PROGRAMA IBERDROLA INNOVA I+D+i 2021/2022

LÍNEA DE INVESTIGACIÓN
Línea de Tecnologías de la información y las Comunicaciones

GUÍA INTRODUCTORIA AL TEMA:
"Fuentes de datos abiertos en Salud a nivel mundial y el impacto de su utilización"

La Inteligencia Artificial necesita datos, no hay Inteligencia Artificial sin datos

En la era del dato, las empresas disponen de gran cantidad de información, a partir de la que tomar decisiones estratégicas. La enorme cantidad de datos que surgen del control y la interacción de miles de millones de dispositivos y sensores necesitan ser analizados, para transformarlos en información útil para distintos sectores de actividad.

Para conseguirlo, se requieren procesos cada vez más automatizados e inteligentes. La inteligencia artificial y el aprendizaje automático juegan un papel crucial. Las tecnologías inteligentes como el aprendizaje automático y el aprendizaje profundo serán claves a la hora de filtrar el volumen de datos, reconocer patrones y analizar los datos más relevantes de forma rápida.

Los datos son actualmente fuente de aprendizaje de sistemas, software y aplicaciones que están presentes en nuestra vida y trabajo diario. El dato en bruto debe ser limpiado, enriquecido y consolidado y así conseguir datos de alta calidad que permitan obtener resultados óptimos en los análisis.

El objetivo de este trabajo es que viváis este proceso desde la búsqueda del dato hasta su refinamiento y a aquellos que estén preparados y se atrevan a su modelización.

1. ¿Qué es el "*Open Data*"?

El concepto *Open Data*, o Datos Abiertos, hace referencia a una filosofía de acumulación de datos electrónicos que persigue su divulgación, con pocas o ninguna restricción y permitiendo su uso por parte de terceros, con el simple objetivo de ayudar a cualquier persona, entidad u organización que los necesite para desarrollar una actividad, un proyecto de investigación, un negocio o con cualquier otra finalidad.

El movimiento de Datos Abiertos se consolidó cuando en diciembre de 2007 una treintena de expertos del gobierno abierto de los Estados Unidos estableció "los 8 principios del *Open Data*", que se resumen en lo siguiente:

1. **Libertad**. Los datos abiertos serán públicos y accesibles, y bajo ningún concepto se restringirá su acceso por privacidad, seguridad, privilegios o derechos de autor.
2. **Originalidad**. Los datos abiertos serán presentados como originalmente se encontraban cuando se recolectaron, lo más detallados posible y sin haber sido modificados ni transformados. En definitiva, los datos abiertos son datos en bruto.
3. **Actualidad**. Los datos abiertos serán presentados lo antes posible, de forma que sean útiles en el momento de su publicación y no "caduquen" por el paso del tiempo.
4. **Accesibilidad**. Los datos abiertos estarán disponibles al mayor número de personas para el mayor número de propósitos. Para su publicación se debe de utilizar Internet y permitir la descarga de estos de manera sencilla, para que todos podamos hacerlo.
5. **Claridad**. Los datos abiertos tendrán una estructura que permitirá a programas informáticos su procesamiento automático. Esto significa que los datos estarán en formato texto limpio ("Bloc de notas" de Windows) y con suficientes explicaciones de qué significa cada característica de los datos recogidos.
6. **Libre descarga**. Los datos abiertos se podrán descargar sin registrarse en ningún formulario ni dejar constancia de quién ha accedido a ellos. Es decir, se garantizará el acceso anónimo a los datos.
7. **Dominio público**. Los datos abiertos no serán propiedad de nadie; serán propiedad de todos. De esta manera, ninguna persona ni organismo podrá restringir quién puede ver y utilizar los datos.
8. **Sin licencia**: Los datos abiertos no pueden ser licenciados por *copyright*, patentes o marcas comerciales. Pueden someterse a ciertas medidas de seguridad y privacidad, pero nunca pueden ser vendidos o intercambiados "en secreto".

En resumen, los datos abiertos son un material que los proyectos de investigación, instituciones gubernamentales y negocios pueden utilizar libremente y sin ninguna restricción, salvo la de mantener y perpetuar el "trato" o "acuerdo" de los datos abiertos (es decir, que sigan estando disponibles para todos y los demás principios descritos anteriormente).

