Módulo 2: Cómo buscar datos en la Web

[00:00:12] **Sandra Crucianelli** ¡Hola a todos! Bienvenidos al segundo módulo de este curso masivo en línea de Introducción al periodismo de datos. Esta semana nos vamos a ocupar de dos temas importantes: de cómo buscar datos, cómo encontrarlos en la web. En la autopista informática circulan millones y millones de datos. Un volumen impresionante, casi diría, muy difícil de medir. Sin embargo, hay técnicas que nos ayudarán a esto. Pero déjenme compartirles un par de consejos que tienen que ver con el ambiente en el que trabajamos. ¿Ven este? ¿Sí? Es un mouse. Este va a ser su principal aliado cuando trabajen con volúmenes de datos muy grandes. Yo sé que muchos de ustedes trabajan con laptops y con la mano sobre el pad, con los dedos sobre el pad, van haciendo sus búsquedas y sus trabajos. Es complicado trabajar de esta manera cuando tenemos volúmenes de datos muy grandes y necesitamos trabajar con celdas. Por eso el puntero del mouse nos sirve muchísimo para ubicar precisamente una celda. De repente tenemos que incorporar una fórmula o tenemos que trabajar con conjuntos muy grandes de celdas en filas, por ejemplo, están contenidas en un Excel y tienen decenas y decenas de celdas. Lo mismo trabajamos con un grupo grande de columnas.

[00:01:36] **Sandra Crucianelli** Entonces yo siempre recomiendo tres cosas: el mouse para trabajar si es un mini mouse, mejor porque se opera, por lo menos para mí, de manera mucho más fácil; una buena silla ergonómica porque seguramente van a pasar horas trabajando en un volumen de datos grande -- muchas veces estamos procesando datos durante siete, ocho horas--, necesitamos la silla cómoda; y también una pantalla grande. Por eso cuando me preguntan "Sandra, ¿vos con qué computadora trabajás?". Pueden trabajar con cualquiera, siempre que sea, tengan una pantalla lo suficientemente cómoda. Yo en este momento trabajo sobre una computadora escritorio que tiene un monitor de 21 pulgadas. Eso me permite visualizar muy bien en un Excel aquellos elementos que quiero destacar, que quiero mover, que quiero correr de un lado a otro, hacer comparaciones. Es decir, cuando uno está trabajando con volúmenes de datos grandes, necesita trabajar no solamente con un buen hardware, sino también con elementos que nos den comodidad. Estos son los que me dan comodidad a mí. Después por supuesto, cada uno puede usar la tecnología que quiera.

[00:02:47] **Sandra Crucianelli** En esta primera parte de esta semana vamos a abordar cómo encontrar datos en la web. Porque sí ya sé, todos me dicen "pero Sandra yo ya sé usar Google". Sí yo sé que todos ustedes lo usan y seguramente lo usan muy bien, pero refrescar algunos conocimientos no viene mal en especial cuando trabajamos con operadores que restringen las búsquedas. Vamos a estar buscando Excel básicamente o tablas contenidas en PDF básicamente porque nos hemos dado cuenta, de acuerdo a la experiencia que hemos recopilado en los últimos diez años, es que la mayoría de los documentos importantes de los gobiernos, de fuentes oficiales o de organismos internacionales están en PDF y en el mejor de los casos, están CSV o en Excel. Por lo tanto, vamos a refinar esas búsquedas para encontrar con mayor precisión lo que estamos buscando.

[00:03:40] **Sandra Crucianelli** Vamos a comenzar con lo que todos conocemos en cuanto a los buscadores que es Google. Bueno, ustedes han trabajado con Google desde sus inicios, imagino en este campo y podemos hacer muchas búsquedas desde la plantilla general de Google. Por ejemplo, si estoy buscando un presupuesto para el caso de Colombia, si yo coloco solamente las dos palabras presupuesto Colombia me voy a encontrar con 95 millones 400 mil resultados. Es una búsqueda muy inespecífica. Yo podría haber colocado comillas y haber colocado la expresión "presupuesto de Colombia",

