Módulo 1: ¿Qué es el periodismo de datos?

[00:00:12] **Sandra Crucianelli** Comencemos este curso sobre Introducción al periodismo de datos, definiendo qué es. Los periodistas venimos usando datos desde hace décadas. En los años 60 se escribió un libro muy conocido por muchos de nosotros llamado "Periodismo de precisión", de Philip Meyer. Philip Meyer fue uno de los primeros periodistas a nivel mundial que comenzó a trabajar con tablas y él se dio cuenta que había un importante volumen de información que para poder ser interpretado necesitaba ser ordenado en tablas y así podía ver mejor una progresión, por ejemplo, una sucesión de hechos a lo largo del tiempo. Con la irrupción de las computadoras este trabajo se simplificó y lo que se conoció inicialmente como periodismo de precisión, luego pasó a conocerse como periodismo asistido por computadoras, o PAC, y es cuando entra la hoja de cálculo en nuestra vida. En los años 90, un grupo de periodistas en todo el mundo, por supuesto en América Latina también, comenzamos a trabajar con hojas de cálculo, comenzamos a trabajar básicamente con Excel y así pudimos llevar, por ejemplo, presupuestos para poder analizar a lo largo del tiempo la evolución de las partidas presupuestarias y una cantidad de información muy grande que estaba contenida en archivo, que también eran importantes en cuanto a su tamaño y que eran susceptibles de ser ordenados en una hoja de cálculo. Estos fueron los primeros pasos del periodismo de datos que se consolida en el año 2011, cuando se crea lo que se conoce como la Alianza de Gobierno Abierto. La Alianza de Gobierno Abierto es una organización que reúne a países de todas partes del mundo, en el que los países asumen compromisos para abrir sus datos.

[00:02:13] **Sandra Crucianelli** Pero, ¿qué significa en sí mismo el periodismo de datos o, como se lo conoce también, el periodismo de base de datos? Se trata del periodismo de investigación que conocemos desde siempre, con la particularidad de que el volumen de datos con el que estamos trabajando es muy grande. Y que además de que el volumen es grande, necesitamos del auxilio de ciertas tecnologías para poder procesar este gran volumen de información. Por ejemplo, las unidades de datos en todo el mundo están compuestas no solamente de periodistas, sino también de mineros de datos, lo que se conoce como Data Mining o minería de datos -- hoy es una especialización, es algo que se estudia en las universidades y hay posgrados sobre esto. Y no solamente mineros de datos, sino también hay programadores, desarrolladores, ingenieros en sistemas, porque los lenguajes de programación nos ayudan a lidiar con volúmenes de datos muy grandes que los periodistas, con nuestro conocimiento, no podríamos procesar.

[00:03:23] **Sandra Crucianelli** Les doy un ejemplo. En plena pandemia en todos los países han estado compartiendo datos abiertos sobre la evolución de las pandemias. Voy a compartir pantalla. Aquí estamos viendo lo que sería el sitio de gobierno abierto de mi país. Y así como hay sitios de gobierno abierto en mi país, lo hay prácticamente en casi todos los países del mundo. Entonces, por ejemplo, si ustedes quisieran ver información sobre COVID en alguno de estos portales, seguramente van a encontrar, por ejemplo aquí, "casos registrados en la República Argentina", lo que van a ver es un archivo muy grande que prácticamente, si bien está comprimido, no lo van a poder descargar en su computadora y se requiere de un lenguaje de programación. En este caso, al ser un volumen tan grande para poder no solamente descargarlo, sino también procesarlo y analizarlo, y en ese caso lo que los programadores hacen es crear una base de datos con esta información a la cual se le pueda hacer consultas. Y esto pasa con la mayoría de los sitios de gobierno abierto en todo el mundo, donde lo que hay es información en distinto tipo de archivos. Si ustedes buscan los conjuntos de datos están disponibles en cada uno de los países, van a ver la misma estructura. Una estructura en formato comprimido --

aquí tenemos el caso de ZIP --, pero también en otros formatos, como pueden ser Excel que es muy frecuente que se compartan archivos en Excel, y también aquí tenemos uno más abajo lo pueden ver -- este que el Registro Nacional de Parques Industriales, el formato en el que se ofrece este conjunto de datos es en Excel--, pero también vemos otros formatos como los comprimidos o esto aquí, las etiquetas en amarillo que ustedes pueden ver, lo que se llama un CSV. En el próximo video vamos a ver qué es un CSV y cómo lo lo podemos procesar.