En este programa, vosotros, los estudiantes, vais a beneficiaros de esta filosofía de los datos abiertos para realizar un pequeño proyecto de investigación en el área sanitaria. Todo proyecto de investigación, especialmente si utiliza Inteligencia Artificial, necesita de muchos datos, como se ha discutido con anterioridad. Es por ello por lo que cada vez más gobiernos europeos y empresas están impulsando este movimiento de **compartir datos**, pues con ello estamos también **compartiendo conocimiento** en beneficio de todos.

2. ¿Y la otra cara de la moneda?

En parte, estaríamos hablando de **datos privados**. Se tratan de datos generalmente sensibles, que pueden contener información personal, han de ser protegidos y custodiados con cautela, pues tienen que permanecer en secreto. Entre ellos pueden incluirse nuestro nombre, apellidos, DNI o dirección de domicilio; esta, lógicamente, es información que no puede aparecer pública por Internet, y se diferencia de los datos abiertos en que **necesitan consentimiento expreso** del propietario de los datos para ser tratados por instituciones, empresas y cualquier persona.

Esto nos da la idea de que los Datos Abiertos tienen que ser **anónimos** y **no comprometer la privacidad de nadie**. Veremos cómo manejar este delicado equilibrio entre libertad y privacidad a continuación.

3. Utilizando *Open Data*

Un ejemplo de un archivo de datos puede ser el siguiente (se muestran tan solo las 12 primeras filas de las 186 que tenía en la fecha de consulta).

	location	date	vaccine	total_vaccinations	source_url
2	Spain	2021-01-04	Pfizer/BioNTech	82834	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
3	Spain	2021-01-05	Pfizer/BioNTech	139339	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
4	Spain	2021-01-07	Pfizer/BioNTech	207323	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
5	Spain	2021-01-08	Pfizer/BioNTech	277976	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
6	Spain	2021-01-11	Pfizer/BioNTech	406091	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
7	Spain	2021-01-12	Pfizer/BioNTech	488122	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
8	Spain	2021-01-13	Pfizer/BioNTech	581638	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
9	Spain	2021-01-14	Pfizer/BioNTech	676186	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
10	Spain	2021-01-15	Moderna, Pfizer/BioNTech	768950	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm
11	Spain	2021-01-18	Moderna, Pfizer/BioNTech	897942	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Informe_
12	Spain	2021-01-19	Moderna, Pfizer/BioNTech	966097	https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Informe_

Ilustración 1: Ejemplo de un archivo de Datos Abiertos por la iniciativa OWID (https://ourworldindata.org/) de las vacunaciones contra el COVID-19 al día en España: número de pinchazos, con qué vacunas, y el enlace de donde se sacó la información.

Este tipo de archivos se componen, como puede verse, de varias columnas con diferentes datos. Este, en particular, cumple con los 8 principios descritos anteriormente, en especial con el de “Claridad”, pues puede ser procesado por un programa informático muy fácilmente. Con el nombre del archivo, un programa puede leer y preparar todo su contenido para comenzar a transformarlo según indique el programador: modificando los datos, transformándolos, dibujando gráficas automáticamente y hasta introduciéndolos en una Inteligencia Artificial para que esta nos ayude a tomar decisiones. Siguiendo esta línea, la Inteligencia Artificial es capaz, con los suficientes datos, de hacer **predicciones** sobre datos futuros **aprendiendo** de los pasados.

Pongamos por caso el consumo eléctrico durante un día de un ordenador. En ese valor influyen otros como: el consumo individual de cada componente, la carga de trabajo del ordenador, el tiempo que permanece encendido, la temperatura ambiente de la habitación… Si recopilamos una cantidad suficiente de estos y más datos, una Inteligencia Artificial podría **aprender** de ellos y **predecir** el consumo del ordenador para un día cualquiera si le indicamos la temperatura ambiente de este día, el tiempo que permaneció encendido y todos los demás datos que habíamos recopilado.

