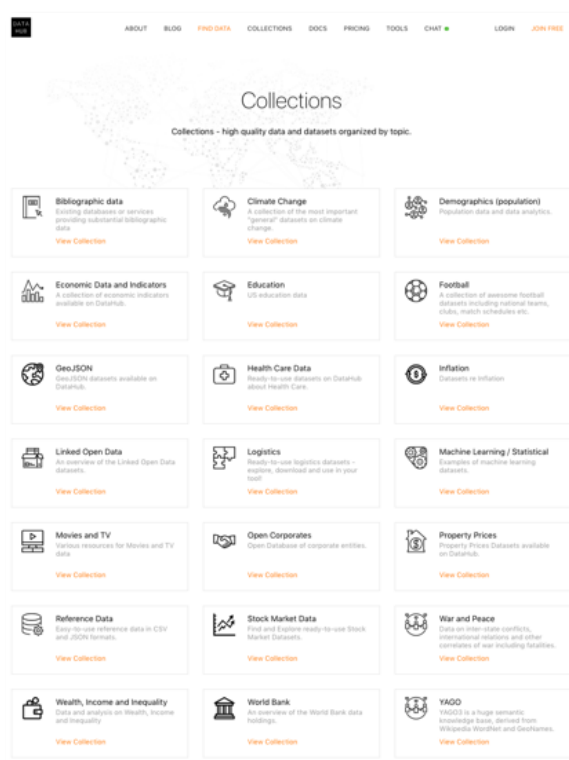
08 Dónde encontrar buenos ejemplos de datos en formato CSV

- Se incluye a continuación la descripción de dos iniciativas que publican conjuntos de Datos Abiertos destacables por la alta calidad de los datasets disponibles en formato CSV:
 - [Datahub.io](https://datahub.io)
 - [Kaggle.com](https://kaggle.com)
- En estas plataformas cobra especial trascendencia la calidad en la forma que se publican los datos, dado el sólido entendimiento sobre la naturaleza multipropósito del uso de los datos y el tratamiento profesional de los mismos, que se encuentra en los fundamentos de ambas iniciativas.
- Datahub, destaca por implementar un soporte integral para transformar, validar y publicar datos de calidad y Kaggle constituye uno de los repositorios de conjuntos de datos y conocimiento en torno al análisis de datos de referencia profesional más importantes.
- Ambas iniciativas constituyen ejemplos reales de buenas prácticas en el tratamiento de archivos CSV que pueden ser tenidas en cuenta para abordar procesos de preparación y publicación de Datos Abiertos.