con lo cual se achica mucho el resultado. Solo entro a obtener 58 resultados porque apliqué un filtro que es muy restrictivo, es decir, usé las comillas para la frase exacta. Ahora también podría haber hecho otra búsqueda. Por ejemplo, si yo saco las comillas, que como les digo es un elemento muy restrictivo, y lo combino con el sitio, por ejemplo el sitio es site en inglés, dos puntos -- estoy aplicando un operador punto gov.co. ¿Qué hice en este caso? En este caso dije bueno, presupuesto y Colombia son dos palabras separadas sin comas y le dije al buscador "devolveme todos aquellos documentos contenidos en sitios gubernamentales de Colombia. ¿Por qué? Porque coloqué site dos puntos punto gov.co [site:gov.co]. Esta es la raíz del dominio, en este caso un dominio gubernamental, para Colombia. Podía haber hecho lo mismo para cualquier país. Aquí la búsqueda del 95 millones de resultado que yo tenía antes, se me restringió bastante y quedó en 4 millones 400 mil. Igual es un número alto. Entonces podría aplicar otro nivel de restricción y poner File Type, filetype -- tipo de archivo -- dos puntos PDF. Entonces lo que le estoy diciendo al buscador -- que cree que soy un robot, pero no lo soy--, es que estoy buscando PDFs, porque he colocado el operador filetype dos puntos PDF dentro del sitio de gobierno de Colombia. Y podría ya esta altura haber eliminado Colombia porque si estoy buscando dentro del sitio gubernamental de Colombia, busco presupuestos en PDF publicados en una web gubernamental, en este caso, la colombiana. Podría haber hecho lo mismo para cualquier país.

[00:06:44] **Sandra Crucianelli** Ahora supongamos que yo elimino estos parámetros de búsqueda y me voy a una ruedita que hay aquí a la derecha en el extremo superior del buscador. Esa ruedita cuando clic, siéntanse curiosos de explorar todas las rueditas, íconos, opciones de herramientas porque uno nunca sabe lo que se va a encontrar. Esta es la configuración de la búsqueda y aquí hay una herramienta importante para todo periodista que es la búsqueda avanzada. La búsqueda avanzada es una plantilla de Google que les permite restringir por palabras, por combinaciones de palabras sin necesidad de usar comillas. Cuando están en esta plantilla de búsqueda avanzada no es necesario que coloquen comillas. Entonces yo aquí podría poner con la palabra presupuesto, podría ir ven que acá dice "luego restringe tus resultados por sitio y dominio", podría colocar la raíz de un dominio y voy a volver al caso anterior punto gov.co, y aquí más abajo -- yo lo estoy marcando -- me dice de qué formato, en qué formato quiero. Bueno, puedo pedirle que me dé en formato PDF y vuelvo al millón 410 mil resultados de la última búsqueda que hicimos. Ven que arriba en el resultado es la misma búsqueda. Presupuestos es la palabra clave. Site dos puntos, es decir, sitio, dos puntos punto gov.co y filetype, tipo de archivo, PDF. Ahora yo también podría volver a la búsqueda avanzada y decirle "OK, no, no quiero los PDF, quiero solamente los Excel". Entonces marco Excel y busco. Y efectivamente, hay 24.600 resultados que tienen un Excel. Es decir, son Excel que están alojados en distintos sitios del gobierno, en este caso de Colombia. Yo podría, como les dije antes, haber cambiado de país. Y pongo acá punto gob.ar, sería mi país, en este caso es con B larga. Y asegúrense de si en su país, en el país donde viven, la raíz del dominio es punto gov con V corta o punto gob con B larga. Mi país es con B larga, en Colombia es con V corta. Es decir, cada país tiene su propio uso. Y yo podría decirle acá "bueno devolveme los PDF que hay publicados dentro de este gobierno". Y también podría aplicar acá dentro de la opción de herramientas una restricción por fecha y decir "OK, devolveme solo el último año".

[00:09:37] **Sandra Crucianelli** Una cosa que no deben hacer en el en la planilla con la que están buscando dentro de la búsqueda avanzada es esto. ¿Sí? Colocar comillas acá. La comilla en realidad está expresada por el cuadro, por lo tanto, no es necesario colocarla en la plantilla de la búsqueda avanzada. Lo mismo cuando quieren buscar dentro de un sitio o dominio. Vamos a suponer el sitio www.datos.gob.ar, si yo hago esto,

no voy a encontrar nada porque cometí un error. La triple doble v no va. Es decir, sólo se coloca el dominio entero. En este caso, coloqué el dominio del sitio de gobierno abierto de Argentina, le voy a cambiar acá la fecha -- ah tengo una restricción acá que me había quedado -- vuelvo a la búsqueda avanzada, y pongo presupuesto y aquí datos.punto.gob.ar. Sin las tres doble v [www]. Y me da 92 resultados.