La recopilación de datos, como es fácil imaginar, es un trabajo muy duro. Entonces, ¿por qué hay que hacerlo una y otra vez si ya se había hecho previamente? Esta es la filosofía de los Datos Abiertos: ahorrarnos trabajo los unos a los otros, sin restricciones. Otra importante ventaja que tienen es que, al trabajar con Datos Abiertos, hay que **preocuparse de bastantes menos cuestiones** que con los datos privados, como hemos visto antes. Normalmente el uso de datos privados está sujeto a muchas condiciones y contratos que habrá que respetar y su uso está pensado para **fines concretos**. Sin embargo, con una fuente de Datos Abiertos, las posibilidades son mucho más amplias porque, para empezar, **(casi) cualquier finalidad está permitida**. Vamos a enumerar las características de los Datos Abiertos que hay que tener en cuenta antes de empezar un proyecto.

1. **Que *sean* Datos Abiertos.** Aunque normalmente una búsqueda por Google nos puede llevar directamente a Datos Abiertos, nunca está de más comprobar por nosotros mismos que, efectivamente, se trata de Datos Abiertos, y bajo qué condiciones. Los podemos identificar si, en la página donde los descargamos, vemos que están licenciados con la familia de licencias "*Creative Commons*" o equivalentes; o si se da consentimiento expreso en algún apartado de "Condiciones". En este último caso, nos tenemos que adherir a estas condiciones, *incluso* si van en contra de alguno de los 8 principios anteriores (en la práctica se verá que esto es más habitual de lo que nos gustaría).

Ilustración 2: Publicación de datos de la Junta de Castilla y León, licenciada bajo "Creative Commons 4.0", lo que permite compartir y adaptar los datos, incluso para fines comerciales.

Fuente: https://datos.gob.es/es/catalogo/a07002862-relacion-de-bibliobuses-de-castilla-y-leon

Ilustración 3: Puntos clave de la licencia Creative Commons 4.0.

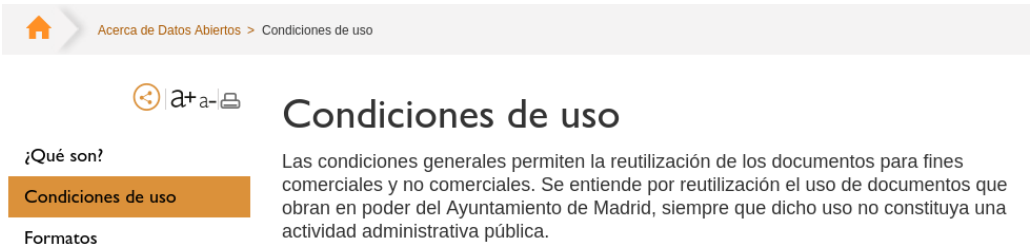

Ilustración 4: Condiciones de uso para los Datos Abiertos del Ayuntamiento de Madrid. Se permite su uso para fines comerciales y no comerciales, pero no para actividades administrativas públicas.

Fuente: https://datos.madrid.es/portal/site/egob/menuitem.400a817358ce98c34e937436a8a409a0/?vgnextoid=b4c412b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextchannel=b4c412b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

2. **Que los datos sean anónimos.** La condición indispensable de los Datos Abiertos es que no amenacen la privacidad de nadie. Comprueba que tus datos no identifiquen, ni puedan identificar, a particulares; y si lo hacen, que sean

particulares anónimos (por ejemplo, para identificar pacientes en un hospital, en vez de utilizar "Carmen", "José" y "Adrián", utilizar "1", "2" y "3", para preservar su anonimato). También hay que cuidar la posible identificación indirecta estadística; por ejemplo, no se pueden dar datos de un barrio si este solo tuviera 6 habitantes, puesto que, al ser un número tan reducido, pueden ser fácilmente identificados con tan solo uno o dos datos suyos, como su edad, género o procedencia.

