Módulo 1: Formatos abiertos y cerrados

[00:00:12] **Sandra Crucianelli** Hola a todos. Vamos a continuar este primer módulo hablando de los formatos de los datos. Todos sabemos en algún momento de nuestra carrera periodística que nos hemos topado en la web con distintas formas en la que se presentan los datos. Me refiero básicamente a los formatos. Y hay formatos abiertos y hay formatos cerrados. En periodismo de datos trabajamos con formatos abiertos. ¿Qué significa un formato abierto? Es una estructura de datos, está libre de copyright, ustedes la pueden descargar fácilmente en su computadora, pueden procesar de manera relativamente fácil esa información, cruzarla con otros datos. Es decir, tienen un nivel de disponibilidad y de accesibilidad mayor al de otros tipos de formatos en los que los datos están generalmente encerrados o atrapados adentro de un formato. Voy a compartir pantalla para mostrarles algunos ejemplos de lo que sería un formato cerrado.

[00:01:19] **Sandra Crucianelli** Esto que están viendo ustedes acá este momento es un documento, un clásico ejemplo de un PDF en el que la información... Bueno, para empezar es un PDF que tiene logos, que tiene sellos, que tiene firmas, evidentemente uno se da cuenta al verlo que esto proviene de una imagen. Por ejemplo, si yo quisiera copiar, estoy marcando ahora como si fuera a copiar, intentar pegar esto en una plantilla de Word resulta imposible, porque fíjense que cuando hago clic sobre el botón derecho para intentar llevarme esos datos a otro formato, me dice "copiar imagen", es decir no me permite copiar texto. Y lo mismo si yo voy bastante más abajo, aquí tengo una tabla. Supongamos que necesito procesar esta tabla. Necesito llevarme los datos contenidos en esta tabla a un Excel. Bueno, de la misma manera no puedo llevarme estos datos porque estos datos vienen de una imagen y están encerrados. El primer nivel de cerramiento es la imagen, porque provienen de una imagen. Sí esto es así para todas las partes del documento. Y el segundo nivel de cerramiento que tiene este documento es que está inserto en un PDF que es un formato que tiene una licencia, tienen la licencia del programa Adobe. Por lo tanto, este es un clásico ejemplo de formatos cerrados, PDF en formatos cerrados.

[00:02:52] **Sandra Crucianelli** Hay otros PDF, como este por ejemplo, que no es tan, tan cerrado, por ejemplo, esta tabla. Esta tabla corresponde a las personas que fueron vacunadas en lo que se conoció en Argentina como "el vacunagate", personas que tuvieron privilegios a la hora de vacunarse. Este es un clásico ejemplo de un PDF en el que la información no proviene de una imagen. Está claro que esto proviene de una tabla que ha sido inserta en Word, pero está cerrado porque al estar bajo licencia de Adobe en un PDF no me permite el procesamiento de una manera más fácil. Igualmente, yo cuando marco los atributos de esta tabla y con el botón derecho del mouse intento copiar, sí me permite -- fíjense que si yo voy a un Excel como estoy yendo ahora, los puedo pegar y los datos como que fueron liberados. Una vez que los tengo en Excel o en algún otro formato de hoja de cálculo, puedo decir que los datos han estado liberados de ese formato en PDF, que no es un formato abierto. Entonces este sería el ejemplo de un PDF y de un documento semiabierto. No está completamente abierto.

[00:04:13] **Sandra Crucianelli** ¿Qué hubiéramos necesitado para que este documento fuera un formato abierto? Que nos lo hubieran compartido en un Excel, por ejemplo, o en algún otro formato, como por ejemplo el CSV. Acá lo que ustedes ven en los sitios de datos abiertos, este es el sitio de datos abiertos de Argentina. Este es el sitio de datos abiertos de Colombia. Y así, si yo voy buscando en sitios de datos abiertos, voy a encontrar la misma estructura. Es decir, por ejemplo, yo voy a hacer clic por un tema, vamos a suponer un tema que tiene que ver con transporte. Si ustedes ven a la derecha,

en la etiqueta que está en amarillo, van a ver tres letras CSV. Esto es un formato abierto, tal vez poco conocido para muchos de ustedes, pero es la forma en la que generalmente se están compartiendo datos a nivel global en plataformas de gobierno abierto. ¿Qué significa CSV? Un CSV es una -- ¿ven? Ahí está el CSV con la etiqueta en amarillo -- es un archivo por sus siglas en inglés Comma Separated Values, ¿sí?, valores separados por comas.

