ANEXO II

# GUÍA PARA LA PUBLICACIÓN DE DATOS EN FORMATOS ABIERTOS

La última versión de este documento se encuentra en https://datosgobar.github.io/paquete-apertura-datos/guia-abiertos/.

## INDICE



## INTRODUCCIÓN

Esta guía busca ayudar a los organismos a instrumentar la Política de Datos Abiertos impulsada desde el Gobierno de la Nación Argentina, a través del Decreto N° 117/2016 del 12 de enero de 2016.

## OBJETIVO DE LA GUÍA

Esta es **una guía de buenas prácticas para la publicación de datos en formatos abiertos.**

Estas recomendaciones se basan en:

- Estándares usados a nivel nacional e internacional.
- La experiencia de trabajo del equipo de la Dirección de Datos Públicos de la Jefatura de Gabinete de Ministros de la Nación.

Esta es **una guía colaborativa y en progreso** . Valoramos, y alentamos, a organizaciones y ciudadanos a plantear ideas, sugerencias, y comentarios que nos ayuden a crear un mejor documento.

El documento se estructura así:

- **Formatos abiertos de archivos:** cuáles son los formatos más usuales en los que se publican datos y cuáles son los más recomendables.
- **Fragmentación de archivos** : cuáles son los criterios para decidir que un archivo es demasiado grande (y hay que fragmentarlo) o demasiado chico (y hay que juntarlo con otros).
- **Nomenclatura de archivos** : cómo nombrar adecuadamente un archivo.
- **Codificación** : cuál es la codificación en que se debe guardar un archivo.
- **Estructura y características de los datos tabulares**
  - Recomendaciones generales: aplican a todos los casos.
  - Recomendaciones para estructurar planillas de cálculo: aplican exclusivamente al trabajo en planillas de cálculo.
  - Exportación a CSV: cómo exportar adecuadamente planillas de cálculo a un archivo de formato abierto.
- **Estándares según el tipo de datos** : cuáles son los estándares recomendados para manejar distintos tipos de datos.

Estos son los primeros aspectos importantes para la estandarización de datos.

Para una discusión sobre los estándares recomendados en el manejo de datos básicos y fundamentales, transversales a distintas áreas temáticas, se puede consultar la **Guía para la identificación y uso de entidades interoperables** .

# FORMATOS ABIERTOS DE ARCHIVOS

Hay una gran variedad de tecnologías disponibles para producir y almacenar datos. Como ser: planillas de cálculo, bases de datos, software estadístico más específico y más. Esto genera una enorme diversidad de formatos, a veces caótica.

Algunos de estos formatos, no siempre se adecuan a los niveles de apertura deseados. Te ofrecemos algunas pautas y recomendaciones que facilitan la adaptación y/o transformación de estos formatos hacia otros más abiertos y fácilmente reutilizables.

En este cuadro consideramos algunos de los formatos más usados y evaluamos su nivel de apertura:

| Formato | Descripción breve | Tipo de datos | Nivel de apertura |
|---|---|---|---|
| PDF | Los PDF son archivos de texto que no se encuentran en formato estructurado. Se utilizan para la generación de documentos, no para publicar o almacenar datos. | Texto | Muy bajo |
| XLS | Los XLS son archivos de planilla de cálculo. Es un formato propietario de Microsoft. | Tabulares | Bajo |
| XLSX | Los XLSX también son archivos de planilla de cálculo cuyo formato fue desarrollado por Microsoft pero su especificación es abierta (ISO/IEC 29500:2008). Es el formato por defecto del Excel 2007 en adelante. | Tabulares | Medio |
| ODS | Los ODS son archivos con la estructura de un XML. Es un formato abierto basado en OASIS OpenDocument Format (ISO/IEC 26300). Es el formato por defecto del procesador de planillas de cálculo Open Office. | Tabulares | Medio |
| CSV | Los archivos CSV son archivos de texto plano donde las columnas se separan por comas y las filas por saltos de línea. Es un formato abierto. | Tabulares | Alto |
| JSON | Es un formato para el intercambio de datos entre sistemas. Es un formato abierto no tabular basado en la especificación RFC 7159. | Estructurados | Alto |
| SHP | ESRI Shapefile (SHP) es un formato propietario de datos espaciales desarrollado por ESRI, quien crea y comercializa software para Sistemas de Información Geográfica. Actualmente se ha convertido en formato estándar de facto para el intercambio de información geográfica entre SIG. | Geográficos | Medio |
| KML | Es un formato abierto para datos geográficos basado en el estándar XML. | Geográficos | Alto |
| GEOJSON | Es un formato estándar abierto diseñado para representar elementos geográficos sencillos, junto con sus atributos no espaciales. | Geográficos | Alto |
| GEOPACKAGE | Es un formato de datos geoespaciales implementado como un contenedor de base de datos SQLite. | Geográficos | Alto |

Antes de seguir, introduciremos dos conceptos que se usarán a lo largo de toda la guía:

- **Distribución o Recurso:** Una distribución o recurso es la unidad mínima en la que se publican datos. Se trata de los archivos que pueden ser descargados y re-utilizados por un usuario. Los recursos pueden tener diversos formatos (.csv, .shp, etc.).
- **Dataset:** Un conjunto de datos o dataset agrupa recursos referidos a un mismo tema que respetan una estructura de la información. Los recursos que lo componen pueden diferir en el formato en que se los presenta (por ejemplo: .csv, .json, .xls, etc.), la fecha a la que se refieren, el área geográfica cubierta, ser tablas de un mismo esquema de base de datos relacional o estar separados bajo algún otro criterio.

