UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

ANALISIS Y RECOLECCIÓN DE DATOS – SEGUNDO ENTREGABLE

DOCENTE:

Ing. Janneth Alexandra Chicaiza Espinoza

MATERIA:

Sistemas Basados en Conocimientos

ESTUDIANTES:

Jefferson Núñez
José Pullaguari

ABRIL- AGOSTO 2021

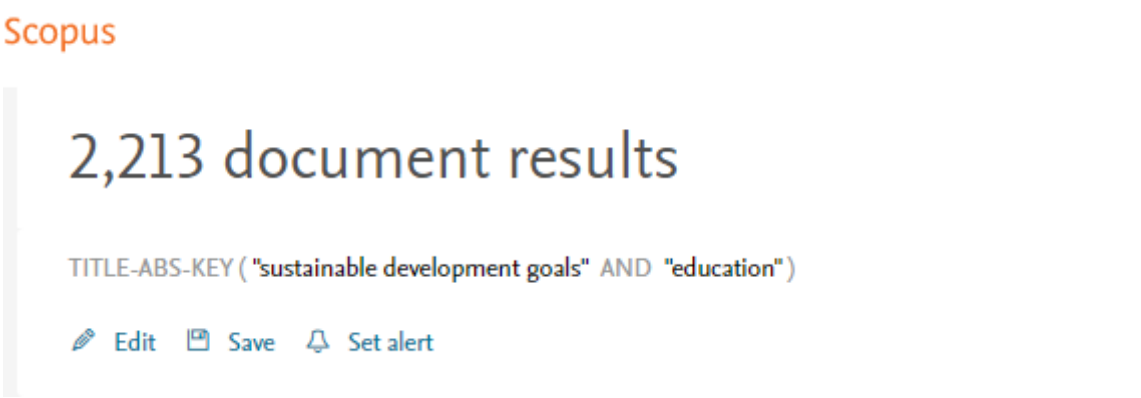
Procesos para la extracción de datos

SCOPUS

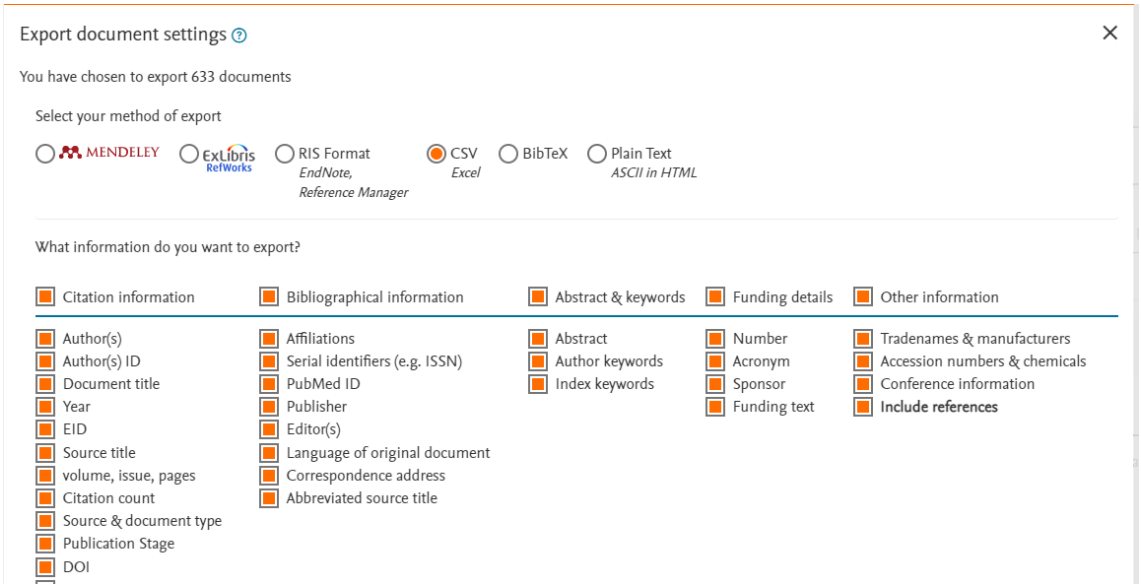
Para el caso de la base de datos SCOPUS, se pudo extraer la información de manera sencilla, ya que dicha plataforma permite la extracción automática en distintos formatos, en nuestro caso se escogió el formato CSV ya que es más flexible.

Proceso

Para esta base de datos bibliográfica se utilizó la ecuación de búsqueda "sustainable development goals" AND "education" en donde se obtuvo un resultado de 2,213



Para la obtención de los datos se seleccionó el formato de exportación en CSV y se marcó los atributos que utilizaremos.



Luego de este procedimiento se obtuvo el archivo CSV, el cual nos permite recuperar el atributo DOi de cada publicación, el mismo que nos servirá para el consumo de datos desde el siguiente motor de búsqueda (Semantic Scholar).

Semantic Scholar

El proceso para el consumo de información desde este motor de búsqueda es diferente, ya que para extracción de datos se utiliza

- Script realizado en Python utilizando la librería **Semanticscholar**
- Librería **Pandas** para trabajar con ficheros CSV

Documentación de la librería

https://api.semanticscholar.org

Recolección de datos

Para la extracción de datos, mediante la librería Pandas(Archivos) se realizó una lectura del archivo CSV, previamente descargado de SCOPUS y luego se aplicó una búsqueda por los DOis de la publicación con la ayuda de la función de paper() de Semantic Scholar.

Por medio de este Script se obtienen los datos en formato JSON para realizar la respectiva limpieza de datos.

```
import semanticscholar as sch
import pandas as pd
import json

data = pd.read_csv('D:\Descargas\data.csv', header=0)
DOIs = data['DOI']

f = open('dataSet', 'w')
try:
    for i in DOIs:
        pub = sch.paper(i)
        print(pub)
        f.write(json.dumps(pub))
finally:
    f.close()
```

Limpieza de datos

El procedimiento para la limpieza de datos es el siguiente: primero se elimina espacios que se han generado durante el proceso de extracción de datos, para este proceso se utilizó la herramienta “PineTools”, luego se realizó un formateo de archivo JSON mediante “JSONFormatter” y finalmente se convierte el archivo en CSV con “Json to CSV”

Modelo Integrado