3. **La legalidad del trabajo.** Mientras que, en lo que a los datos respecta, la finalidad no suele estar limitada, no por ello hemos de pensar que podemos hacer *cualquier cosa* con los Datos Abiertos. Hay que asegurarse de que el trabajo en sí **cumple con la ley**, especialmente si va a ser divulgado públicamente o comercializado. Es decir, hemos de tener cuidado de no incurrir en delitos como el de acoso o el de odio, o de plagiar toda o parte de una obra o estudio que sí contenga derechos de autor; entre muchas otras cosas.
4. **La mezcla entre Datos Abiertos y datos privados.** Muchas veces algunos proyectos utilizan Datos Abiertos y privados. Estos casos son complejos, porque el trato que les podemos dar a unos y a otros es completamente distinto y frecuentemente se mezclan sin distinguirlos: no podemos publicar alegremente ni siquiera una parte de los datos privados si no tenemos permiso para hacerlo.
5. **El sentido común.** La filosofía de los Datos Abiertos se basa, en última instancia, en el sentido común del investigador. Aunque encuentres una fuente de datos abierta, **permítete juzgar si es ético** utilizarlos para tu propósito, y si no estás amenazando la privacidad o imagen de nadie con ello. Aquí tenéis un link a la normativa

4. Los Datos Abiertos en el sector salud

A priori parece que el concepto de Datos Abiertos queda algo fuera del sector de la salud, especialmente por los problemas de privacidad que presenta. Sin embargo, por lo que ya sabemos, a los pacientes 1, 2 y 3 de varios hospitales A, B o C, se les puede **anonimizar**, como se ha demostrado antes. Es más: hay mucho que ganar en el sector sanitario de los datos abiertos, como vamos a comprobar a continuación.

Ello quedó patente cuando Alemania notificó en 2011 un **brote epidémico** de la bacteria Escherichia Coli (E. Coli) y, en contra del procedimiento habitual científico, liberó los datos genómicos de la misma el primer día a través de, ni más ni menos, la red social *Twitter*. El resultado fue un proceso colaborativo de análisis del genoma (que algunos llamaron con el divertido término de "*Tweenoma*") en el que participaron microbiólogos de todo el mundo, que rastreó en cuestión de pocos días el origen de la epidemia. Gracias a

ello, **muchos agricultores españoles se liberaron del boicot** al que vieron sometidos sus productos tras el brote, pues al principio se pensaba que el origen estaba en ellos.

De igual manera, España cuenta con pequeñas aplicaciones que también se alimentan de Datos Abiertos. Dos farmacias de guardia en Navarra y el buscador de centros de salud de la Comunidad de Madrid los utilizan para encontrar proveedores cercanos de atención sanitaria. *ZaraHealth* monitoriza la calidad del agua, el estado del aire y los niveles de polen de Zaragoza con Datos Abiertos. Y, evidentemente, todos los que se han ido publicando desde 2020 sobre la pandemia del COVID-19 de incidencia acumulada, muertes, etc., por comunidad autónoma y municipios, nos han ayudado, y nos ayudan, tanto a los ciudadanos como a los gobernantes a tomar decisiones que, en definitiva, afectan a nuestra salud.

Sin embargo, en la situación actual nos encontramos ante un problema que tiene tres vertientes:

- **En las empresas de la salud, hospitales y administraciones:** Los médicos y personal del sector salud desconocen qué potencial tienen los datos que manejan todos los días y cómo utilizarlos. Necesitan el apoyo de empresas tecnológicas, pero...
- **En las empresas tecnológicas:** Saben desarrollar los programas y algoritmos, pero no poseen los datos para alimentarlos ni los medios para recopilarlos. Además, también les faltan infraestructuras (potencia computacional) para llevar todos sus proyectos a cabo.
- **En la sociedad en su conjunto:** Los ciudadanos tienen miedo al manejo de datos por parte de las empresas. Las empresas, a compartir sus datos, por ser un elemento de ventaja en la competitividad. Todo ello es una consecuencia de la falta de un marco regulatorio y de canales y herramientas de compartición de datos.