Un recurso en formato tabular es un archivo plano que se ajusta a un esquema predefinido de columnas, incluyendo el nombre de la columna y el tipo de datos.

En la mayoría de los casos, corresponde a datos que llegan de bases de datos, reportes y planillas de cálculo en general. A diferencia de los formatos tabulares, los archivos JSON siguen una estructura diferente donde se definen listas de objetos con pares "clave" - “valor”.

**Recomendamos con énfasis la publicación de los datos en formato CSV y/o JSON** . En caso de utilizar formatos propietarios o aún no estandarizados, es útil indicar software, versión y aplicación que permite procesar esos formatos.

## CSV

El CSV es un formato estándar de archivo de texto plano donde:

- Los campos (columnas) se separan por comas `,` .
- Los registros (filas) se separan por saltos de línea.
- Los números decimales utilizan `.` para separar la parte entera de la parte

decimal.

- Se utilizan las comillas dobles `"` como caracter de entrecomillado. Los valores en tablas CSV que incluyen dentro de sí caracteres especiales como `,` o `"`, deben estar encerrados entre `"` para su correcta interpretación.

Algunas versiones alternativas de esta forma de publicar datos usan otros separadores como punto y coma ( `;` ) o *pipe* ( `|` ), pero la recomendación para toda la Administración Pública Nacional se basa en la versión de CSV más estándar, indicada por la especificación RFC4180 y las pautas de la W3C .

Otros elementos a tener en cuenta:

- La primera fila siempre contiene los nombres de los campos.
- No se deben repetir nombres entre los campos.
- No se debe colocar espacios al principio ni al final del nombre de un campo, o de un valor.
- Tanto los campos como los valores deben estar separados por comas ( `,` ).
- En el caso de que un valor contenga el caracter separador ( `,` ) o cualquiera de los caracteres que separan las líneas ( `\n` ), el valor completo debe ser encerrado entre comillas dobles `""` . Esto indica que el caracter no cumple el rol de separar columnas o filas, sino que es parte de un valor.

Ejemplo:

```
col1,col2\n
"La tasa de Juan, está vacía",La tasa de Pablo está llena\n
"La tasa de Juan\nestá vacía",La tasa de Pablo está llena\n
"La tasa de Juan\nestá vacía",La tasa de Pablo está llena\n
```

*Nota: Posiblemente te preguntes qué es el caracter '/n'. En la mayoría de los casos no te vas a enterar que existe ya que usás "Enter" en tu editor de texto y este por detrás de escena aplica un '/n'. No te preocupes por esto, a menos que lo veas en algún lado. Usualmente, el ejemplo anterior lo vas a ver como:*

```
col1,col2
"La tasa de Juan, está vacía",La tasa de Pablo está llena
"La tasa de Juan está vacía",La tasa de Pablo está llena
"La tasa de Juan está vacía",La tasa de Pablo está llena
```

- En el caso de que un valor contenga el caracter comilla doble ( `"` ), el valor debe ser encerrado entre comillas dobles como en el caso anterior ( `""` ) y, además, los caracteres comilla doble que se encuentren dentro del valor deben escribirse dos veces ( `""` ).

Ejemplo:

```
col1,col2
"La tasa de ""Juan"" está vacía",La tasa de Pablo está llena
```

- Para todos los tipos de datos se considera válido el valor indefinido. Este se expresará con la ausencia de todo caracter y no con un caracter o string especial como podrían ser ".", "null", "none", "nan", "SD", "S/D", etc.

Ejemplo:

```
col1,col2,col3
```

```
a,,b
a,"",b
```

Cabe destacar que un archivo CSV puede leerse desde una planilla de cálculo como MS Office o similar, donde los campos se separarán en columnas independientes a través de (,). Para hacerlo en Microsoft Excel, se debe ir `Archivo > Abrir > Archivos de texto` y allí usar las comas (,) como separadores de columnas.

### JSON

JSON es un formato de texto popular para el intercambio de datos, es un acrónimo de JavaScript Object Notation. Por su característica de ser un formato de tipo estructurado es especialmente útil para el intercambio de datos entre sistemas ( *machine readable format* ).

El formato JSON ha sido definido por la especificación RFC 7159 y, tal como CSV, también es un estándar abierto.