[00:05:28] **Sandra Crucianelli** Les voy a mostrar un ejemplo, aquí tenemos un CSV. Yo me he descargado del sitio abierto, del sitio de gobierno abierto de mi país, lo que sería un CSV clásico es un un archivo en el que la información no está estructurada, pero es fácil de estructurar. Una vez que ustedes tienen un archivo en este formato, lo abren usando la aplicación Excel. Lo primero que deben hacer es guardarlo como Excel, ¿sí?, para que puedan trabajarlo mejor bajo ese formato. Yo aquí estoy en CSV, voy a proceder a guardarlo como un Excel, lo guardo como Excel para poder trabajar con mayor facilidad. Y para convertir este archivo que está en CSV, que es el formato que más se está utilizando a nivel mundial para compartir grandes volúmenes de datos, el procedimiento en Excel es bastante sencillo: sólo tienen que marcar la primera columna, ir a la pestaña aquí arriba estoy marcando con el puntero datos, hacer clic donde se indica texto en columna -- le estoy indicando a Excel que me separe los atributos en columnas, la aplicación me pregunta si quiero que los datos estén delimitados, sí quiero que estén, por eso dejo marcado delimitados --- hago clic en siguiente y como veo que esos datos están separados no por barras, no por espacios, sino por comas, marco la coma (ya puedo previsualizar aquí que me va a separar en columnas esos datos), hago clic en siguiente. Le digo que me le dé formato a los datos en columnas y hago clic en Finalizar. Una vez que hice clic, fíjense que me tomó apenas unos segundos, yo acá tengo la tabla perfectamente organizada y puedo marcar para ver mejor los encabezados. Yo siempre recomiendo que hagan esto al principio, porque de esta manera yo puedo ver mucho mejor qué clase de datos tengo y cómo los tengo. ¿Ven? Ya esto es una tabla, esta es una tabla sobre aplicaciones de vacunas, primera dosis y segunda dosis por jurisdicción. Entonces yo puedo borrar lo que creo que no me interesa. Por ejemplo, el código de cada provincia lo elimino porque no es de mi interés, yo no voy a trabajar con esa columna, y dejo solamente lo que voy a utilizar. Y esto es muy útil tenerlo de esta manera porque yo puedo marcar estas columnas, puedo -yendo a la pestaña datos- aplicar un filtro, y buscar solamente por ejemplo, aquí voy a buscar solamente por tipo de vacuna. Digo bueno, devolveme, le digo al Excel, ¿no?, le doy la indicación, devolveme todo lo que es AstraZeneca y ahí me aparece en todas las aplicaciones de un tipo de vacuna. Y si no quiero visualizar esta voy a ir a otra, Sinopharm, entonces me va a devolver sólo lo que yo le pedí.

[00:08:41] **Sandra Crucianelli** Y esta es la manera de trabajar con CSV que son formatos abiertos que me permiten trabajar con mucha mayor libertad a la hora de moverme con los datos. Es decir, Excel y CSV van a ser nuestras principales herramientas. Vamos a encontrar en muchas bases de datos que vamos a ver en la próxima clase, archivos en CSV y archivos en Excel. Una vez que los encontramos en Excel, que los localizamos en ese formato, es mucho más sencillo de trabajar. Podemos descargarlo directamente a nuestra computadora y comenzar a trabajar con ellos. Cuando encontramos los archivos en formato CSV, es Comma Separated Values - valores separados por comas-, en ese caso lo que hacemos es proceder a la conversión de CSV en Excel, como yo recién les conté, es un procedimiento relativamente sencillo. Y yo he encontrado historias de periodistas que me dicen "uy, descargué un archivo que era un CSV y cuando lo abrí no supe qué hacer". Bueno, esta es una pregunta muy frecuente en periodismo de datos porque no saben cómo empezar hasta que lo tienen en un formato abierto, ¿no?, hasta

que lo lo pueden estructurar en columnas y pasar de ese formato cerrado que tenían originalmente, como puede ser un CSV a un Excel.

[00:10:08] **Sandra Crucianelli** En el Módulo 3 vamos a ver técnicas de scraping. Significa cómo extraer los datos cuando están encerrados, por ejemplo en un PDF que es la situación más común de todas. Pero no se preocupen. Vamos paso a paso. Aquí les presenté los formatos abiertos y los formatos cerrados. Ahora, en la próxima semana, en el próximo módulo vamos a ver cómo encontrar datos abiertos en la web, y no solamente en sitios gubernamentales, sino también en bases de datos. ¡Hasta la próxima!