Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

HERRAMIENTA PARA EL ANÁLISIS Y EVALUACIÓN DE BASES DE DATOS DEL PORTAL DE DATOS ABIERTOS

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.

Sector	**Lenguaje**	**Fuente de datos**
Planeación	Python	Metadatos de las bases de datos de datos.gov.co

Presentación
El portal de Datos Abiertos de Colombia busca acercar a la ciudadanía con los datos que poseen las entidades públicas colombianas. Con el fin de facilitar el acceso y evaluación de estos datos, la UCD creó una librería en Python de uso gratuito. El primer módulo de la librería consiste en funciones que permiten a un usuario evaluar la calidad de cualquier base de datos, una vez ingresada a Python, independientemente si se encuentra en Datos Abiertos o no. La segunda librería facilita el acceso a los metadatos y microdatos que se encuentran en datos.gov.co. Estas librerías fueron desarrolladas para un usuario ya sea de la ciudadanía o del sector público y pueden ser utilizadas como insumo para mejorar la calidad de las bases de datos de datos.gov.co

Colombia´s Open Data webpage goal is to publish the databases of public entities for the benefit of the citizenship. In order to facilitate their access and evaluate quality, the Data Science Unit in DNP created two free-to-use Python libraries. The first one consists of functions that evaluate the quality of any database, not only the ones from datos.gov.co, and the second one connects users to the databases on the webpage and gives information of its metadata. These libraries were developed for independent users as well as for those in the public sector. They can be used as an input to improve the quality of databases from datos.gov.co.

Objetivo general
Desarrollar herramientas computacionales en Python que le ayuden a los usuarios (ciudadanía y entidades públicas) a conocer y evaluar la calidad de las bases de datos de Datos Abiertos

Objetivos específicos
1. Desarrollar un módulo en Python con funciones que evalúen la calidad y analicen estadísticas descriptivas de cualquier base de datos
2. Desarrollar un módulo en Python con funciones que permitan a los usuarios conocer las bases de datos y metadatos del portal de Datos Abiertos y evaluar la calidad de los metadatos
3. Desarrollar un reporte automático que contenga las funciones de las librerías de la herramienta de calidad de datos

Metodología
La metodología del proyecto consistió en crear las librerías de evaluación de bases de datos y la de búsqueda y evaluación de metadatos de datos.gov.co. La primera se puede utilizar para cualquier base de datos y la segunda para las bases y metadatos en datos.gov.co. Ambas se construyeron en Python y los usuarios utilizarían este lenguaje de programación para hacer uso de ellas. A continuación se describe cómo se construyeron, qué métricas de calidad utilizan y sus funciones.

Módulo 1. Análisis de bases de datos y evaluación de su calidad

Este módulo se construyó con base en las librerías de *pandas* y *numpy* de Python. Las funciones que contiene analizan y evalúan la calidad de los datos de cualquier base de datos, no necesariamente las del portal de Datos Abiertos. La evaluación de la calidad de esta librería se basa en 3 criterios: (1) completitud; (2) veracidad y (3) consistencia**.** Estos se describen a continuación.

Completitud
La completitud mide si los campos en las bases están completamente diligenciados. Más específicamente, mide el tamaño de las bases de datos, las filas y columnas, y la cantidad de datos faltantes. Las funciones de la librería contabilizan los datos faltantes como su número total dentro de cada columna y también como el porcentaje dentro de cada una.

Unicidad
La unicidad se refiere a las filas y columnas duplicadas, que no aportan información relevante y aumentan el tamaño de las bases de datos de forma innecesaria. Con la librería se puede calcular el número de las columnas y filas que no son únicas y también emparejar los nombres de las columnas y filas (o su número) duplicadas para que el usuario sepa exactamente cuáles son iguales.

Consistencia
La consistencia evalúa si los datos son coherentes y libres de contradicción, es decir, si la base de datos contiene información acorde con lo que en efecto debería tener. Para que el usuario verifique la consistencia se pueden utilizar varias funciones. La primera muestra el tipo de cada columna en la base de datos, que por lo general suelen ser numéricas o de texto. Si encuentra alguna columna de texto que debería ser numérica, o viceversa, entonces sabría que algo debería corregirse.