## FRAGMENTACIÓN DE ARCHIVOS

Para garantizar la accesibilidad a los datos, **puede ser necesario fragmentar los archivos excesivamente grandes** , que superen el millón de filas.

Sin embargo, **se recomienda no fragmentarlos (manteniendo una única versión completa) si los archivos tienen menos de 500 mil filas** y su actualización en los próximos años no los acercaría al millón de filas.

Si fragmentar resulta necesario, recomendamos usar conceptos simples:

1. **por períodos** en caso de tratarse de información temporal (Ej. Años, semestres, trimestres, meses, semanas, días),
2. **por zonas** en caso de tratarse de información geográfica (Ej. provincias, municipios, barrios, secciones, o manzanas) o
3. **por dimensiones temáticas** propias del dominio particular de la información.

Sin embargo, siempre que se decida fragmentar un archivo para garantizar su accesibilidad, **recomendamos publicar también una versión no fragmentada que contenga el conjunto de datos completo** (incluso aunque sea muy grande), a los fines de evitar la tarea de consolidación a los usuarios que requieren usar todos los datos.

**Cuando la versión completa resulta muy pesada (más de 50MB) se recomienda comprimir el archivo** usando protocolos de compresión sin pérdida, y abiertos.

## NOMENCLATURA DE ARCHIVOS

Nombrar bien los archivos es muy importante. Un buen nombre de archivo es claro respecto del contenido, es fácil de leer por cualquier software o terminal, y es resistente al paso del tiempo y las modificaciones que pueda sufrir el contenido.

Recomendamos estas convenciones para nombrar archivos:

- Usar palabras siempre en minúsculas.
- No usar artículos ni preposiciones.
- Usar únicamente letras y números ASCII, siempre en minúsculas, comprendidos en el rango "a-z" y “0-9”.

- Separar las palabras con guión medio "-".
- En caso de corresponder, ubicar la referencia temporal o del atributo de fragmentación siempre al final.

Ejemplos:

- *acceso-informacion-publica.csv* : Versión completa del recurso.
- *acceso-informacion-publica-2013.csv* : Versión del recurso fragmentada por año.
- *acceso-informacion-publica-201302.csv* : Versión del recurso fragmentada por mes.
- *acceso-informacion-publica-caba.csv* : Versión del recurso fragmentada por división político-territorial (provincia o caba).
- *acceso-informacion-publica-caba-2013.csv* : Versión del recurso fragmentada por división político-territorial (provincia o caba) y año.
- *acceso-informacion-publica-jujuy-20130208.csv* : Versión del recurso fragmentada por división político-territorial (provincia o caba) y fecha.

Para la fragmentación temporal, recomendamos el estándar de los ejemplos, ya que es compacto y ordena los recursos por tiempo: YYYYMMDD. Por favor, recordá mantener siempre dos dígitos para el mes y el día, incluso si el número es menor a 10.

Para la fragmentación por zonas, consultá la **Guía para la identificación y uso de entidades interoperables** , y mirá cómo nombrarlas adecuadamente.

En el caso de usar dimensiones temáticas propias del dominio particular de la información, podés ver esa guía o usar el mejor estándar identificado para esa temática particular.

# CODIFICACIÓN

Todos los recursos de datos, incluyendo los geográficos, deben publicarse usando la codificación UTF-8 siguiendo las recomendaciones de la W3C .

Una de las principales razones es que UTF-8 soporta una gran variedad de lenguajes. Según la W3C es un *"estándar en el que se definen todos los caracteres necesarios para la escritura de la mayoría de los idiomas hablados en la actualidad. Su objetivo es ser -y en gran medida ya lo ha logrado- un superconjunto de todos los sets de caracteres que se hayan codificado"* .

# ESTRUCTURA Y CARACTERÍSTICAS DE LOS DATOS TABULARES

En esta sección veremos:

- Recomendaciones generales para el trabajo con datos.
- Recomendaciones para el trabajo con planillas de cálculo, orientadas tanto a facilitar su exportación a formatos abiertos, como a su propia usabilidad en el contexto de cualquier aplicación de planillas de cálculo.

## RECOMENDACIONES GENERALES

Estas son recomendaciones generales para el trabajo con datos tabulares.

Sugerimos adoptarlas sea cual sea la tecnología usada.

Muchas de las recomendaciones aquí presentadas se encuadran en los principios de Tidy Data delineados por Hadley Wickham . Éstos establecen, por ejemplo, que en una tabla de datos *"cada variable es una columna, cada observación es una fila, y cada tipo de unidad observacional es una tabla"* . Sugerimos complementar la lectura de esta guía con la del trabajo mencionado.

## Nomenclatura de los campos (nombres de las columnas)

La "nomenclatura de los campos" es el nombre de las columnas en los datos de estructura tabular. Nombrar bien los campos es extremadamente importante (tal vez incluso más que nombrar bien los archivos!) porque **permiten al usuario entender qué información hay en esos campos y cómo usarla adecuadamente** . Los nombres codificados, poco claros o ambiguos pueden generar graves errores en el uso de la información o incluso la imposibilidad de usarla.