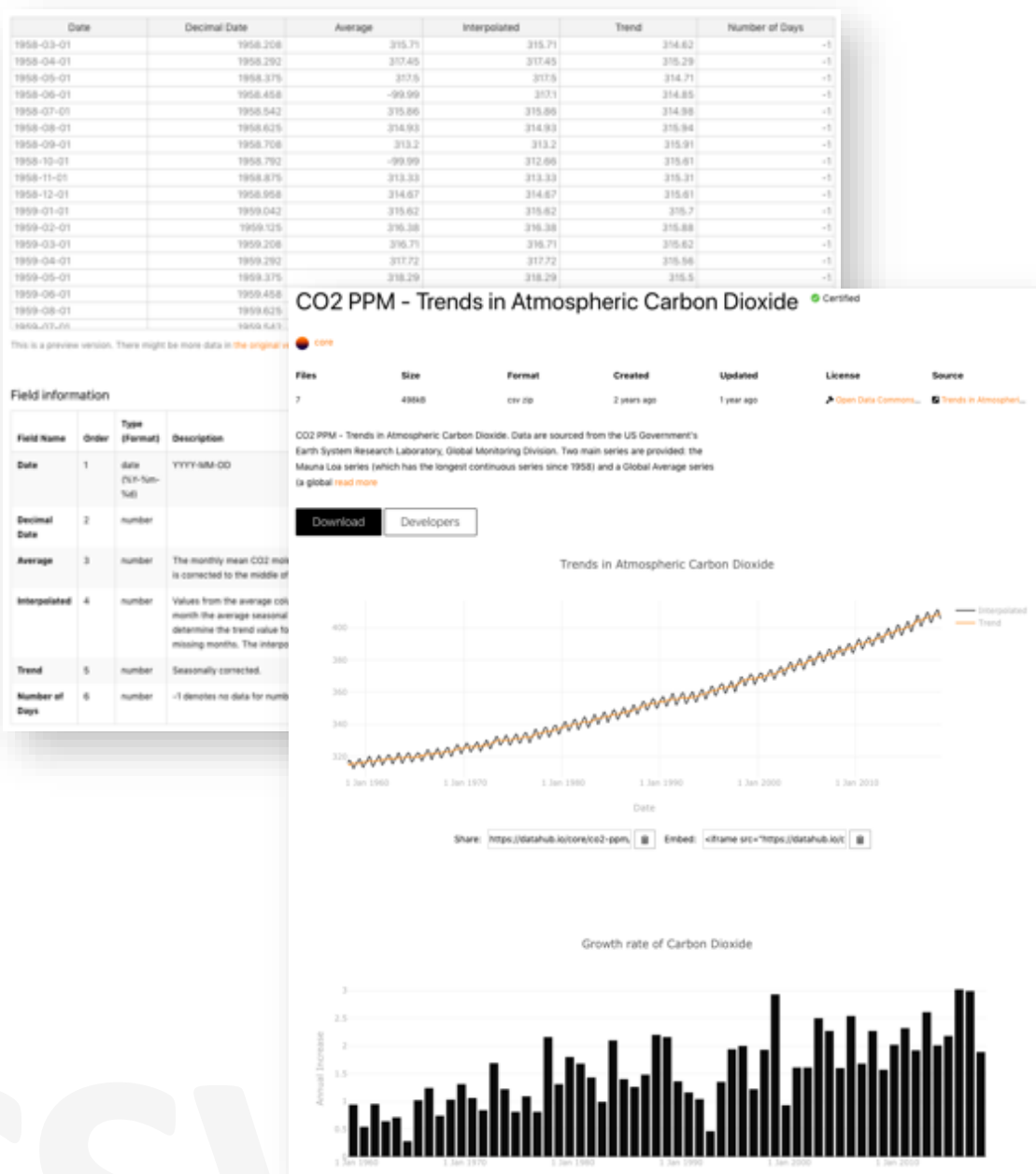
CSV

datahub.io

- [Datahub](#) es una plataforma web que da soporte integral a flujos de trabajo para la preparación y publicación de Datos Abiertos. Está diseñada para preparar, catalogar y publicar datos de alta calidad haciendo uso del conjunto de [herramientas Frictionless Data](#).
- El conjunto de herramientas Frictionless Data, es una colección de especificaciones y aplicaciones para la preparación de archivos de datos, entre las que se encuentra “Goodtables”, descrita en el apartado “Toolbox para archivos CSV” de esta guía.
- Datahub, contiene colecciones de datos compatibles con Datos Abiertos de alto valor como: cambio climático, datos económicos e indicadores, estadística, logística, registros empresariales con origen en fuentes oficiales.
- Cada entrada de datos disponible contiene un conjunto de elementos para mostrar las propiedades del dataset (esquema y recursos de datos), opciones de descarga de datos en diversos formatos, entre ellos CSV, vistas de las tablas de datos y visualizaciones simples.
- Proporciona, además, acceso directo para importar los datos utilizando una variedad de herramientas habituales en el contexto profesional: R, Python, JavaScript y SQL.



- Un ejemplo de [dataset](#) en formato CSV disponible en la plataforma es el que muestra la tendencia de Dióxido de Carbono en la atmósfera, con origen en el “US Government's Earth System Research Laboratory”.