[00:10:55] **Sandra Crucianelli** Esta es una manera fácil de buscar. Como les digo, utilizando esta plantilla ustedes pueden jugar con esta plantilla y hacer restricciones de idioma, de región, del momento, es decir lo mismo que hacíamos con la opción de las herramientas, decidir el momento de la publicación. Y lo más importante, por lo menos en la técnica que yo utilizo, es restringiendo por dominio que puedo colocar el dominio completo -- en este caso coloqué datos.gob.ar--, pero también podría haber usado solamente la raíz cuando voy a la pesca de un documento. Es decir no sé en qué sitio de gobierno específico está, sino sé que está en una oficina gubernamental, pero no exactamente en cuál, entonces coloco la raíz del dominio. Y generalmente hago dos búsquedas la que es en PDF porque generalmente da muchos resultados. Y si tengo, estoy muy convencida de que los datos van a estar y quiero tener suerte porque van a estar en Excel. En este caso encontré 1.250 Excel dentro de oficinas de gobierno. En este caso de mi país.

[00:12:11] **Sandra Crucianelli** Esto sería Google. Tienen que jugar, tienen que animarse a entrar a la plantilla de búsqueda avanzada porque les va a permitir encontrar mucha información. Recuerden que cuando ustedes colocan una palabra, la forma de entrar a la búsqueda avanzada es a través del acceso de preferencias aquí a la derecha, en el extremo superior derecho, y allí encuentran la plantilla con la cual ustedes pueden empezar a hacer sus búsquedas, restringiendo por formato y restringiendo por dominio.

[00:12:49] **Sandra Crucianelli** Pero no sólo existe Google, también hay otros buscadores. Déjenme decirles que Bing también es una excelente opción. Cuando ustedes no encuentran y no encuentran algo, prueben otros buscadores, en este caso Bing. Voy a colocar corrupción en Panamá, y tengo 578 mil resultados. Pero podría haber puesto las comillas para la frase exacta y se restringe un poco la búsqueda de esa manera. Aunque aquí lo que yo sugiero es para que no esté tan restringida la búsqueda con el uso de las comillas es que puedan trabajar con los operadores, como les comenté en la primera parte de este video. Por ejemplo, presupuesto Colombia sin comillas, porque son dos palabras separadas, y por tipo de archivo (filetype) punto XLS, y por la raíz del dominio, es decir, punto gov.co, en el caso de Colombia. Aquí tengo 12.400 resultados que son y fíjense qué interesante me devuelve una cantidad muy importante de archivos en Excel, que es lo que siempre estamos buscando, son archivos de la Fiscalía General de la Nación, también hay del Ministerio de Economía y así pueden encontrar un gran volumen de información utilizando estos operadores. Yo podría haberle dicho "no, dame solamente los que están en PDF", por el tipo de archivo PDF, en vez de que me devuelva sólo con Excel, digamos un formato más cerrado, porque a veces mucha de la información que buscamos obviamente la buscamos para que esté en un formato descargable, como sería un Excel que nos permita procesarlo en columnas. Pero da la casualidad que en muchos de nuestros países de América Latina eso no existe literalmente en Excel, sino que todos esos Excel están encerrados en un PDF, entonces hay que salir a buscarlos en PDF. Y restringir por tipo de archivo, filetype, o por raíz del dominio site punto gov punto el país en el que ustedes están viviendo y trabajando. Es una manera muy efectiva de hacer las búsquedas.

[00:15:17] **Sandra Crucianelli** Pero, pero hay mucho más en búsquedas. Es decir, una buena cantidad de información la podemos recolectar haciendo búsquedas específicas, búsquedas avanzadas o como les comenté recién sobre Bing aplicando los operadores, ¿no?, por tipo de archivo y por la raíz del dominio. Pero también hay un gran volumen de información que está como, no digo oculta, pero sí atrapada en bases de datos, entonces los buscadores no devuelven bien esos niveles de profundidad. Esas bases de datos están en un nivel bastante profundo de la web y vamos a necesitar accesarlas para poder ver qué información importante hay allí. Pero eso, en el próximo video.