5. Carencia de datos en el sector salud

La realidad es que para estas aplicaciones tan importantes **no existen abundantes datos** con los que “alimentarse”. Además, tenemos que añadir a esto que muchas organizaciones relacionadas con la salud no utilizan aún, ni publican, Datos Abiertos. Según el estudio “The Open Data Impact Map”, un proyecto de Open Data for Development Network (OD4D), de 124 organizaciones encuestadas en 2018, **tan solo 19** aseguraron que aprovechaban los Datos Abiertos para desarrollar sus servicios y productos sanitarios, **y únicamente 13** para su investigación. Asimismo, tal y como

relatan dos científicos y un profesor[1] de Broad Institute y la Universidad de Harvard, durante la crisis del **ébola** en África en 2014, se liberaron una gran cantidad de secuencias genómicas en los primeros meses que fueron vitales para rastrear el origen y asegurar cómo se transmitía el virus, entre muchas otras conclusiones. Sin embargo, durante el pico más dramático de la crisis, entre septiembre y noviembre de 2014, a pesar de que se sabía que los laboratorios en Estados Unidos y demás países habían conseguido secuenciar muchos más genomas, **no se hicieron públicos** porque los científicos no tenían un **estándar** para hacerlo. Los autores de este relato concluyen que aquello fue una oportunidad perdida: se podrían haber salvado muchas más vidas si solo hubiéramos tenido una **manera común** de hacer públicos aquellos datos.

Es por este motivo por el que empiezan a despegar iniciativas de esta índole: plataformas de Datos Abiertos y repositorios públicos, alimentados por los datos de usuarios, empresas y, especialmente, gobiernos y universidades. Es por ello que en este programa los estudiantes **vais a ir a la búsqueda de estos Datos Abiertos,** para luego realizar un estudio sobre ellos. Estos repositorios públicos y plataformas de datos, como se había mencionado antes, se encuentran muy fácilmente con una simple búsqueda en Google. Pongamos por caso la plataforma de datos abiertos del Gobierno de España.

Ilustración 5: Página de inicio de la plataforma de Datos Abiertos del Gobierno de España, https://datos.gob.es/

Navegando por esta página, cualquier persona puede acceder al **catálogo de datos** y empezar a curiosear sobre estos datos que tiene disponible el propio Gobierno para que cualquiera, usuario o empresa, los utilice en sus proyectos o negocios. Los datos de una **plataforma de datos** generalmente vienen de diferentes formas:

[1]Nathan L. Yozwiak, Stephen F. Schaffner y Pardis C. Sabeti, respectivamente.

- **Archivos.** Suele ser la forma más sencilla y práctica de difundir datos. Un usuario puede **descargarse estos archivos** y ya tiene en posesión sus datos abiertos. Los formatos de estos archivos varían y muchas plataformas los ofrecen en varios de ellos, pues cada uno responde a necesidades distintas. Algunos de estos formatos pueden ser el clásico Excel (.xlsx), pero más interés tienen para los programas los archivos "Valores Separados por Coma" (.csv, del inglés, "*Comma-separated values*) que, como su nombre indica, no es más que texto separado por comas, legible para cualquier máquina. **Este va a ser el más interesante para nosotros**. Otros formatos de archivo comúnmente utilizados son el tipo JSON y hasta texto plano (TXT), separado por comas, punto y comas o tabuladores.
- **API.** Este es un método más avanzado y muy interesante para los programadores. En este caso, es el programa informático el que **"pregunta" a la plataforma de datos** mediante fórmulas ya programadas sobre los datos que quiere específicamente. Esto permite conseguir siempre los datos más actualizados, pues el programa se estará "descargando constantemente" los datos, en vez de guardarse un archivo.
- **SPARQL.** Se trata de un lenguaje de consulta de base de datos unificado que tiene una ventaja sobre la API, y es que las fórmulas para hacer "preguntas" son iguales para todos. Algunas plataformas de Datos Abiertos también permiten consultas de este tipo.