Adicionalmente, se desarrollaron funciones que muestran análisis descriptivos para las columnas numéricas y de texto. El análisis descriptivo de las columnas numéricas tiene 5 métricas principales: (1) promedio; (2) desviación estándar; (3) valores en la distribución (mínimo, máximo y percentiles 25, 50 y 75).; (4) valores faltantes; (5) datos extremos. Estas estadísticas se generan en una tabla donde se pueden visualizar estos valores para cada columna numérica. Por su lado, las columnas de texto se analizan de acuerdo con la frecuencia de sus valores únicos, con una función que crea una tabla donde se visualizan los 10 valores más frecuentes para cada columna y su frecuencia, tanto en valor numérico como en porcentaje. También se muestra la frecuencia de los demás valores y los valores faltantes dentro de cada columna.

Módulo 2. Conexión al portal de Datos Abiertos y evaluación de metadatos

El segundo módulo se desarrolló para ayudar a los usuarios con la búsqueda de bases de datos en el portal de Datos Abiertos y evaluar la calidad de los metadatos que se encuentran en él. Para averiguar la información de los metadatos y buscar las bases que sean de su interés, esta librería contiene una tabla con la información de los metadatos en el portal de Datos Abiertos con información sobre el nombre de la base, su descripción, el dueño, las fechas de creación y actualización, la categoría, el tipo, el código API para descargarla con código, entre otros. Se cuenta con una función construida a partir de la API de Socrata para descargar un conjunto de datos de *datos.gov.co* como *dataframe* a Python y un buscador personalizado que permite al usuario filtrar la tabla con los metadatos para encontrar bases que sean de su interés. Este filtro se puede hacer por palabras o términos clave en las variables de texto (título, descripción, etc.), según su número de filas o columnas y fechas de creación o actualización.

En cuanto a las métricas de evaluación, estas buscan verificar la calidad de los metadatos del portal y si la información de los metadatos concuerda con aquella de las bases de datos. Las métricas de evaluación en esta librería son las siguientes: (1) portabilidad, (2) trazabilidad, (3) actualidad y (4) credibilidad. Se presentan a continuación.

Portabilidad

La portabilidad se cumple si la base de datos tiene un formato sin restricciones para la utilización de los datos. En el caso de esta librería se cumple si se logra descargar la base de datos a Python por medio de la API de Socrata y si se puede trabajar con la base.

Trazabilidad
La condición de trazabilidad se cumple si hay información disponible sobre el histórico del conjunto de datos disponible: publicación, actualizaciones y fechas de creación. Esta información se encuentra en la tabla de los metadatos que se puede obtener de la librería 2.

Actualidad
Se cumple si los datos se encuentran vigentes y actualizados. Esta información se puede observar en la tabla con la información de los metadatos y también se escribió una función que calcula si los microdatos y metadatos superan o no el límite de actualización reportado en los metadatos.

Credibilidad
La credibilidad es la información veraz y confiable para los usuarios. Se cumple si existe información sobre la institución dueña de la base de datos y si se puede contactar. La información sobre el dueño de la base de datos está en la tabla con los metadatos, la cual también tiene el link a la página principal de los metadatos en el portal de Datos Abiertos, desde donde se puede contactar con los dueños de la base a través del botón "Contactar con dueño del conjunto de datos"

Reporte automático
Como valor agregado de la librería se ofrece la funcionalidad de generar un reporte de calidad de datos en formato HTML que contenga de manera resumida un análisis descriptivo de la base de datos analizada, utilizando las funciones descritas en las secciones anteriores (data_summary, descriptive_stats y demás) para el análisis. Para le generación de este reporte se utilizó la librería Jinja en Python, la cual, mediante una plantilla implementada en HTML, CSS y javascript permite generar un reporte dinámico en función de los resultados obtenidos al realizar el análisis de calidad de datos, dicho reporte es independiente y solo requiere de un navegador web para su visualización.