- El archivo CSV descargable del dataset "*CO2 PPM - Trends in Atmospheric Carbon Dioxide*" posee las siguientes características:

- Diccionario de datos procesable en formato JSON según la especificación Data Package.
- Fila única de cabecera.
- Registro único por fila.
- Comprensible nombrado de columnas.
- Estructura de datos vertical.
- Tratamiento de valores desconocidos, indicados por valores del tipo -99.99 (para el atributo '*average*') y -1 (para el atributo '*días*').
- No contiene totales ni agrupaciones.
- Correcto tipado de campos.
- Campo de fechas codificado siguiendo el estándar ISO-8601.
- No contiene datos con coordenadas geográficas o campos codificados.

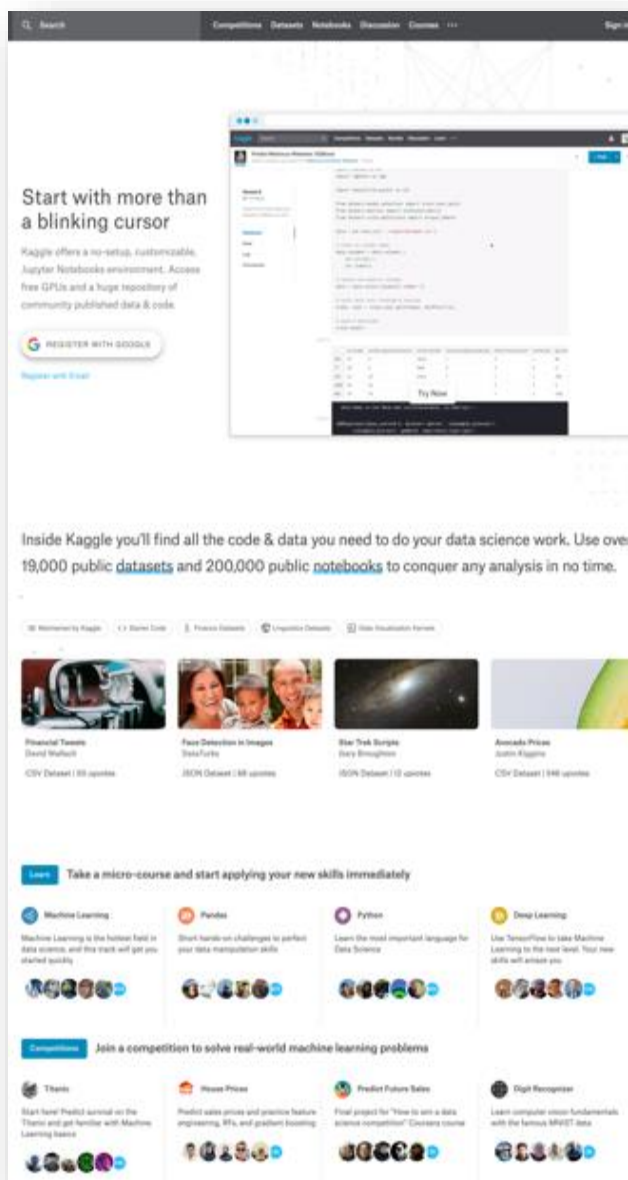
```
Date,Decimal Date,Average,Interpolated
,Trend,Number of Days
1958-03-01,1958.208,315.71,315.71,314.62,-1
1958-04-01,1958.292,317.45,317.45,315.29,-1
1958-05-01,1958.375,317.50,317.50,314.71,-1
1958-06-01,1958.458,-99.99,317.10,314.85,-1
1958-07-01,1958.542,315.86,315.86,314.98,-1
1958-08-01,1958.625,314.93,314.93,315.94,-1
1958-09-01,1958.708,313.20,313.20,315.91,-1
1958-10-01,1958.792,-99.99,312.66,315.61,-1
1958-11-01,1958.875,313.33,313.33,315.31,-1
1958-12-01,1958.958,314.67,314.67,315.61,-1
1959-01-01,1959.042,315.62,315.62,315.70,-1
1959-02-01,1959.125,316.38,316.38,315.88,-1
1959-03-01,1959.208,316.71,316.71,315.62,-1
1959-04-01,1959.292,317.72,317.72,315.56,-1
1959-05-01,1959.375,318.29,318.29,315.50,-1
1959-06-01,1959.458,318.15,318.15,315.92,-1
1959-07-01,1959.542,316.54,316.54,315.66,-1
1959-08-01,1959.625,314.80,314.80,315.81,-1
1959-09-01,1959.708,313.84,313.84,316.55,-1
1959-10-01,1959.792,313.26,313.26,316.19,-1
1959-11-01,1959.875,314.80,314.80,316.78,-1
1959-12-01,1959.958,315.58,315.58,316.52,-1
...
2014-12-01,2014.958,398.91,398.91,399.64,29
2015-01-01,2015.042,399.98,399.98,399.69,30
2015-02-01,2015.125,400.28,400.28,399.51,27
2015-03-01,2015.208,401.54,401.54,400.05,24
2015-04-01,2015.292,403.28,403.28,400.49,27
2015-05-01,2015.375,403.96,403.96,400.63,30
2015-06-01,2015.458,402.80,402.80,400.50,28
2015-07-01,2015.542,401.31,401.31,400.92,23
2015-08-01,2015.625,398.93,398.93,400.84,28
2015-09-01,2015.708,397.63,397.63,401.15,25
2015-10-01,2015.792,398.29,398.29,401.59,28
```