También existen, como habíamos mencionado antes, **repositorios de datos**. Se diferencian de las plataformas de datos en que son **más sencillos y grandes en número**, porque gracias a ellos **cualquier persona** puede subir Datos Abiertos en cuestión de minutos. No requieren del despliegue de toda una página web como se había visto con el Gobierno de España, pero a cambio, no tienen tantas funciones (o pueden parecer no tan accesibles para todo el mundo). En cualquiera de los casos, no por ello es menos fácil descargarse datos de un repositorio. La red de repositorios más grande en el mundo es **GitHub**, y uno de sus repositorios de Datos Abiertos podría tener el siguiente aspecto:

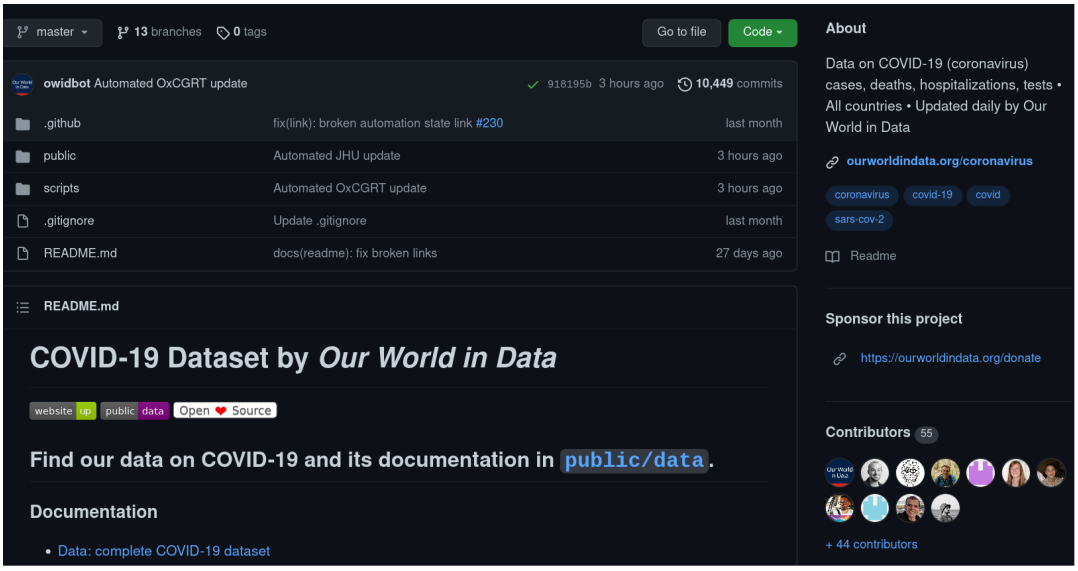

Ilustración 6: Repositorio de Datos Abiertos sobre el COVID-19 de la organización Our World in Data (OWID).

Este formato quizá se haga menos atractivo que el anterior, pero gracias a una gran cantidad de usuarios y de organizaciones que operan de este modo, se pueden difundir y obtener Datos Abiertos muy útiles. GitHub se puede asemejar a **un explorador de archivos**: en este repositorio particular, si entramos en la carpeta "*public*" y luego en "*data*", ya podemos empezar a ver algunos archivos CSV como los de antes. Hay que tener en cuenta que cada usuario organiza sus archivos de distinta manera, así que en estos casos habrá que leer bien el archivo "*README.md*" que suele dar información sobre el repositorio de datos.

El objetivo al final es que los Datos Abiertos sean accesibles al mayor número de personas posible, en el mayor número de formatos posible; y que sea fácil subir estos datos, de manera homogénea y utilizables por cualquiera.

6. Preguntas fundamentales

Como objetivo de proyecto queremos contestarnos las siguientes preguntas:

- Dentro del sector Salud ¿qué datos abiertos están disponibles a nivel mundial que puedan ser utilizados por algoritmos de IA?
    - Buscamos fuente de datos abiertas de:
        - Pacientes con datos estructurados y texto.
        - Fuentes de datos de radiografías y/o pruebas radiológicas, como mamografías.

    - Documentación de medicamentos.
    - Guías de enfermedades.
    - Datos demográficos de salud.
    - Estándares de referencia.

- ¿Existe una manera sencilla de catalogar las fuentes de datos encontradas? ¿Qué atributos deberíamos utilizar para describirlas?
- ¿Cuáles son las 3 fuentes de datos más relevantes que se han encontrado?
    - Por contenido.
    - Por volumen de datos.
    - Por facilidad de utilización.
- Seleccionando una fuente de datos concreta:
    - ¿Puedo describir la información que contiene solo ejecutando *queries* y visualizando gráficos?
    - ¿Hay suficientes datos históricos que nos ayuden a entrenar un modelo?
    - ¿Cuál sería el objetivo que queremos alcanzar con el modelo?