Resultados
Para presentar los resultados se toma como ejemplo una base de datos del portal de Datos Abiertos con el nombre *CONSOLIDADO CONTRATACIÓN ANTIOQUIA OCTUBRE 2019* y con código API "9pt8-42xj". A continuación se mostrará el proceso de importación de las librerías en Python, la descarga de la base de datos y las diferentes métricas de evaluación aplicadas a esta base de datos. Estas funciones sirven también para bases de datos distintas a la que se mostrará como ejemplo. Asimismo, se mostrará un ejemplo de cómo buscar una base de datos a partir de la tabla con los metadatos de datos.gov.co.

Importar las librerías
Las librerías se importan con los siguientes comandos en Python:

```
- import datos
- import metadatos
```

Donde *datos* es la librería que evalúa cualquier base de datos y *metadatos* se encarga de conectar al usuario con el portal de Datos Abiertos

Importar la base de datos a Python
La base de datos se importa con la librería *metadatos* con la función *sodapy_data,* la cual utiliza la API de Socrata para conectarse con las bases de datos.gov.co.

- `base=metadatos.sodapy_data("9pt8-42xj",token=`*`código`*`)`

Donde "9pt8-42xj" es la identificación de la API del conjunto de datos que se desea importar y *código* es el código de usuario para eliminar las restricciones de descarga. Este código es opcional y lo puede obtener cada usuario en el portal de Datos Abiertos luego de crear una cuenta allí. La base de datos obtenida por este medio se guarda en este ejemplo como la variable "base" de Python, la cual será utilizada en las métricas de evaluación a continuación.

Descripción general de las bases de datos

Con la función *data_summary* se crea la descripción general de la base de datos ingresada. El código es el siguiente:

- `datos.data_summary(`*`base`*`)`

Donde *base* es el nombre de la base de datos en Python. El ejemplo de la tabla de resumen se encuentra abajo.

Métrica de evaluación	Valor
Número de filas	8250
Número de columnas	25
Columnas numéricas	6
Columnas de texto	19
Número de filas duplicadas	0
Número de columnas duplicadas	0
Columnas con más de la mitad de los datos faltantes	1
Columnas con más del 10% de datos como extremos	5

Tabla 1. Resumen base de datos

Tipos de columnas

Con la función *col_type* de la librería *datos* se pueden visualizar los tipos de columnas de la base de datos. La función se escribe de la siguiente manera:

- `datos.col_type(base)`

La tabla que se crea con esta función es la siguiente:

Columna	Tipo
sujeto_de_control	object
evento	object
fecha_evento	object
tipo_de_registro	object
c_digo_contrato	object
identificaci_n_contratista	object
nombre_contratista	object
c_digo_del_proyecto	object
nombre_del_proyecto	object
sector_del_proyecto	object
valor_del_proyecto	float64

Columna	Tipo
valor_ejecutado_del_proyecto	float64
objeto_del_contrato	object
fecha_suscripci_n	object
fecha_inicio	object
plazo_estimado_d_as	float64
valor_contrato	float64
proceso_de_contrataci_n	object
tipolog_a	object
identificaci_n_interventor	object
nombre_del_interventor	object
tipo_interventor	object
disponibilidades	float64
registros_presupuestales	float64
no_contrato_interventor	object

Tabla 2. Tipos de columnas

Donde *float64* son las columnas numéricas y *object* representa otro tipo, que usualmente son las columnas de texto.

Estadísticas descriptivas de las columnas numéricas

Las estadísticas descriptivas de las columnas numéricas se consiguen con la función *descriptive_stats:*

- `datos.descriptive_stats(base)`

La tabla 3 muestra los resultados. La columna *count* cuenta el número de filas sin valores faltantes, *mean* es el promedio, *std* es la desviación estánda, *min* es el valor mínimo, las columnas de porcentajes corresponden a los percentiles de cada columna¸ *max* es el valor máximo, *missing* es la proporción de valores faltantes y *oultiers* es la proporción de valores extremos.