08

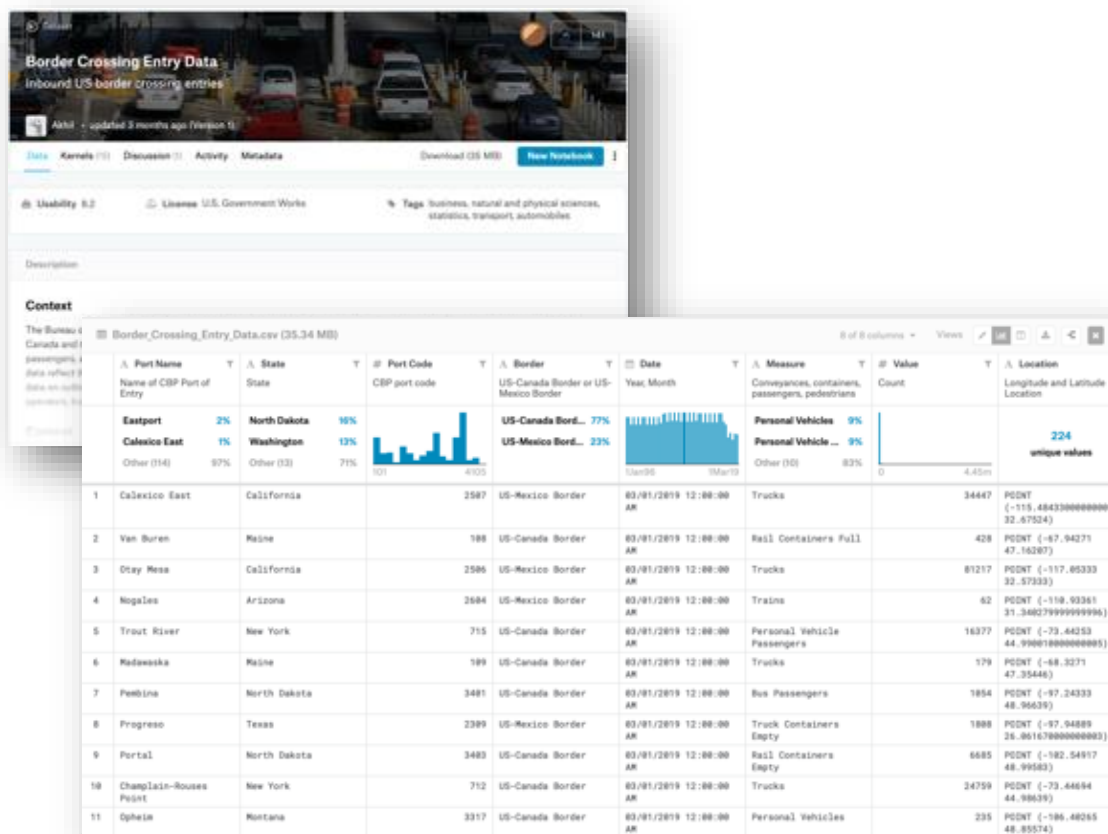
CSV

/// kaggle.com

- [Kaggle](#) es una plataforma web donde empresas e instituciones proponen problemas estratégicos o de negocio para que grupos con interés o experiencia en análisis de datos compitan, de forma remunerada, para crear y proponer las mejores soluciones.
- En la plataforma se presenta cualquier tipo de problema a resolver que pueda encontrarse en distintos ámbitos del mundo real, tales como servicios financieros, energía, sociedad, tecnología de la información, etc.
- Además de las competiciones propuestas y la disponibilidad de recursos de código para analizar cualquier conjunto de datos, Kaggle alberga cientos de conjuntos de datos de todo tipo y tamaño que se pueden descargar y usar de forma gratuita, mayoritariamente en formato CSV.
- Cada conjunto de datos contiene una descripción detallada de su contenido y dispone de una vista previa del contenido del archivo mediante un explorador de datos.



- Los archivos incluyen la especificación del esquema de datos asociado incluyendo descripciones de columna con sus correspondientes metadatos: tipos de datos, rango de valores, unidades, histograma, etc. Las descripciones se asignan a nivel de tabla y a nivel de columna individual facilitando su comprensión.
- Cada dataset, además, posee un valoración otorgada por la comunidad de usuarios sobre su grado de usabilidad en base a aspectos como la facilidad para entender el contenido mediante la calidad de descriptores esenciales: de archivo, de columnas, etiquetas, licencia, procedencia, frecuencia de actualización etc.
- Un **ejemplo** es el [dataset](#) que contiene los datos de entrada del cruce de fronteras del US Bureau of Transportation Statistics (BTS).



- El archivo CSV descargable del dataset "*Border crossing entry data*" posee las siguientes características:
 - Diccionario de datos especificado en forma de tabla pero no procesable por máquina. Un aspecto de mejora es su disponibilidad en formato JSON enlazado desde el sitio web de descarga.
 - Fila única de cabecera.
 - Registro único por fila.
 - Comprensible nombrado de columnas.
 - Estructura de datos vertical.
 - No contiene totales ni agrupaciones.
 - Correcto tipado de campos.
 - Campo de fechas codificado siguiendo el estándar ISO-8601.
 - Codificación de códigos postales.
 - Coordenadas geográficas en grados decimales indicadas como punto geográfico de latitud-longitud.

```

Port Name,State,Port Code,Border,Date,Measure,Value,Location
Calxico East,California,2507,US-Mexico Border,03/01/2019 12:00:00 AM,Trucks,34447,POINT (-
115.484330000000001 32.67524)
Van Buren,Maine,108,US-Canada Border,03/01/2019 12:00:00 AM,Rail Containers Full,428,POINT (-
67.94271 47.16207)
Otay Mesa,California,2506,US-Mexico Border,03/01/2019 12:00:00 AM,Trucks,81217,POINT (-
117.05333 32.57333)
Nogales,Arizona,2604,US-Mexico Border,03/01/2019 12:00:00 AM,Trains,62,POINT (-110.93361
31.3402799999999996)
Trout River,New York,715,US-Canada Border,03/01/2019 12:00:00 AM,Personal Vehicle
Passengers,16377,POINT (-73.44253 44.9900100000000005)
Madawaska,Maine,109,US-Canada Border,03/01/2019 12:00:00 AM,Trucks,179,POINT (-68.3271
47.35446)
Pembina,North Dakota,3401,US-Canada Border,03/01/2019 12:00:00 AM,Bus
Passengers,1054,POINT (-97.24333 48.96639)
Progreso,Texas,2309,US-Mexico Border,03/01/2019 12:00:00 AM,Truck Containers
Empty,1808,POINT (-97.94889 26.0616700000000003)
Portal,North Dakota,3403,US-Canada Border,03/01/2019 12:00:00 AM,Rail Containers
Empty,6685,POINT (-102.54917 48.99583)
Champlain-Rouses Point,New York,712,US-Canada Border,03/01/2019 12:00:00
AM,Trucks,24759,POINT (-73.44694 44.98639)
Opheim,Montana,3317,US-Canada Border,03/01/2019 12:00:00 AM,Personal Vehicles,235,POINT (-
106.40265 48.85574)