Las siguientes recomendaciones aplican a la generalidad de los casos. Cuando existan convenciones particulares según la temática o rubro de datos de que se trate -y éstas entraran en conflicto- puede ser conveniente privilegiar primero la convención de la temática específica y luego la convención general.

Los nombres de los campos deben:

- Estar en español.
- Ser lo más explícitos, descriptivos y declarativos como sea posible.
    - Es preferible que el nombre de un campo sea claro antes que corto, pero se recomienda no superar los 50 caracteres (en todos los casos, **el nombre de un campo no debería superar los 60 caracteres** que es el límite de varias aplicaciones de software o bases de datos como PostGres).
    - No usar abreviaturas si no es estrictamente necesario o recomendado por una convención ampliamente difundida. En caso de usarlas, incluir su explicación en la descripción de la columna documentada.
- Estar en minúsculas, no incluir caracteres especiales, ni estar subrayados.
    - Usar palabras compuestas únicamente de caracteres en minúsculas comprendidos en el rango a-z (letras sin tildes) y en el rango 0-9 (dígitos).
    - Las palabras deben unirse con guión bajo " _ ".
    - No contener espacios.
    - Las palabras deben separarse siempre con " _ ", en lugar de no tener separación alguna: *fecha_audiencia_solicitada* en lugar de *fechaaudienciasolicitada*
- Referirse a un sólo atributo de los datos, indivisible en más de un campo.
    - Los campos deben separar los atributos de los datos en la forma más desagregada que sea posible.
    - Se debe evitar definir campos que contengan más de un tipo de información (por ejemplo: e-mail y sitio web, número de teléfono, etc bajo " *datos_contacto* ").
- Si existe una entidad que engloba varias características separadas en campos diferentes, comenzar nombrando los campos con esa entidad y luego con los atributos más específicos (de lo más general a lo más específico).
    - Ej.: solicitante y solicitante_documento son entidades más generales que se

repiten en varios campos, que corresponden a atributos más específicos.

- solicitante_nombre
- solicitante_cargo
- solicitante_documento_tipo
- solicitante_documento_numero

- Resulta más fácil identificar qué campos están relacionados entre sí porque configuran atributos de una misma entidad, en lugar de parecer campos conceptualmente independientes. Además, el ordenamiento alfabético de los campos los dejaría automáticamente agrupados por su pertenencia a una entidad más importante.
- Incluso cuando la entidad de un atributo parezca evidente (ej.: un dataset llamado "audiencias" donde todos los campos son atributos de la entidad “audiencia”), se recomienda nombrar el campo incluyendo la entidad a la que hace referencia el atributo.

- **No recomendado** : “fecha_hora”
- **Recomendado** : "audiencia_id", “audiencia_fecha_hora”.

Los campos que sean identificadores o códigos, deberán 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.

**Recomendado**

| sucursal_id | sucursal_nombre |
|---|---|
| 1 | Nueva Pompeya |
| 2 | Barracas |
| 3 | La Quiaca |

## Nivel de granularidad de los datos

Por favor, no incluir totales, subtotales ni agrupamientos de datos. Un dataset debe ser consistente en el nivel de granularidad de los datos que contiene. Está bien tener un dataset con la cantidad de convenios firmados por provincia y está bien tener un dataset con la cantidad de convenios firmados por municipio. No está bien tener un dataset que mezcle ambos.

Dicho esto, el dato agregado "convenios firmados por provincia" siempre se puede calcular a partir de un proceso del dataset más desagregado, pero esto no es así a la inversa (es imposible recuperar los datos a nivel de municipio desde el dataset provincial).

**No recomendado** - datos con subtotales y/o totales incluidos (diferentes niveles de granularidad)

| provincia_nombre | municipio_nombre | convenios_firmados_anio | convenios_firmados_numero |
|---|---|---|---|
| | Provincia X | Municipio W | 2011 |
| Provincia X | Municipio X | 2011 | 15 |
| Provincia X | Subtotal | 2011 | 25 |
| Provincia Y | Municipio Z | 2011 | 5 |
| Provincia Y | Subtotal | 2011 | 5 |
| | TOTAL | 2011 | 30 |

**Recomendado** - datos de un mismo nivel de granularidad

| provincia_nombre | municipio_nombre | convenios_firmados_anio | convenios_firmados_numero |
|---|---|---|---|
| Provincia X | Municipio W | 2011 | 10 |
| Provincia X | Municipio X | 2011 | 15 |
| Provincia Y | Municipio Z | 2011 | 5 |

## Usar orientación vertical en lugar de horizontal

Es preferible que la orientación de los datos sea "vertical" en lugar de “horizontal” en los casos en que esto sea posible. La principal razón es que los datos orientados de manera vertical facilitan el tratamiento y análisis de los datos.