Columna	count	mean	std	min	25%	50%	75%	max	missing	outliers
valor_del_proyecto	8250	8,61E+10	1,69E+11	1	2,67E+08	2,6E+09	4,31E+10	5E+11	0	0,212242
valor_ejecutado_del_proyecto	8250	6,42E+10	9,08E+10	0	4,44E+08	3,59E+09	1,65E+11	1,03E+12	0	0,000121
plazo_estimado_d_as	8250	40,464	90,80965	1	30	30	47	3653	0	0,188485
valor_contrato	8250	30560083	6,83E+08	0	700000	2600000	7971819	6,02E+10	0	0,135394
disponibilidades	8250	1,82E+08	6,5E+08	0	934506	5000000	47455000	3,01E+10	0	0,187515
registros_presupuestales	8250	16401971	1,22E+08	0	428550	1920000	6163050	6,18E+09	0	0,131152

Tabla 3. Estadísticas descriptivas

Tabla de frecuencias de valores de columnas de texto

En la siguiente tabla se muestran las frecuencias de valores únicos de 2 columnas de texto. Se presentan los 10 valores más frecuentes para cada una, la frecuencia del resto de valores y valores faltantes, tanto numéricamente como proporción del total.

Columna	Valor	Frecuencia	Proporción del total de filas
sujeto_de_control	Universidad De Antioquia	1228	0,148848485
sujeto_de_control	Sociedad Televisión De Antioquia Ltda. - Teleantioquia	775	0,093939394
sujeto_de_control	E.S.E. Hospital La Maria	495	0,06
sujeto_de_control	E.S.E. Hospital Regional San Juan De Dios - Santafe De Antioquia	201	0,024363636
sujeto_de_control	E.S.E. Hospital Regional San Rafael - Yolombo	194	0,023515152
sujeto_de_control	E.S.E. Hospital Marco Fidel Suarez - Bello	182	0,022060606
sujeto_de_control	E.S.E. Hospital Regional San Vicente De Paul - Caldas	124	0,015030303
sujeto_de_control	Admon La Estrella	119	0,014424242
sujeto_de_control	Instituto Para El Deporte Y La Recreacion Indesa - Sabaneta	118	0,01430303
sujeto_de_control	Instituto Municipal Del Deporte Y Recreación De La Estrella - Indere	113	0,01369697
sujeto_de_control	Demás categorías	4701	0,569818182
sujeto_de_control	Datos faltantes	0	0
sector_del_proyecto	SALUD	3094	0,375030303
sector_del_proyecto	EDUCACIÓN	1439	0,174424242
sector_del_proyecto	GASTOS DE OPERACIÓN	957	0,116
sector_del_proyecto	FORTALECIMIENTO INSTITUCIONAL	507	0,061454545
sector_del_proyecto	GASTOS DE FUNCIONAMIENTO	387	0,046909091
sector_del_proyecto	RECREACIÓN Y DEPORTE	371	0,044969697
sector_del_proyecto	APOYO A LA GESTIÓN ADMINISTRATIVA	192	0,023272727
sector_del_proyecto	ATENCIÓN GRUPOS VULNERABLES	185	0,022424242
sector_del_proyecto	AGUA POTABLE Y SANEAMIENTO BÁSICO	148	0,017939394
sector_del_proyecto	GASTOS DE INVERSIÓN	140	0,016969697
sector_del_proyecto	Demás categorías	830	0,100606061
sector_del_proyecto	Datos faltantes	0	0

Tabla 4. Tabla de frecuencias de columnas de texto

Duplicados de filas y columnas

Con las funciones *duplic_col* y *duplic_row* se crean tablas que muestran los nombres de las columnas duplicadas y los números de las filas duplicadas:

- `datos.duplic_col(base)`

- `datos.duplic_row(base)`

En este caso no se encontraron columnas ni filas duplicadas, por lo cual las funciones de la librería muestran los siguientes resultados:

- *"No hay columnas duplicadas"*
- *"No hay filas duplicadas"*

Información de las columnas según los metadatos

La información de las columnas según los metadatos se consigue con el comando *info_cols_meta* de la librería *metadatos:*

- `metadatos.info_cols_meta("9pt8-42xj")`

La tabla 5 muestra los resultados, donde se observa el nombre de las columnas, la descripción y su tipo, tal como se encuentran en la página web de Datos Abiertos.

