Especificaciones técnicas de Datos Abiertos

Este documento está dirigido a las dependencias y entidades de la Administración Pública responsables de cumplir con la implementación de la Política de Datos Abiertos, para lo cual el Gobierno del Estado de Colima celebró el Convenio de Colaboración para Facilitar el Acceso, Uso, Reutilización y Redistribución de los Datos considerados de Carácter Público Puestos a Disposición de Cualquier Interesado en el Sitio de Internet www.datos.gob.mx, y permitirá cumplir con la normatividad vigente ligada a la Política de Datos Abiertos.

Glosario.

Dato: es el registro informativo simbólico, cuantitativo o cualitativo, generado u obtenido por las dependencias y entidades de la Administración Pública.

Conjunto de Datos: la serie de datos estructurados, vinculados entre sí y agrupados dentro de una misma unidad temática y física.

Recurso de Datos: son los archivos descargables en formatos abiertos y accesibles mediante diversos medios de distribución.

Ejemplos:
Conjunto de Datos: Presupuesto anual
Recurso de Datos: Presupuesto 2018, presupuesto 2019, Glosario de claves presupuestales.

Conjunto de Datos: Alumnos Inscritos en la Institución.
Recurso de Datos: Alumnos Inscritos ciclo 2016-2017, Alumnos Inscritos ciclo 2017-2018

¿Cómo facilitar el procesamiento por máquinas?

Los recursos deben estar estructurados para el cómputo, no para la presentación de información para un tomador de decisión (esto último se debe ver como un producto secundario del procesamiento de los datos).

Es así que se deben evitar formatos no estructurados como PDFs, imágenes, y hojas de cálculo con decoraciones como logotipos y encabezados.

Para facilitar el procesamiento por máquinas es importante tener en consideración los siguientes principios de estructuración de las bases de datos a publicar:

1. Los valores no deben presentar agregación estadística o pre procesamiento, de ser así, esto deberá indicarse en los metadatos apropiados.

2. Los nombres de columnas o propiedades de cada registro deben ser altamente descriptivas, p. ej.
“Fecha de solicitud”, “Monto entregado”, etc. En caso de no serlo, se deberá crear un diccionario de datos y agregar su URL de acceso en los metadatos adicionales de la base de datos (ver “Definición de metadatos”).

3. Cuando los valores representan magnitudes, es necesario que permanezcan como datos numéricos y que la unidad de medida se agregue a la descripción del título del campo, p. ej. “Distancia en KM”.

4. Los campos numéricos, incluyendo los monetarios, deben permanecer en un formato numérico de tipo entero o flotante. En este último caso se debe evitar el uso de símbolos monetarios y mejor indicarlo en el título del campo, p. ej. “Monto en pesos mexicanos”, “Monto en USD”, o “Monto en €”.

5. Cuando un campo no tenga valor se debe evitar registrar valores para indicar la ausencia de éste.
Malos ejemplos son las comillas vacías, textos como “No disponible”, “N/A”, etc. La simple ausencia de un valor será el indicador de la falta de dicho dato.

6. Procurar la consistencia de tipos de valores por campo, atributo o columna. En otras palabras, es considerado mala práctica tener valores de diferentes tipos (como texto y número) para una columna en diferentes registros.

7. Los campos de tiempo, como fechas, horas, y rangos temporales, deben seguir el estándar ISO 8601, como se indicó en la sección de metadatos.

8. Para permitir un amplio rango de caracteres (como los acentos), la codificación de los documentos debe ser UTF-8. A pesar de que la codificación ISO 8859-1 (latin-1) cubre los caracteres especiales del español, el estándar UTF-8 además de incluirlos (ISO 10646 base de UTF-8 contiene los caracteres de Latin-1: http://tools.ietf.org/html/rfc3629) extiende a un mayor rango de caracteres y se ha convertido en el estándar de mayor utilización en la Web. Como referencia ver:
http://w3techs.com/technologies/overview/character_encoding/all y
http://googleblog.blogspot.ca/2012/02/unicode-over-60-percent-of-web.htm

Convertir los datos a formatos abiertos.

Publicar datos en formatos abiertos asegura que los datos estén disponibles para ser utilizados por computadoras y personas.

Formatos abiertos
Formatos de archivo para los cuales su especificación se encuentra disponible abiertamente y, por lo cual, no se requiere de licencias o software especializado para su interpretación.

Existen **diversos tipos de formatos para publicar Datos Abiertos**
Por lo anterior, se recomienda seguir los siguientes formatos de publicación:

Datos tabulares, se recomienda el formato CSV, en los cuales los valores o cadenas de caracteres que conforman los datos, se acomodan en filas -separadas por saltos de línea- y columnas -separadas por comas-.