**No recomendado** - Orientación horizontal

| municipio_nombre | solicitudes_anio | solicitudes_poda_y_arbolado_numero | solicitudes_recoleccion_residuos_numero |
|---|---|---|---|
| Municipalidad X | 2015 | 340 | 198 |

**Recomendado** - Orientación vertical

| municipio_nombre | solicitudes_anio | solicitudes_categoria | solicitudes_numero |
|---|---|---|---|
| Municipalidad X | 2015 | Poda y arbolado | 340 |
| Municipalidad X | 2015 | Recolección de residuos | 198 |

## Incluir sólo un atributo por campo

Se recomienda definir los campos de forma atómica de modo de incluir un sólo atributo por elemento, en lugar de datos múltiples, generando campos adicionales de ser necesario.

**No recomendado** - múltiples datos en una celda

| municipio_nombre | solicitudes_anio | categoria_solicitudes_numero_y_tipo |
|---|---|---|
| Municipalidad X | 2015 | Poda y arbolado - 340 Recolección de residuos - 198 |

**Recomendado** - un dato por celda

| municipio_nombre | solicitudes_anio | solicitudes_categoria | solicitudes_numero |
|---|---|---|---|
| Municipalidad X | 2015 | Poda y arbolado | 340 |
| Municipalidad X | 2015 | Recolección de residuos | 198 |

## Valores nulos, desconocidos o en blanco en campos numéricos

Los valores de los datos deben ser siempre explícitos y respetando el tipo de datos del campo de que se trate. Los elementos o celdas en blanco se interpretarán siempre como "valor ausente".

Si existen distintas interpretaciones posibles de un "valor ausente", éstas deben ser explicitadas en un campo aparte. Si sólo hay “valores ausentes” (no hay distintas interpretaciones, es siempre un “valor ausente”) no es necesario agregar una

columna adicional.

Es importante destacar, por ejemplo, que cuando un valor numérico sea "0" siempre debe ponerse un “0” como dato (y no un valor nulo, en blanco o vacío).

**No recomendado** - texto presente en campos numéricos

| solicitudes_numero |
|---|
| 34 |
| NA |
| ... |
| |
| 567 |

**Recomendado** - texto excluido de campos numéricos

| solicitudes_numero | solicitudes_numero_missing_nombre |
|---|---|
| 34 | |
| | No disponible |
| | No sabe / No contesta |
| | Valor ausente |
| | Valor ausente |
| 0 | |
| 567 | |

## RECOMENDACIONES PARA ESTRUCTURAR PLANILLAS DE CÁLCULO

Las recomendaciones de esta sección aplican exclusivamente al trabajo sobre planillas de cálculo.

### Usar celdas simples

Recomendamos usar celdas simples y, en ningún caso, combinar celdas.

**No recomendado** - celdas combinadas

| proveedor_nombre | contacto_datos | |
|---|---|---|
| | correo_electronico | telefono |
| Ejemplo Sociedad Anónima | ejemplo@ejemplo.com.ar | 1143XXXXXX |

**Recomendado** - celdas simples, sin combinar

| proveedor_nombre | contacto_correo_electronico | contacto_telefono |
|---|---|---|
| Ejemplo Sociedad Anónima | ejemplo@ejemplo.com.ar | 1143XXXXXX |

### Fila de encabezado

Los datos deben contener sólo una fila de encabezado. Desde la segunda fila en adelante, sólo debe haber datos, pero nunca un encabezado.

**No recomendado** - múltiples filas de encabezado

| Nombre del | Dirección de correo | Teléfono de |
|---|---|---|
| proveedor | electrónico de contacto | contacto |
| Ejemplo Sociedad Anónima | ejemplo@ejemplo.com.ar | 1143XXXXXX |

**Recomendado** - encabezado de una única fila

| proveedor_nombre | contacto_correo_electronico | contacto_telefono |
|---|---|---|
| Ejemplo Sociedad Anónima | ejemplo@ejemplo.com.ar | 1143XXXXXX |

### Celdas vacías en filas para agrupar conceptos

Recomendamos no dejar celdas vacías en filas bajo la presunción de que valores en blanco posteriores a un valor positivo contienen implícitamente a ese mismo valor en una suerte de "agrupamiento conceptual".

Este error es muy común en la construcción de planillas de cálculo y suele generar problemas graves cuando cambia el orden original de las filas. Además, impide el uso de tablas dinámicas y otras formas de analizar los datos.