Nombre de la clumna	Descripión	Tipo
'CÓDIGOCONTRATO'	''	'Textosimple'
'CÓDIGODELPROYECTO'	''	'Textosimple'
'DISPONIBILIDADESPRESUPUESTALES'	''	'Número'
'EVENTO'	''	'Textosimple'
'FECHAEVENTO'	''	'Fechayhora'
'FECHAINICIO'	''	'Fechayhora'
'FECHASUSCRIPCIÓN'	''	'Fechayhora'
'IDENTIFICACIÓNCONTRATISTA'	''	'Textosimple'
'IDENTIFICACIÓNINTERVENTOR'	''	'Textosimple'
'NOMBRECONTRATISTA'	''	'Textosimple'
'NOMBREDELINTERVENTOR'	''	'Textosimple'
'NOMBREDELPROYECTO'	''	'Textosimple'
'NoCONTRATOINTERVENTOR'	''	'Textosimple'
'OBJETODELCONTRATO'	''	'Textosimple'
'PLAZOESTIMADO(DÍAS)'	''	'Número'
'PROCESODECONTRATACIÓN'	''	'Textosimple'
'REGISTROSPRESUPUESTALES'	''	'Número'
'SECTORDELPROYECTO'	''	'Textosimple'
'SUJETODECONTROL'	''	'Textosimple'
'TIPODEREGISTRO'	''	'Textosimple'
'TIPOINTERVENTOR'	''	'Textosimple'
'TIPOLOGÍA'	''	'Textosimple'
'VALORCONTRATO'	''	'Número'
'VALORDELPROYECTO'	''	'Número'
'VALOREJECUTADODELPROYECTO'	''	'Número'

Tabla 5. Columnas según los metadatos

Número de columnas y filas en los microdatos y en los metadatos

Con las funciones *rows_vs_meta y cols_vs_meta* de la librería *metadatos* se busca verificar si las filas y columnas en los metadatos son en efecto las mismas que en los microdatos. Las funciones se escriben de la siguiente manera:

- `metadatos.rows_vs_meta("9pt8-42xj")`
- `metadatos.cols_vs_meta("9pt8-42xj")`

Para la base de datos que se tomó como ejemplo, se verificó que los números de las filas y las columnas sean las mismas en los metadatos y microdatos. Las funciones muestran los resultados de abajo:

- *"Filas en metadatos: 8250.0. Filas en microdatos: 8250"*
- *"Columnas en metadatos: 25.0. Columnas en microdatos: 25"*

Actualización de los datos y metadatos

Para revisar la actualización de los datos y los metadatos se escriben las dos siguientes funciones:

- `metadatos.updated_data("9pt8-42xj","datos")`
- `metadatos.updated_data("9pt8-42xj","metadatos")`

La primera especifica que se quiere revisar la actualización de los datos, o microdatos, y la segunda la de los metadatos. Ambas funciones se corrieron el 29 de noviembre de 2019 y tanto los datos como los metadatos se encontraban actualizados, dado que fueron actualizados el 13 de noviembre y se deben actualizar cada mes. Los resultados de las funciones son las siguientes:

- *"La base de datos fue actualizada hace 16 días, por lo tanto sigue vigente para su período mensual"*
- *"Los metadatos fueron actualizados hace 16 días, por lo tanto siguen vigentes para su período mensual"*

Tabla con los metadatos

La tabla con la información de los metadatos de datos.gov.co se obtuvo a partir de la tabla *Asset Inventory.* La tabla abajo tiene la información con las columnas de la tabla con metadatos, su descripción y los valores para el ejemplo de la base de datos *CONSOLIDADO CONTRATACIÓN ANTIOQUIA OCTUBRE 2019* (para la fecha 18 de noviembre de 2019).