```

09 Enlaces de interés

La guía para el tratamiento de archivos CSV que se describe en este documento se construye en base a las aportaciones de los siguientes documentos de referencia:

- Especificación del formato CSV: <https://tools.ietf.org/html/rfc4180>
- Especificación del formato TSV: <https://www.iana.org/assignments/media-types/text/tab-separated-values>
- Grupo de trabajo de W3C sobre CSV en la Web: https://www.w3.org/2013/csvw/wiki/Main_Page
- Modelo de datos tabulares de W3C: <https://www.w3.org/TR/2015/REC-tabular-data-model-20151217/>
- Norma Técnica de Interoperabilidad de Reutilización de Recursos de Información: <https://datos.gob.es/es/documentacion/norma-tecnica-de-interoperabilidad-de-reutilizacion-de-recursos-de-informacion>
- Datos abiertos: guía estratégica para su puesta en marcha y conjuntos de datos mínimos a publicar, de la FEMP: <https://datos.gob.es/es/documentacion/datos-abiertos-guia-estrategica-para-su-puesta-en-marcha-y-conjuntos-de-datos-0> y Datos abiertos FEMP 2019: 40 conjuntos de datos a publicar por las Entidades Locales: [http://femp.femp.es/files/3580-1937-fichero/DATOS ABIERTOS FEMP 2019.pdf](http://femp.femp.es/files/3580-1937-fichero/DATOS_ABIERTOS_FEMP_2019.pdf)
- Guía de datos “New York State Open Data dataset submission guide”: <https://data.ny.gov/download/c3zp-wr9j/application/pdf>
- Guía para la publicación de datos en formatos abiertos de Argentina: <https://datosgobar.github.io/paquete-apertura-datos/guia-abiertos/#guia-para-la-publicacion-de-datos-en-formatos-abiertos>

Anexo I:

Taxonomías y listas de códigos de uso común

En este anexo se recoge una serie de referencias a esquemas de conceptos, listas de códigos (code lists) y taxonomías de términos que se pueden utilizar como valores prescritos para propiedades o atributos de datos tabulares.

- Clasificaciones y estándares armonizados a nivel nacional e internacional del INE: <http://www.ine.es/ss/Satellite?L=0&c=Page&cid=1254735839296&p=1254735839296&pagename=MetodologiaYEstandares%2FINELayout>
- Tesoro multilingüe de la Unión Europea (EUROVOC): <http://open-data.europa.eu/cs/data/dataset/eurovoc>
- Nomenclatura de actividades económicas (NACE): https://ec.europa.eu/eurostat/statistics-explained/index.php/NACE_background
- Nomenclatura de las unidades territoriales estadísticas (NUTS): <https://ec.europa.eu/eurostat/web/nuts/background>
- ISO 639-2 - códigos de idioma: http://www.loc.gov/standards/iso639-2/php/code_list.php
- ISO 4217 - códigos de moneda: http://www.iso.org/iso/home/standards/currency_codes.htm
- ISO 3166-1 - códigos de país: http://www.iso.org/iso/country_codes.htm
- Unidades de medida UN / CEFAC: http://www.unece.org/fileadmin/DAM/cefact/recommendations/rec20/rec20_rev3_Annex3e.pdf