**No recomendado** - primera celda de la segunda fila vacía

| municipio_nombre | solicitudes_anio | solicitudes_categoria | solicitudes_numero |
|---|---|---|---|
| Municipalidad X | 2015 | Poda y arbolado | |
| | 2015 | Recolección de residuos | 198 |

**Recomendado** - primera celda de la segunda fila completa

| municipio_nombre | solicitudes_anio | solicitudes_categoria | solicitudes_numero |
|---|---|---|---|
| Municipalidad X | 2015 | Poda y arbolado | 340 |
| Municipalidad X | 2015 | Recolección de residuos | 198 |

### Formato de celdas

Las celdas de una planilla de cálculo deben estar formateadas 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”.

| audiencia_fecha_hora | audiencia_sujeto_obligado_nombre | audiencia_numero |
|---|---|---|
| 1/3/16 | Juan | 3456 |
| (Fecha) | (Texto) | (Número) |

Mantener el formato correcto de las celdas según el tipo de datos que contengan:

- Mejora las probabilidades de que una exportación a otro formato salga correctamente.
- Hace que los datos sean más operables en la propia planilla de cálculo, aprovechando mejor sus funcionalidades.

## EXPORTACIÓN A CSV

Insistimos: CSV es el formato más recomendado para la publicación de archivos tabulares. A la hora de exportar una planilla de cálculo a CSV hay 3 parámetros que deben ser especificados durante la exportación, independientemente del software de que se use:

- **Codificación** (encoding, en inglés): siempre debe ser UTF-8.
- **Caracter separador** (separator character, en inglés): siempre debe ser "," (coma).
- **Caracter calificador** (quote character o enclosing character, en inglés): siempre debe ser " “ “ (comillas dobles).

# ESTÁNDARES SEGÚN EL TIPO DE DATOS

El formato recomendado para los distintos tipos de datos está mayormente basado en las especificaciones de la W3C . En los otros casos, las recomendaciones surgen de la experiencia de trabajo del equipo de la Dirección de Datos Públicos y del esfuerzo realizado en la búsqueda de estándares más adecuados.

## TEXTO

- Los campos de texto no deben contener espacios en blanco innecesarios al principio ni al final.

### Entidades

Las entidades que aparezcan entre los datos de un campo textual deben tener una descripción única. Es decir, toda mención que se realice a una entidad dada debe hacerse usando exactamente la misma cadena de caracteres cada vez:

- **Las descripciones de entidades** deberían elegirse siempre de forma tal que cumplan con el estándar específico que las describe, en caso de que este exista.
- **Cuando este estándar no existe** y hay dudas respecto del criterio a adoptar para elegir la descripción única de una entidad, debe **privilegiarse siempre aquella que sea lo más explícita** , descriptiva y declarativa posible.

**No recomendado**

| localidad_nombre |
| --- |
| Ciudad Autónoma de Buenos Aires |
| CABA |
| Capital Federal |
| Ciudad de Buenos Aires |

**Recomendado**

| localidad_nombre |
| --- |
| Ciudad Autónoma de Buenos Aires |
| Ciudad Autónoma de Buenos Aires |
| Ciudad Autónoma de Buenos Aires |
| Ciudad Autónoma de Buenos Aires |

En el ejemplo anterior, los cuatro valores de texto refieren a la misma entidad. Debe elegirse una única forma de referirse a la misma y usarla en todos los casos.

Siempre que sea posible, la elección deberá fundamentarse en el estándar establecido para ese tipo de entidad (para más información ver la **Guía para la identificación y uso de entidades interoperables** ). En el caso de no existir un estándar, deberá adecuarse a las pautas generales contexto del dataset de que se trate.

### Nombres propios

**Se capitalizan** (primera letra de cada palabra es mayúscula, el resto de las letras son minúsculas) **todas las palabras significativas** , salvo las siglas. Las palabras significativas son aquellas que no cumplen la función de artículos o preposiciones.

### Siglas

Todas las siglas se escriben en mayúsculas, sin usar puntos ni espacios intermedios.

## NÚMERO

- El separador decimal debe ser el caracter ".".
- No se usará separador de miles.
- No se deberán usar espacios en blanco.
- Para los números negativos debe incluirse el símbolo menos "-" inmediatamente antes del número, sin espacio en blanco intermedio.

### Moneda

Los valores numéricos que sean además valores monetarios se consideran números y, por lo tanto, valen las mismas recomendaciones que para ellos. Además, agregamos las siguientes recomendaciones:

- La cantidad de decimales debe limitarse a dos, salvo que el uso de una mayor cantidad de decimales sea significativo para el caso particular.
- En ningún caso se incluirán símbolos o letras en el campo numérico -ya sea "$", "DOL", "USD", etc.

Si en el recurso los valores monetarios están expresados en diferentes monedas, se recomienda indicarlo en un campo aparte (que puede llamarse "moneda_id") usando los códigos alfabéticos definidos en la ISO 4127 .

Ejemplo:

- ARS, para el peso argentino.
- USD, para el dólar estadounidense.

### Números telefónicos

En este apartado, proponemos una solución para incluir números telefónicos nacionales en los recursos de datos.

A nivel internacional, el estándar para los números telefónicos fue desarrollado por el "Sector de Normalización de las Telecomunicaciones de la Unión Internacional de Telecomunicaciones" (ITU Telecommunication Standardization Sector) bajo la recomendación E.164 .

