Universidad Nacional Autónoma de México

INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

Licenciatura en Ciencia de Datos

Minería de Datos

Tarea 2

Esquema de datos del banco de COVID-19

Integrante:

- Villalón Pineda Luis Enrique

¿Cómo se llegó al banco de datos final?

1) El origen de los datos: múltiples fuentes confiables Our World in Data (OWID) no se limita a una sola fuente de información. Para construir su base de datos integral sobre COVID-19, el equipo combina datos de diferentes orígenes según el tipo de información:
    - Para casos confirmados y fallecimientos, confían en el repositorio de Johns Hopkins CSSE, actualizando esta información de manera diaria para mantener la precisión.
    - Para las métricas más específicas como pruebas diagnósticas, campañas de vacunación y hospitalizaciones, el equipo de OWID tomo un enfoque más directo: van país por país recolectando datos directamente desde las fuentes oficiales de cada gobierno.

      Pero no se detienen ahí. También integran información complementaria que ayuda a entender el panorama completo: la tasa de reproducción del virus (R), las políticas públicas implementadas por cada gobierno (utilizando el Oxford COVID-19 Government Response Tracker), y variables contextuales importantes de organizaciones internacionales como la ONU y el Banco Mundial.

2) La herramienta que hace posible todo: "cowidev"

   Para manejar esta enorme cantidad de información de manera eficiente, OWID desarrolló su propia herramienta interna llamada "cowidev". Esta librería funciona como una navaja suiza digital, con comandos específicos que automatizan todo el proceso:

    - `cowid vax get` para obtener datos de vacunación
    - `cowid test get` para extraer información de pruebas
    - Y muchos otros comandos especializados

   Esta herramienta se encarga de tres tareas fundamentales: extraer los datos (ya sea descargándolos o extrayéndolos de sitios web), transformarlos al formato necesario, y generar conjuntos de datos intermedios listos para el siguiente paso.

3) Dando orden al caos: procesamiento y normalización

   Una vez que tienen todos estos datos en bruto, comienza el trabajo de orfebrería digital. Cada conjunto de datos pasa por un proceso de refinamiento que incluye:

- Estandarización de nombres: convertir "Estados Unidos", "USA" y "United States" en una nomenclatura uniforme
- Conversión de unidades: asegurar que todos los números estén en las mismas escalas
- Cálculo de indicadores derivados: como las cifras per cápita que permiten comparaciones justas entre países
- Adición de metadatos: información descriptiva que ayuda a entender qué representa cada dato

4) La gran fusión: creando el dataset global

   Aquí es donde la magia realmente sucede. Todos esos datasets individuales —casos, muertes, vacunaciones, pruebas, hospitalizaciones, tasas de reproducción, políticas públicas y otros indicadores— se combinan en un solo conjunto de datos global que se actualiza religiosamente cada día.