Nombre de la columna	Descripción	Valores de la base de datos ejemplo
titulo	Título de la base de datos	CONSOLIDADO CONTRATACIÓN ANTIOQUIA OCTUBRE 2019
categoria	Categoría económica (algunos ejemplos): Educación, Salud y Protección Social, Función Pública, Ambiente y Desarrollo Sostenible, Transporte, Agricultura y Desarrollo Rural, Estadísticas Nacionales	Organismos de Control
sector	Categoría según los metadatos	Organismos de control y vigilancia

Nombre de la columna	Descripción	Valores de la base de datos ejemplo
tipo	Conjunto de datos, enlace externo, gráfico, vista filtrada, mapa, data lens o archivo o documento	Conjunto de Datos
descripcion	Descripción de la base de datos	Reporte Consolidado de Contratación del mes de octubre de 2019 de las entidades públicas que audita la Contraloría General de Antioquia.
url	Ruta de internet para acceder a la página de la base de datos en datos.gov.co	https://www.datos.gov.co/Organismos-de-Control/CONSOLIDADO-CONTRATACI%C3%93N-ANTIOQUIA-OCTUBRE-2019/9pt8-42xj
url_api	Ruta de internet para acceder a la página de la APIi de cada base de datos (aplica solo para los conjuntos de datos)	https://dev.socrata.com/foundry/www.datos.gov.co/9pt8-42xj
api_id	Código de la API: string de nueve dígitos . Por ejemplo: 9pt8-42xj	9pt8-42xj
departamento	Departamento de la entidad dueña de la base de datos	Antioquia
municipio	Municipio de la entidad dueña de la base de datos	Medellín
dueno	Dueño de la base de datos	ContraloriaAntioquia
entidad	Nombre de la entidad	Contraloría General de Antioquia
area_dependencia	Área o dependencia	Dirección de Sistemas de Buen Gobierno y las Tic
atribucion	Datos ofrecidos por	Contraloría General de Antioquia
cobertura	Cobertura geográfica	Nacional
orden	Territorial o nacional	Territorial
idioma	Idioma dentro de la base de datos	Español
fecha_emision	Fecha de emisión/creación de la base, no necesariamente en datos.gov.co	Fecha Emisión (aaaa-mm-dd) 2019-11-12
creacion	Creación de la base en datos.gov.co	13/11/2019
actualizacion_datos	última fecha cuando se actualizaron los microdatos	13/11/2019
actualizacion_metadatos	Última fecha cuando se actualizaron los metadatos	13/11/2019

Nombre de la columna	Descripción	Valores de la base de datos ejemplo
actualizacion_frec	Frecuencia con la que deberían estar actualizados los datos	Mensual
meta_columnas	Número de columnas según los metadatos	25.0
meta_filas	Número de filas según los metadatos	8250.0
meta_columnas_nombre	Lista con los nombres de las columnas, según los metadatos	['SUJETO DE CONTROL', 'EVENTO', 'FECHA EVENTO', 'TIPO DE REGISTRO', 'CÓDIGO CONTRATO', 'IDENTIFICACIÓN CONTRATISTA', 'NOMBRE CONTRATISTA', 'CÓDIGO DEL PROYECTO', 'NOMBRE DEL PROYECTO', 'SECTOR DEL PROYECTO', 'VALOR DEL PROYECTO', 'VALOR EJECUTADO DEL PROYECTO', 'OBJETO DEL CONTRATO', 'FECHA SUSCRIPCIÓN', 'FECHA INICIO', 'PLAZO ESTIMADO (DÍAS)', 'VALOR CONTRATO', 'PROCESO DE CONTRATACIÓN', 'TIPOLOGÍA', 'IDENTIFICACIÓN INTERVENTOR', 'NOMBRE DEL INTERVENTOR', 'TIPO INTERVENTOR', 'No CONTRATO INTERVENTOR', 'DISPONIBILIDADES PRESUPUESTALES', 'REGISTROS PRESUPUESTALES']
meta_columnas_descr	Lista con la descripción de las columnas, según los metadatos	['', '']
meta_columnas_tipo	Lista con los tipos de las columnas, según los metadatos	['Texto simple', 'Texto simple', 'Fecha y hora', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Número', 'Número', 'Texto simple', 'Fecha y hora', 'Fecha y hora', 'Número', 'Número', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Número', 'Número']