Para el caso de los números nacionales, el ENACOM tiene la competencia sobre el sistema de numeración telefónica. Este organismo determina que el número nacional de abonado debe estar compuesto por 10 dígitos. Estos 10 dígitos están conformados por un indicativo interurbano más un número de abonado. Pudiendo el indicativo interurbano tener entre 2 y 4 dígitos, y el número de abonado entre 6 y 8 dígitos.

| Indicativo interurbano (ámbito geográfico al que corresponde) | Número de abonado | Número Nacional interurbano = Indicativo interurbano + Número de abonado |
|---|---|---|
| 11 (AMBA) | 4343XXXX | 114343XXXX |
| 351 (Ciudad de Córdoba) | 434XXXX | 351434XXXX |
| 3837 (Tinogasta) | 43XXXX | 383743XXXX |

Esta numeración es válida para los teléfonos móviles, pero dado que para llamar a un móvil desde un teléfono fijo es necesario anteponer "15" al número de abonado, es necesario que el registro del número telefónico especifique de alguna manera si se trata de un móvil o de un teléfono fijo.

Al no existir estándares nacionales para la inclusión de números telefónicos en recursos de datos, los números telefónicos suelen indicarse de múltiples maneras. Por ejemplo:

**No recomendado** - Múltiples formas de indicar un número telefónico

| proveedor_nombre | contacto_correo_electronico | contacto_telefono |
|---|---|---|
| Ejemplo Sociedad Anónima | ejemplo@ejemplo.com.ar | 01143XXXXXX |
| Ejemplo2 Sociedad Anónima | ejemplo2@ejemplo2.com.ar | 011-45XXXXXX |

Ejemplo3 Sociedad Anónima ejemplo3@ejemplo3.com.ar 351 434-XXXX Ejemplo4 Sociedad Anónima ejemplo4@ejemplo4.com.ar 011 15 6344-XXXX

Para los recursos que contengan números telefónicos nacionales recomendamos como mínimo:

- Respetar el estándar definido por el Número Nacional Interurbano utilizando la conformación de números mediante 10 dígitos.
- Asegurarse de indicar si el teléfono es móvil o fijo.
- Omitir el agregado de dígitos adicionales en el indicativo interurbano. Recomendamos no indicar cero inicial antes del código de área.

Con las salvedades que comentaremos al final de este apartado, un posible abordaje sería el de la tabla a continuación:

**Recomendado** - adecuado al estándar del Número Nacional Interurbano

| proveedor_nombre | contacto_correo_electronico | tipo_numero_telefono | contacto_telefono_indicativo_interurbano | contacto_telefono_numero_abona |
|---|---|---|---|---|
| Ejemplo Sociedad Anónima | ejemplo@ejemplo.com.ar | Fijo | 11 | 43XXXXXX |
| Ejemplo2 Sociedad Anónima | ejemplo2@ejemplo2.com.ar | Fijo | 11 | 45XXXXXX |
| Ejemplo3 Sociedad Anónima | ejemplo3@ejemplo3.com.ar | Fijo | 351 | 434XXXX |
| Ejemplo4 Sociedad Anónima | ejemplo4@ejemplo4.com.ar | Móvil | 11 | 6344XXXX |

Esta recomendación no contempla estos casos específicos:

- No será aplicable a números de uso público, ejemplo: 100, Bomberos; 911, Policía Federal; etc.
- Para casos que requieran la inclusión de más de un número telefónico. Deberán agregarse campos o modificar la estructura de la base de datos.
- Para teléfonos que requieran la inclusión de un número de interno, y al estar éste definido por la persona u organización específica, deberá considerarse la inclusión de otro campo de tipo texto. Ya que los números de interno pueden incluir texto. Ejemplo: "*86", “#36”, etc.
- Para el casos de números internacionales recomendamos contemplar el estándar internacional E.164 .

### Coordenadas

Para registrar datos de coordenadas geográficas de puntos, usamos números decimales. Los campos deberán llamarse "latitud" y “longitud”. Cuando sea conveniente especificar el nombre de la entidad de la cual se consignan las coordenadas, se usarán los sufijos “_latitud” y “_longitud”.

**No recomendado**

| latitud | longitud |
|---|---|
| 34º 11' 30" | 58º 43' 20" |

**Recomendado**

| latitud | longitud |
|---|---|
| -34.6043222 | -58.4134862 |

Para datos geográficos que no sean coordenadas/puntos (por ejemplo líneas o polígonos) recomendamos su especificación en WKT (Well Known Text) . Los puntos/coordenadas también se pueden representar en WKT, pero en estos casos recomendamos utilizar latitud y longitud para representarlos.

También para datos geográficos, recomendamos incluir por una parte un CSV con campos geográficos (en WKT o puntos de coordenadas), y por otra parte el archivo en su formato geoespacial (en alguno de los formatos recomendados anteriormente).