7. Programa

El programa tiene dos objetivos:

- Aprender a identificar y elaborar los datos.
- Aprender a procesar los datos utilizando la analítica descriptiva y predictiva.

Las actividades a realizar serían las siguientes:

- Búsqueda de datos abiertos de salud en internet:
    - Internet es muy amplia, hay que definir una estrategia de búsqueda.
    - Aplicar la estrategia e ir recolectando las fuentes de datos identificadas, dejando almacenado en un soporte digital: el nombre, la URL para llegar, nº de registros y demás atributos que ayuden a catalogar la información. Utilizar una hoja Excel para ir almacenando la información
    - Validar las fuentes de datos encontradas: ¿son datos abiertos?, ¿son de salud?, ¿están dentro del tipo de datos requerido? Y, si no lo está, ¿por qué nos parece interesante?

- Formarnos en:
  - Python
  - El uso de Notebooks de JUPYTER
  - Manejo de bases de datos
- Catalogar las fuentes de datos encontradas:
  - Diseñar la forma en que se va a catalogar las fuentes de datos encontradas
  - Definir una base de datos
  - Construir un notebook que lea de la hoja Excel y cargue la base de datos
  - Construir una *query* para seleccionar las 3 fuentes de datos más relevantes
  - Construir una presentación que explique:
    - Qué fuentes de datos se han encontrado
    - Cómo se han catalogado
    - La selección de las tres fuentes de datos más relevantes
- Utilizar el Notebook para:
  - Seleccionar una fuente de datos concreta
  - Cargar los datos de esta fuente de datos
  - Realizar las *queries* necesarias para describir mediante gráficos la información que contienen
  - Aplicar modelos con apoyo del investigador.

8. ¿Cómo y dónde buscar datos?

A continuación, se listan unas cuantas plataformas y repositorios de Datos Abiertos que pueden ser de utilidad para iniciar la búsqueda. Hay que tener en cuenta que el **objetivo final** de la búsqueda es haber conseguido uno o varios **archivos de datos** que pueden ser de formato **CSV, TSV, JSON** o **TXT**, este último siempre que sea texto separado por comas, puntos y comas o tabuladores, o cualquier otro archivo de formato similar.

Los recursos nacionales más fiables están en:

- **Iniciativa de Datos Abiertos del Gobierno de España:** https://datos.gob.es/
  - Hay que acceder a través de: Catálogo de datos → Conjuntos de datos.
- **Página del Instituto Nacional de Estadística:** https://www.ine.es/
- **Portal de Datos Abiertos del Ayuntamiento de Madrid:** https://datos.madrid.es/

Existen recursos internacionales igualmente interesantes (algunos en inglés):

- **Portal Oficial de Datos Europeos:** https://data.europa.eu/es
- **Eurostat:** https://ec.europa.eu/eurostat/
- **Banco Mundial de Datos Abiertos:** https://datos.bancomundial.org/
- **Datahub.io:** https://datahub.io/
- **Kaggle:** https://www.kaggle.com/

Para los más avanzados, existen repositorios llevados por usuarios con datos de lo más variados (generalmente en inglés):

- **Datos del COVID-19 por OWID** (ya presentado): https://github.com/owid/covid-19-data
- **Colección de Datasets de OWID:** https://github.com/owid/owid-datasets
- **Datos recopilados por Google Trends:** https://github.com/GoogleTrends/data
- **Datasets indexados por AwesomeData:** https://github.com/awesomedata/awesome-public-datasets
- **Buscador de GitHub**: https://github.com/. En general, se pueden encontrar muchos repositorios de diferentes áreas de *Open Data* si se busca por el término en la propia página de GitHub.