CSV Es un formato simple y adecuado para datos tabulares, que estructura los datos en filas y columnas separadas por comas.

Ejemplo:

Estado	Municipio	Localidad
Colima	Armería	Armería
Colima	Colima	Tepames
Colima	Comala	Suchitlán
Colima	Coquimatlan	Pueblo juarez
Colima	Cuauhtemoc	El Trapiche
Colima	Ixtlahuacán	Las Conchas
Colima	Manzanillo	El Chavarín
Colima	Minatitlán	Agua Salada
Colima	Tecomán	Cofradía de Morelos
Colima	Villa de Álvarez	Juluapan

La tabla de arriba puede ser representada de la siguiente forma en CSV:

```
Estado,Municipio ,Localidad
Colima,Armería,Armería
Colima,Colima,Tepames
Colima,Comala,Suchitlán
Colima,Coquimatlan,Pueblo juarez
Colima,Cuauhtemoc,El Trapiche
Colima,Ixtlahuacán,Las Conchas
Colima,Manzanillo,El Chavarín
Colima,Minatitlán,Agua Salada
Colima,Tecomán,Cofradía de Morelos
Colima,Villa de Álvarez,Juluapan
```

Datos estructurados, se recomienda el uso de los formatos JSON o XML cuya especificación se encuentra disponible abiertamente.

JSON Acrónimo de JavaScript Object Notation, es un formato de texto ligero para el intercambio de datos. JSON es un subconjunto de la notación literal de objetos de JavaScript.

Ejemplo:

Estructura del JSON

JSON Content

```json
{
    "marcadores": [
      {
        "latitude": 40.416875,
        "longitude": -3.703308,
        "city": "Madrid",
        "description": "Puerta del Sol"
      },
      {
        "latitude": 40.417438,
        "longitude": -3.693363,
        "city": "Madrid",
        "description": "Paseo del Prado"
      },
      {
        "latitude": 40.407015,
        "longitude": -3.691163,
        "city": "Madrid",
        "description": "Estación de Atocha"
      }
    ]
}
```

Ejemplo Json minificado.

{"marcadores":[{"latitude":40.416875,"longitude":-3.703308,"city":"Madrid","description":"Puerta del Sol"},{"latitude":40.417438,"longitude":-3.693363,"city":"Madrid","description":"Paseo del Prado"},{"latitude":40.407015,"longitude":-3.691163,"city":"Madrid","description":"Estación de Atocha"}]}

Datos espaciales, se recomienda el uso de los formatos SHP, GeoJSON, o KML;

SHP El formato ESRI Shapefile (SHP) es un formato de archivo informático propietario de datos espaciales desarrollado por la compañía ESRI, quien crea y comercializa software para Sistemas de Información Geográfica como Arc/Info o ArcGIS. Originalmente se creó para la utilización con su producto ArcView GIS, pero actualmente se ha convertido en formato estándar de facto para el intercambio de información geográfica.

Es un formato multiarchivo, es decir está generado por varios ficheros informáticos. El número mínimo requerido es de tres y tienen las extensiones siguientes:

.shp - es el archivo que almacena las entidades geométricas de los objetos.
.shx - es el archivo que almacena el índice de las entidades geométricas.
.dbf - es la base de datos, en formato dBASE, donde se almacena la información de los atributos de los objetos

https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

KML Es un estándar abierto aprobado por la OGC y con notación XML. Comúnmente son agrupados y comprimidos en formato KMZ.

Ejemplo:

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2"> <Placemark>
 <name>Marca de posición simple</name>
 <description>Pegada al suelo. Se coloca de forma inteligente a la altura del relieve subyacente.</description>
 <Point>
 <coordinates>-122.0822035425683,37.42228990140251,0</coordinates>
 </Point>
```

</Placemark> </kml>

Este archivo tiene la siguiente estructura:

- Un encabezado XML. Es la línea número 1 de todos los archivos KML. Antes de esta línea no puede haber caracteres ni espacios.
- Una declaración de espacio de nombres de KML. Es la línea número 2 de todos los archivos KML 2.2.
- Un objeto de marca de posición (Placemark) que contiene los siguientes elementos:
- un nombre (name) que se utiliza como etiqueta para la marca de posición,
- una descripción (description) que aparece en una "viñeta" junto a la marca de posición,
- un punto (Point) que especifica la posición de la marca de posición en la superficie de la Tierra (la longitud, la latitud y, opcionalmente, la altitud).

Documentos de texto, se recomienda el uso del formato ODT. Dicho formato forma parte del estándar ODF (del inglés, *Open Document File Format*).

ODT Formato con notación XML, del estándar abierto OpenDocument.

Finalmente, se podrán publicar **datos en formatos abiertos de base de datos** contenidos en un solo archivo, como es el caso de los archivos de base de datos SQLite. Adicionalmente, se podrán publicar datos que contengan en un solo archivo las sentencias necesarias en sintaxis del lenguaje SQL para que los mismos puedan ser importados en una base de datos que soporte el estándar de SQL.
Además, **se podrán publicar formatos adicionales para cada conjunto** para asegurar la mayor inclusión y entendimiento posible, por ejemplo: en el caso de datos tabulares hacer disponible la versión XLS o XLSX del archivo junto con el CSV.

Documentar de acuerdo al estándar DCAT y publicar el Catálogo Institucional de Datos Abiertos.

Documentación.

Asegurar que los conjuntos de Datos Abiertos incluyen metadatos consistentes y en formatos legibles para humanos y máquinas facilita el entendimiento del origen, tratamiento y significado de los conjuntos y recursos de datos.

Metadatos solicitados (los metadatos marcados con * son obligatorios).

ds:identifier*	Identificador único del conjunto de datos, utilizado para agrupar recursos dentro de éste, por ejemplo "rezago-social", "unidades médicas", "adquisiciones". Utilizar caracteres ASCII (p. ej. sin acentos).
ds:title*	Título descriptivo del conjunto de datos, por ejemplo "Programa de fomento a la agricultura", "Vuelos comerciales".
ds:description*	Una explicación de los datos, con suficiente detalle para que los usuarios puedan entender si es de su interés, por ejemplo "Apoyos otorgados a través del programa Opciones Productivas, desglosado a nivel localidad".
ds:keyword	Lista de términos clave separados por coma, que facilitarán al usuario la búsqueda del conjunto de datos. Es importante considerar el uso de términos no técnicos, por ejemplo "salud, medicinas, compras, agricultura".
ds:modified*	Fecha y hora de la última modificación del conjunto de datos; en formato ISO 8601, por ejemplo "2014-05-27T01:42:05-05:00"
ds:contactPoint*	Nombre de la persona de contacto que atenderá dudas y comentarios sobre el conjunto de datos.

ds:mbox*	Correo electrónico de contacto para atender dudas y comentarios sobre el conjunto de datos.
ds:temporal	La fecha o fechas que cubren los datos, por ejemplo "2013", "2010/2012", "2014-01/2014-04". Si es un rango de fechas, deberán ordenarse ascendentemente.
ds:spatial	El espacio geográfico que cubre el conjunto de datos. Puede ser una región, el nombre de un lugar, una clave INEGI, un polígono o un cuadro delimitador de coordenadas geográficas (bounding box) en GML. Por ejemplo "Baja California", 002, http://www.geonames.org/4017700/baja-california.html, "estatal", o "32.71,-112.32 27.99, -118.45".
ds:landingPage*	Dirección electrónica para obtener mayor documentación o información sobre el conjunto de datos, como lo puede ser un manual, un sitio web, o un diccionario de datos. Este documento sirve como guía adicional para que el usuario entienda con mayor detalle los datos.
ds:accrualPeriodicity*	Frecuencia con la cual el conjunto de datos será publicado o actualizado, por ejemplo "mensualmente".
ds:quality	Información adicional sobre la calidad y procesos de control de calidad de los conjuntos de datos de los procesos de control de calidad del conjunto de datos

Metadatos (descriptores) del recurso o descargable (dcat:Distribution)

ds:identifier*	La clave que identifica al conjunto de datos al que pertenece (y bajo el que se agrupa) este recurso. Ver ds:identifier.
rs:title*	Título descriptivo del recurso o descargable, por ejemplo "Otorgamientos del 2013", "Otorgamientos del 2014", "Apoyos por municipio", "Apoyos por localidad".
rs:description	Ver ds:description. Esta explicación es adicional a la que existe en el conjunto de datos.
rs:downloadURL*	Dirección electrónica (enlace) para la descarga del recurso.
rs:mediaType*	Formato de archivo del recurso a descargar, por ejemplo "text/csv", "application/rss+xml". Este campo permite al usuario buscar conjuntos de datos por formato en datos.gob.mx.
rs:byteSize	El tamaño en bytes del recurso o descargable, por ejemplo 3145728.
rs:temporal	Ver ds:temporal
rs:spatial	Ver ds:spatial
rs:codelists	Información y documentación sobre los códigos utilizados en el recurso de datos, por ejemplo Marco Geoestadístico Nacional o ISO 8601.
re:codelistlink	Hipervínculos a las fuentes oficiales de los códigos y estándares utilizados.
rs:copyright	En caso de ser necesario, describir los derechos de copyright de los datos o información contenida dentro del recurso de datos, por ejemplo Fotografías de un acervo.
rs:tools	Especificar las herramientas recomendadas para el uso, visualización o análisis ligadas al recurso de datos. Por ejemplo JsonView, Herramienta de Datos para Cambio Climático.