En caso de publicar datos con geometrías (lineas o figuras) en diversos formatos, recomendamos incluir siempre el formato SHP ya que es uno de los más difundidos entre diversas comunidades de usuarios de datos geográficos.

## TIEMPO

### Fecha

Se usará el estándar ISO 8601 (YYYY-MM-DDTHH:MM:SS[.mmmmmm][+HH:MM]). A menos que se indique lo contrario, se asumirá que la zona horaria es UTC-03:00 (Argentina).

- Fecha: YYYY-MM-DD
- Hora: HH:MM:SS[.mmmmmm][+HH:MM]
- Fecha y Hora: YYYY-MM-DDTHH:MM:SS[.mmmmmm][+HH:MM]
- Duración: YYYY-MM-DDTHH:MM:SS[.mmmmmm]

### Rangos horarios

- Los rangos estarán divididos en dos partes separadas por un doble guión bajo "__", la primera indica el día y la segunda, la hora.
- Se puede omitir la parte del día o bien de la hora pero nunca ambas.
- Si se omite la parte que indica el día se asumirá que el rango abarca todo el horario indicado.
- Si se omite la parte que indica el horario se asumirá que el rango abarca todo el día indicado.
- El día se puede indicar tanto mediante rangos separando los días con guiones

medios "-" o bien como particulares con el guión bajo "_".

Ejemplos de formatos válidos para días:

```
DAY: Un solo día
DAY1-DAY2: Entre entre DAY1 y DAY2
DAY1_DAY2: DAY1 y DAY2
DAY1-DAY2_DAY3: DAY1 a DAY2 y DAY3
```

- La hora se indica mediante rangos, separando los horarios con guiones medios ("-"). También se pueden indicar varios horarios con el guión bajo "_".

Ejemplos de formatos válidos para horas:

```
HH:MM-HH:MM : Rango simple
HH:MM-HH:MM_HH:MM-HH:MM : Dos rangos
```

Más ejemplos de formatos válidos completos:

```
HH:MM-HH:MM para indicar un rango que ocurre todos los días.
DAY para indicar que el rango ocupa todo el día DAY.
DAY__HH:MM-HH:MM para indicar un rango que ocurre los días D
AY entre HH:MM y HH:MM.
DAY__HH:MM-HH:MM_HH:MM-HH:MM para indicar mas un rango hora
rio en el mismo día
DAY1-DAY2__HH:MM-HH:MM para indicar un rango que ocurre los
días DAY1 a DAY2 entre HH:MM y HH:MM
DAY1-DAY2__HH:MM-HH:MM_HH:MM-HH:MM para indicar mas un rango
 horario en el mismo rango de días
```

- En caso de que se necesite cubrir más de una franja horaria y esta sintaxis sea insuficiente, se pueden incluir varias separadas por espacios.
- Los días se indicarán con sus iniciales en castellano: LUN, MAR, MIE, JUE, VIE, SAB y DOM

Ejemplos:

```
24hs -> "00:00-23:59"
Jueves 24hs -> "JUE"
Jueves de 14:30 a 17 hs -> "JUE__14:30-17:00"
Jueves de 8 a 12 hs y de 16 a 20 hs -> "JUE__08:00-12:00_16:
00-20:00"
Jueves de 8 a 15 hs y Viernes de 8 a 15 hs -> "JUE__08:00-15
:00 VIE_08:00-15:00"
Lunes a Viernes 7:30 a 17 hs y Sábados 8 a 12 hs -> "LUN-VIE
__07:30-17:00 SAB__08:00-12:00"
Lunes a Viernes 8 a 11 y 14 a 18 hs -> "LUN-VIE__08:00-11:00
_14:00-18:00"
Lunes y Miercoles 8 a 11 y 14 a 18 hs -> "LUN_MIE__08:00-11:
00_14:00-18:00"
Lunes a Miercoles y Viernes 8 a 11 y 14 a 18 hs -> "LUN-MIE_
VIE__08:00-11:00_14:00-18:00"
Lunes a Miercoles 8 a 11 y de Viernes a Domingo 9 a 10 -> "L
UN-MIE__08:00-11:00 VIE-DOM__09:00-10:00"
```

## BOOLEANO

- A menos que se indique lo contrario, se identificarán con los valores *true* o *false* .
    - Esta convención puede variar en algunos rubros específicos de datos, pero en caso de no existir una convención clara y definida aplicable al rubro o contexto del dataset, se recomienda utilizar *true* o *false* .
- Este campo puede contener "valores ausentes". En ese caso, el campo deberá estar totalmente vacío, no conteniendo ningún caracter.
- Si existe la posibilidad de que haya otro valor que no sea *true* , *false* o "valor ausente" significa que se eligió un tipo de datos incorrecto: este no es booleano, el tipo de dato booleano es binario y sólo admite 2 valores de verdad (aparte del caso del “valor ausente”).