Extracción de metadata de ficheros

Control de versiones

Versión	Fecha de modificación	Responsable	Aprobador	Resumen de cambios
1.0	08/01/2024	Anjana Producto	Anjana Producto	Creación del documento

Introducción

El objetivo de este documento es explicar sobre qué tipos de ficheros es posible extraer el metadato, las peculiaridades y qué información se extrae de cada uno.

El nombre que se indica en cada fichero es el nombre del atributo que debería existir en Anjana (name de la tabla attribute_definition) en las plantillas de objetos de las que se quiera extraer información.

Tipos de ficheros

CSV

Los ficheros de tipo csv se distinguen por su extensión ".csv". Se interpretará cada columna como un dataset_field del que se rellenará la siguiente información:

- **name** con el nombre del campo
- **physical_name** con el nombre del campo
- **fieldDataType** con el tipo de dato definido para el campo (puede ser boolean, number, string o date)
- **position** posición que ocupa el campo

Los separadores admitidos para la extracción de csv son coma (,), punto y coma (;) y tabulación.

AVRO

Los ficheros de tipo avro se distinguen por su extensión ".avro". Estos ficheros pueden ser únicos o particionados.

Se interpretará cada columna como un dataset_field del que se rellenará la siguiente información:

- **name** con el valor del campo
- **physical_name** con el nombre del campo
- **defaultValue** con el valor por defecto definido para el campo
- **fieldDataType** con el tipo de dato definido para el campo (puede ser record, enum, array, map, union, fixed, string, bytes, int, long, float, double, boolean o null)
- **position** posición que ocupa el campo
- **description** con la descripción del campo
- **alias** los alias que el campo tiene

Además de estos valores presentes en todo campo de un fichero avro se pueden poner propiedades extras, todas las propiedades que se incluyan se recogerán y extraerán.

EXCEL

Los ficheros de tipo excel se distinguen por su extensión ".xls" y ".xlsx". Se interpretará cada columna como un dataset_field del que se rellenará la siguiente información:

- **name** con el valor del campo
- **physical_name** con el nombre del campo
- **fieldDataType** con el tipo de dato definido para el campo (puede ser string, boolean, number)
- **position** posición que ocupa el campo
- **description** con la descripción del campo

PARQUET

Los ficheros de tipo parquet se distinguen por su extensión ".parquet". Estos ficheros pueden ser únicos o particionados.

Se interpretará cada columna como un dataset_field del que se rellenará la siguiente información:

- **name** con el valor del campo
- **physical_name** con el nombre del campo
- **fieldDataType** con el tipo de dato definido para el campo (puede ser int64, int32, boolean, binary, float, double, int96 o fixed_len_type_array)
- **position** posición que ocupa el campo
- **nullable** indicando si el campo es nullable
- **length** indicando longitud del campo

Solo aquellos campos que pertenezcan a tipos primitivos serán extraídos.