5) Compartiendo el conocimiento: acceso abierto para todos

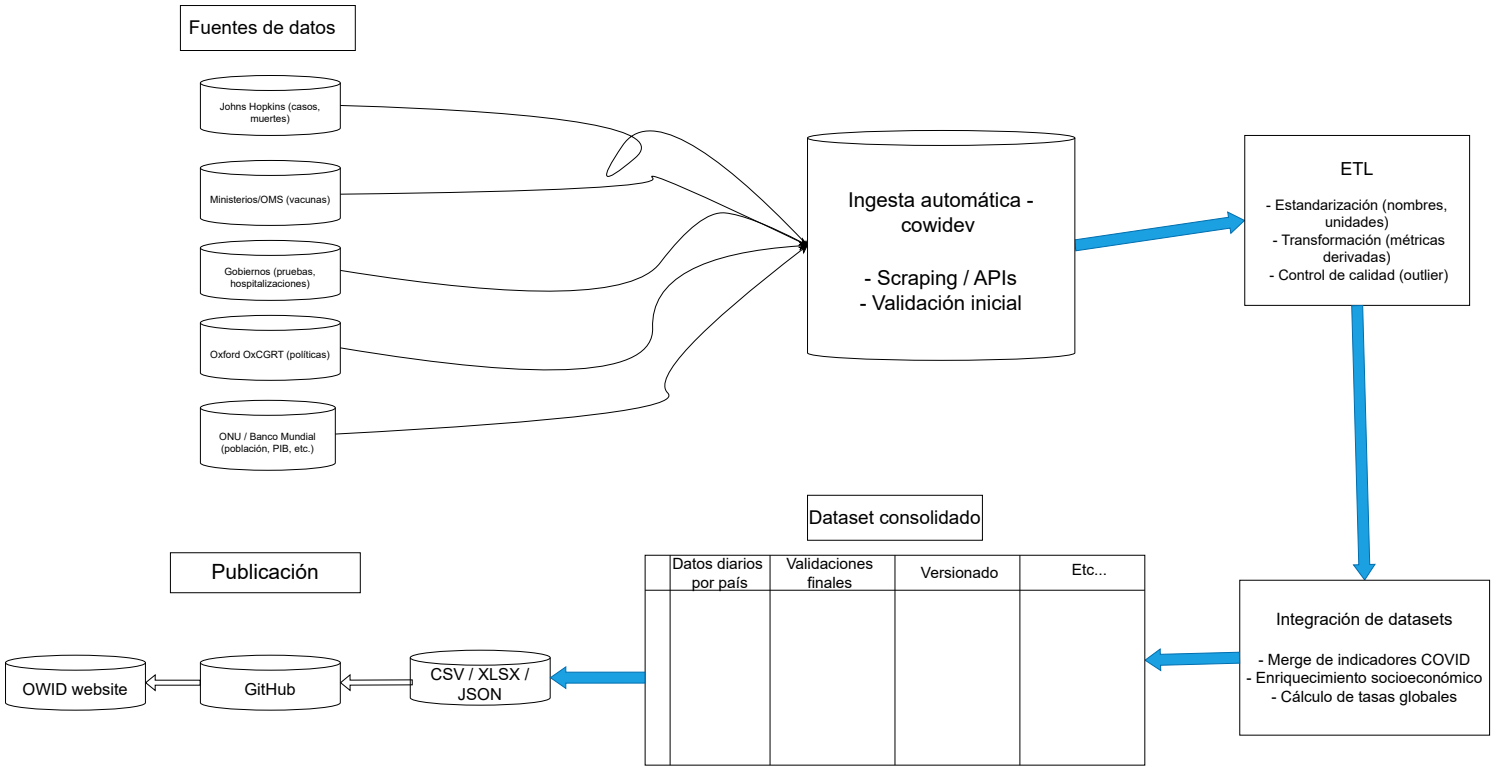
   El resultado final está disponible para cualquiera que lo necesite. OWID publica el dataset consolidado en múltiples formatos (CSV, XLSX y JSON) con acceso completamente abierto. El formato JSON está especialmente bien organizado, estructurado por país usando códigos ISO, con datos diarios e información estática de cada nación.

   Y lo mejor de todo: el código completo y todos los scripts están disponibles en su repositorio público de GitHub. Esto significa total transparencia sobre cómo se procesan los datos.

6) Una historia de evolución constante

   Este sistema no surgió de la noche a la mañana. OWID comenzó a trabajar con datos de COVID-19 desde los primeros días de la pandemia, en marzo de 2020, y ha ido desarrollando y perfeccionando este proceso conforme la situación mundial evoluciona.

   Aunque detuvieron las actualizaciones del dataset de pruebas en junio de 2022 (cuando muchos países redujeron sus programas de testing), el resto de los indicadores continúa actualizándose diariamente, proporcionando una ventana confiable a la evolución global de la pandemia.


Fuentes de datos
Johns Hopkins (casos, muertes)
Ministerios/OMS (vacunas)
Gobiernos (pruebas, hospitalizaciones)
Oxford OxCGRT (políticas)
ONU / Banco Mundial (población, PIB, etc.)
Ingesta automática - cowidev
- Scraping / APIs
- Validación inicial
ETL
- Estandarización (nombres, unidades)
- Transformación (métricas derivadas)
- Control de calidad (outlier)
Dataset consolidado
Datos diarios por país
Validaciones finales
Versionado
Etc...
Integración de datasets
- Merge de indicadores COVID
- Enriquecimiento socioeconómico
- Cálculo de tasas globales
Publicación
OWID website
GitHub
CSV / XLSX / JSON