Tabla 6. Metadatos

Búsqueda de bases de datos dentro de la tabla de metadatos

El buscador de bases de datos de la librería *metadatos* permite buscar palabras o términos clave dentro de las columnas de texto (como *título*, *descripción* o *api_id*), según el número de filas o columnas y la fecha de creación o actualización. El buscador se define con dos diccionarios de Python, uno que contenga las palabras o elementos clave y otro que especifique la manera como se quiere hacer el filtro. Abajo se observa un ejemplo:

- ```
  columnas_valor={
        "titulo":["SECOP"],
        "descripcion":["compra"],
        "meta_filas":[100,10000],
        "meta_columnas":[5,30],
        "creacion":["01/01/2019","31/12/2019"]
        }
  ```
- ```
  columnas_operacion={
        "titulo":"contiene",
        "descripcion":"contiene",
        "meta_filas":"entre",
        "meta_columnas":"entre",
        "creacion":"entre",
        }
  ```

En el diccionario *columnas_valor* se define que se busca el carácter "SECOP" en la columna "titulo" y "compra" en la columna "descripción". En el diccionario *columnas_operacion* se especifica la palabra "contiene" para estas dos columnas, es decir, que las columnas contengan los caracteres especificados, pero no se busca que el título de la base de datos sea exactamente "SECOP" y la descripción "compra". El buscador funciona de tal manera que no importa si se escribe la palabra o término que se quiere buscar en mayúsculas o con tildes, ya que se quitan las tildes y se pasan las palabras a minúsculas tanto en los términos buscados como en las columnas de interés antes de hacer el filtro.

Las listas de *meta_filas* (`[100,10000]`) y *meta_columnas* (`[5,30]`) dentro del diccionario *columnas_valor* indican el número de columnas y filas deseadas en las bases de datos buscadas. En el diccionario *columnas-_operacion* se especifica que se quieren bases con entre 100 y 10.000 filas y entre 5 y 30 columnas. Si se buscaran valores iguales a un número se escribiría la opción "igual".

Por último, se especifica que la fecha de creación de las bases debería estar entre el primero de enero de 2019 y el 31 de diciembre de 2019. Se escriben la fecha inicial y final en el diccionario *columnas_valor* y en columnas_operacion se resalta que sean fechas entre estos valores (no iguales a un día en específico).

Estos diccionarios se ingresan a la función *table_search*. Se escribe de la siguiente manera:

- `metadatos.table_search(columnas_valor,columnas_operacion)`

El filtro con los diccionarios ingresado da como resultado una base de datos titulada "SECOP I – Proponentes". Los valores de esta base de datos se encuentran en la tabla abajo.

Nombre de la columna	Valores dentro de la columna
titulo	SECOP I - Proponentes
categoria	Gastos Gubernamentales
sector	Planeación
tipo	Conjunto de Datos
descripcion	Información de proponentes por proceso de compra y su calificación
url	https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-I-Proponentes/tauh-5jvn
url_api	https://dev.socrata.com/foundry/www.datos.gov.co/tauh-5jvn
api_id	tauh-5jvn
departamento	Bogotá D.C.
municipio	Bogotá D.C.
dueno	Colombia Compra Eficiente
entidad	Agencia Nacional de Contratación Pública Colombia Compra Eficiente
area_dependencia	Subdirección de IDT
atribucion	Colombia Compra Eficiente
cobertura	Nacional
orden	Nacional
idioma	Español
fecha_emision	Fecha Emisión (aaaa-mm-dd) 2019-10-18
creacion	18/10/2019
actualizacion_datos	6/11/2019
actualizacion_metadatos	18/10/2019
actualizacion_frec	Diaria
meta_columnas	6.0
meta_filas	3837.0
meta_columnas_nombre	['ID Proceso', 'Tipo Doc Proponente', 'Num Doc Proponente', 'Digito Verificación Proponente', 'Proponente', 'Calificacion']
meta_columnas_descr	['', '', '', '', '', '']
meta_columnas_tipo	['Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple', 'Texto simple']
creacion_fecha	18/10/2019