Por último, merece especial mención el servicio de Google de **buscador de datos**, encontrado en https://datasetsearch.research.google.com/. Como pista para utilizarlo para el propósito del programa, al buscar cualquier término, se puede activar la opción “Tabular” en el “Formato de descarga”:

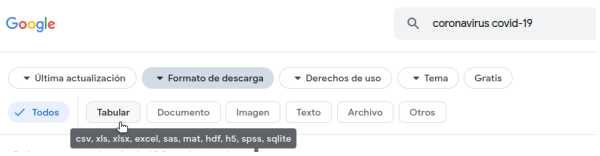

Ilustración 7: Filtros de "Google Research Dataset"

Bibliografía

- The Open Data Handbook. (2021). Open Data Handbook. https://opendatahandbook.org/ (consultado en 21/09/2021).
- The 2007 Workshop. (2007, 8 de diciembre). *The 8 Principles of Open Government Data*. opengovdata. https://opengovdata.org/ (consultado en 21/09/2021).
- Datos abiertos. (2021, 18 de julio). *Wikipedia, La enciclopedia libre*. Fecha de consulta: septiembre 21, 2021 desde https://es.wikipedia.org/w/index.php?title=Datos_abiertos&oldid=137089280.
- Rebaque, B. R. (2019, 20 junio). *Sobre Datos Abiertos: definición, principios, tipos y ejemplos para Educación*. Ciberimaginario. https://ciberimaginario.es/2018/07/25/sobre-datos-abiertos/ (consultado en 21/09/2021).
- Creative Commons. *About The Licenses*. https://creativecommons.org/licenses/ (consultado en 22/09/2021).
- The Open Data Handbook. (2021). *Conformant Licenses*. https://opendefinition.org/licenses/ (consultado en 22/09/2021).
- Warner, T. (2016, 10 mayo). *Guide to Open Data: Using it, Sharing it, and Creating a Portal*. Safe Software Inc. https://www.safe.com/blog/2016/05/open-data-portals/ (consultado en 22/09/2021).
- Junta de Castilla y León. (2021, 13 de septiembre). *Relación de bibliobuses de Castilla y León*. datos.gob.es. https://datos.gob.es/es/catalogo/a07002862-relacion-de-bibliobuses-de-castilla-y-leon (consultado en 22/09/2021).
- Ayuntamiento de Madrid. (2021). *Condiciones de uso*. datos.madrid.es. https://datos.madrid.es/portal/site/egob/menuitem.400a817358ce98c34e937436a8a409a0/?vgnextoid=b4c412b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextchannel=b4c412b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default (consultado en 22/09/2021).
- OWID. (2021). *COVID-19 Data*. GitHub. https://github.com/owid/covid-19-data (consultado en 22/09/2021).
- The ODI. (2020, 3 de julio). *Covid-19: Identifying and managing ethical issues around data*. https://theodi.org/article/covid-19-identifying-and-managing-ethical-issues-around-data/ (consultado en 22/09/2021).
- Edmunds, S. C. (2021). *Hong Kong/China - open sourcing genomes / crowdsourcing killer outbreaks*. The Open Data Handbook.

https://opendatahandbook.org/value-stories/en/open-sourcing-genomes/ (consultado en 22/09/2021).

- Open Data For Development Network. (2021). *Index*. https://opendataimpactmap.org/index (consultado en 23/09/2021).
- Gobierno de España. (2021, 9 septiembre). *¿Cómo se utilizan los datos abiertos en el sector salud y bienestar?* datos.gob.es. https://datos.gob.es/es/blog/como-se-utilizan-los-datos-abiertos-en-el-sector-salud-y-bienestar (consultado en 23/09/2021).
- Yozwiak, N., Schaffner, S. & Sabeti, P. Data sharing: Make outbreak research open access. *Nature* **518,** 477–479 (2015). https://doi.org/10.1038/518477a (consultado en 23/09).
- Sáez-Domingo, D. (2021). *Espacios de Datos en España: visión y casos de uso.* Jornada Espacios de Datos en Entornos Federados – VIII Asamblea General, 24 de septiembre de 2021. https://www.planetic.es/sites/planetic.es/files/public/content-files/presentations/2021/06.%20Daniel%20Saez_Espacios%20de%20Datos%20PLANETIC.pdf