Publicacion de paquete de datos tabulareses (CSV)

Traducido y adaptado de: http://dataprotocols.org/tabular-data-package/ por J.P. Aulet - Content licensed under a CC Attribution 4.0 International License.

[Información] Las palabras clave "**DEBE**", "**NO DEBE**", "**REQUERIDO**", "**DEBERÁ**", "**NO DEBERÁ**", "**DEBERÍA**", "**NO DEBE**", "**recomendada**", "**DEBERÍA**", y "**OPCIONAL**" en este documento se DEBE interpretarse como se describe en el RFC 2119 (https://www.ietf.org/rfc/rfc2119.txt).

Resumen

Este documento define un sencillo formato de publicación de datos (paquete de datos tabulares) para la publicación de los datos y el intercambio de cuadros de estilo.

Las características principales de este formato son los siguientes:

- CSV (variables separadas) para los datos
- Archivo JSON Individual (datapackage.json) para describir el conjunto de datos que incluye un esquema para archivos de datos
- Reutilizar siempre que sea posible del trabajo existente, incluyendo otros protocolos de datos

Inicio rápido

Un paquete de datos tabulareses contiene:

- Los archivos de datos en formato CSV
- (Mínimo) información de conjunto de datos en JSON (incluyendo un esquema para el CSV)

He aquí un ejemplo de un simple conjunto de datos mínimos formato de datos:

2 archivos: data.csv y datapackage.json

```
data.csv
--------
var1, var2, var3
A, 1,2
B, 3,4
```

```
datapackage.json
----------------------

{
  "Name":"mi-conjunto de datos",
  "recursos":[
    {
      "path":"data.csv",
      "schema":{
```

```
        "campos":[
          {
            "Name":"var1",
            "Tipo":"cadena"
          },
          {
            "Name":"var2",
            "Tipo":"entero"
          }
        ]
      }
    }
  ]
}
```

Especificación

Esta especificación se basa en la especificación de paquetes de datos. Se define un perfil para la publicación de los paquetes de datos que establece algunas restricciones adicionales sobre los formatos de metadatos y datos.

Un paquete de paquete de datos tabulares válido DEBE ser un paquete de datos válido según se define en dichas especificaciones. Esto significa que **DEBE**:

- Contener un descriptor de paquete (*datapackage.json*)
- Proporcionar al menos el paquete mínimo REQUERIDO metadatos como se describe en la especificación del paquete de datos
- Incluir una descripción de cada archivo de datos en el paquete en la matriz de los recursos del paquete

Además de dichas normas básicas, un paquete tabular paquete de datos válidos deben cumplir con los siguientes requisitos adicionales:

- DEBE contener al menos un archivo de datos
- Todos los archivos de datos deben estar en formato CSV, véase más adelante nuevas normas aplicables por CSV denominación y estructura de archivos
- Cada recurso DEBE tener un esquema siguiendo la especificación JSON tabla de esquema

Los archivos CSV

Los archivos CSV incluidos en un paquete tabular paquete de datos DEBE ajustarse a RFC 4180 "Common Format and MIME Type for Comma-Separated Values (CSV) Files" sujetas a las siguientes excepciones:

- Los archivos deben ser codificados como UTF-8 (el RFC requiere ASCII de 7 bits)
- El carácter estándar de terminación de línea puede ser LF o CRLF (RFC permite que sólo

CRLF)
- Los archivos pueden (pero no DEBERÍA) desviarse de CSV estándar en términos de varios parámetros que incluyen delimitadores de campo

Además de estos requisitos:

- Los nombres de archivo DEBE terminar con `.csv`

- Los archivos deben tener una sola fila de encabezado. Esta fila DEBE ser la primera fila del archivo.
  - *Terminología*: cada columna en el archivo CSV que se denomina un campo y su nombre es la cadena en esa columna en la fila de encabezado.
  - El nombre DEBE ser único entre los campos y DEBE contener al menos un carácter
  - No hay más restricciones a la forma del nombre, pero se recomienda que contiene sólo caracteres alfanuméricos junto con ″.-_″

- Filas del archivo NO DEBE contener más campos de los que están en la fila de encabezado (aunque pueden contener menos)

- Cada archivo DEBE tener una entrada en la matriz de recursos en el archivo `datapackage.json`

- Los metadatos de recursos DEBE incluir un atributo de esquema cuyo valor DEBE ajustarse a la tabla de esquema JSON

- Todos los campos en los archivos CSV deben ser descritos en el esquema

Los archivos CSV generados por las diferentes aplicaciones a menudo varían en su sintaxis, por ejemplo, citando uso de caracteres, delimitadores, etc. Para fomentar el cumplimiento, archivos CSV en un paquete de datos tabulares DEBERÍA

- Uso "," como delimitadores de campo (según RFC)

- Use "\ r \ n" (CRLF) o "\ n" (LF) como terminadores de línea

Si un archivo CSV no sigue estas reglas, entonces su dialecto específico CSV DEBE ser documentada. El hash de recursos para el recurso en el mosto descriptor `datapackage.json`:

¿Por que CSV?

- CSV es muy simple - es quizás el formato de datos más sencillo

- CSV es orientado a datos tabulares. La mayoría de las estructuras de datos son o bien tabulares o pueden ser transformados a una estructura tabular por alguna forma de normalización

- Es abierto y el "estándar" es bien conocido

- Es ampliamente apoyado - prácticamente todos los programas de hoja de cálculo, base de

datos relacional y el lenguaje de programación en existencia puede manejar CSV en una forma u otra

- Es basado en texto y por lo tanto susceptibles de manipulación y acceso entre una amplia gama de herramientas estándar (incluidos los sistemas de control de revisiones, como git, mercurial y subversión)

- Se alineación orientada que significa que puede ser procesado de forma incremental - usted no necesita leer un archivo completo para extraer una sola fila. Por razones similares, significa que el formato es compatible con streaming.

¿Por que JSON para el esquema?

- JSON es simple

- JSON es compatible con la estructura rica incluyendo tipos de anidación y básicos
  JSON es muy ampliamente utilizado y apoyado (todos los principales lenguajes de programación pueden manejar JSON)

- JSON es web-nativo (cada navegador puede acceder y manipular JSON)

- JSON es legible como texto simple por lo que es susceptible de manejo y procesamiento de texto usando herramientas simples