Tabla 7. Resultados del buscador de bases de datos

Reporte automático

El reporte automático muestra en un archivo HTML los resultados de las funciones descritas arriba. A continuación se muestran ejemplos de cómo se ve el reporte en este archivo (se utiliza otra base de datos de ejemplo a la de arriba). La ilustración 1 tiene las estadísticas generales o resumen de la base de datos

Ilustración 1. Reporte automático – estadísticas generales

Reporte perfilamiento

Reporte generado automáticamente 03-06-2020 10:51:58 AM

Estadísticas generales

Categoria	Valor
Número de filas	1239
Número de columnas	51
Columnas numéricas	25
Columnas de texto	24
Columnas boolean	2
Columnas de fecha	0

Categoria	Valor
Otro tipo de columnas	0
Número de filas no únicas	8
Columnas con más de la mitad de datos faltantes	17
Columnas con más del 10% de datos como extremos	14
Tamaño de la base en megabytes (redondeado)	0

La ilustración 2 tiene el visualizador de las 10 primeras filas de la base de datos. Existe la opción de observar también las 10 últimas filas y de ver los resultados de cada columna al arrastrar el cursor en la parte de abajo.

Ilustración 2. Reporte automático – muestra de datos

Muestra de datos

Primeras 10 filas | Últimas 10 filas

Id	Nombre de la simulación	TipoDocumentoId	NumIdTributaria	TipoUsuarioId	Moneda Local	FechaCreacion	Tip
28233	CAFE	1.0	x	3	COP	NaN	Otr
13380	bolso3	1.0	NaN	3	COP	NaN	Otr
11524	Exportación China	1.0	NaN	3	COP	NaN	Otr
11562	COSTO DE EXPORTACION PETROLEO	1.0	NaN	3	COP	NaN	Otr
11562	COSTO DE EXPORTACION PETROLEO	1.0	NaN	3	COP	NaN	Otr
11562	COSTO DE EXPORTACION PETROLEO	1.0	NaN	3	COP	NaN	Otr
11562	COSTO DE EXPORTACION PETROLEO	1.0	NaN	3	COP	NaN	Otr
11562	COSTO DE EXPORTACION PETROLEO	1.0	NaN	3	COP	NaN	Otr
11562	COSTO DE EXPORTACION PETROLEO	1.0	NaN	3	COP	NaN	Otr
26780	COMERCIALIZADORA J&D	1.0	NaN	3	COP	NaN	Otr

La ilustración 3 muestra el cuadro en el que se pueden ver las estadísticas descriptivas, información general de cada variable, la frecuencia de las categorías principales de variables categóricas y los datos duplicados. En la imagen están las estadísticas descriptivas. El resto de las funciones se pueden visualizar al oprimir los botones en la parte superior.

Ilustración 3. Reporte automático – estadísticas descriptivas

Estadística	Id	TipoDocumentoId	TipoUsuarioId	FechaCreacion	Valor unidad cor
count	1239.00	1207.00	1239.00	38.00	718.00
mean	15769.97	1.01	3.01	43144.81	10236884.85
std	8567.03	0.10	0.21	332.31	106098046.40
min	51.00	1.00	2.00	42521.90	0.06
25%	8652.00	1.00	3.00	42883.87	4350.00
50%	15937.00	1.00	3.00	43159.51	45000.00
75%	23364.00	1.00	3.00	43409.63	192000.00
max	31536.00	2.00	4.00	43648.51	1683348362.00
missing	0.00	0.03	0.00	0.97	0.42
outliers	0.00	0.01	0.04	0.00	0.12

Trabajo futuro

Se harán varias pruebas con usuarios para recibir comentarios sobre las funciones de la librería de calidad de datos. Entre los comentarios que ya se han recibido, se agregarán funciones para enriquecer el análisis descriptivo, se revisará si se deben juntar funciones con métricas similares y añadir parámetros a ellas. Sobre el reporte automático, se agregará la sección de análisis de metadatos del Portal de Datos Abiertos. Se especificará si se quiere analizar una base de datos del portal y de esta manera se generará el reporte con la información ya presentada arriba y también una sección que muestre los metadatos de esa base. Por último, se hará el trabajo para que la librería quede en un repositorio en la web y pueda ser accedida por cualquier usuario.