[00:05:26] **Sandra Crucianelli** En general, la apertura de los archivos gubernamentales que se dio en Estados Unidos, en Reino Unido, en Australia, en Nueva Zelanda hace muchos años, ya más de diez años, fue el que ha incentivado el uso periodístico de estos datos. Y uno de los disparadores que hubo fue las filtraciones de Wikileaks, es decir, que fue una de las primeras grandes filtraciones que hubo a nivel mundial y de alguna manera se dieron los primeros pasos allí, en lo que sería el uso de grandes volúmenes de información para tratar encontrar verdades sociales importantes que nos llevan, nos conducen a un título y en el mejor de los casos, una primicia. Uno de los diarios, y que fue pionero en esta materia, fue The Guardian en el Reino Unido y él la compartió, este diario ha compartido con todos nosotros el proceso que utiliza en usar los datos públicos. Se parte generalmente de datos compartidos, es decir, de datos que encontramos en la web, que se han compartido a través de la web, sea a través de una filtración en una plataforma específica como puede ser la de Panama Papers o Paradise Papers, que el Consorcio Internacional de Periodistas de Investigación ha compartido un volumen enorme de información en la web -- lo vamos a ver cuando veamos en la semana próxima bases de datos. También podemos encontrar estos datos compartidos en plataformas de gobierno abierto, no solamente de los gobiernos, sino también de las organizaciones de la sociedad civil. Y también podemos encontrar datos a través de lo que hemos pedido, usando leyes de acceso a la información pública.

[00:07:07] **Sandra Crucianelli** Una vez que tenemos esos datos, tenemos que preguntar qué significan los datos. Es decir, si estos necesitan o no, generalmente sí, validarse y qué medida podemos usar para compararlos, para montar cambios en esos datos y qué otros conjuntos de datos podemos usar para cruzar esa información inicial que tenemos para encontrar una nueva. Todo esto se hace usando planillas de cálculos, hojas de cálculo como lo conocen ustedes seguramente, lo hacemos a través de Excel que es la herramienta de Microsoft que nos permite una singular cantidad de operaciones que podemos hacer, que son muy útiles a la hora de poner los datos en un contexto y darles un significado. Para esto necesitamos hacer cálculos con los datos. Los datos pueden estar medidos en diferentes unidades. Podemos tener columnas que no sean necesarias. Podemos tener celdas que están unidas o fusionadas. Podemos tener datos en formatos equivocados. Todo esto nos lleva a lo que sería el proceso de limpieza de datos, de normalización de los datos. Y una vez que los limpiamos, podemos hacer cálculos con esos datos, podemos recalcular si fuera necesario, verificar que esto se hace prácticamente en todos los procesos. Y todo esto nos lleva a que esos datos son susceptibles de ser montados en una plataforma de visualización que nos permite ver mejor ese volumen de datos, que es enorme --generalmente son volúmenes de datos muy grandes. Por lo tanto, lo que tenemos es una gran cantidad de datos, las definiciones y cómo se ha estado trabajando con grandes volúmenes de datos en Latinoamérica, lo encontramos en el Manual de Periodismo de Datos Iberoamericano, que yo les voy a compartir en nuestra plataforma de recursos donde ustedes van a poder leer sobre periodismo y nuevas tecnologías, sobre cómo encontrar historias en grandes volúmenes de datos. También en toda la experiencia que hemos tenido las salas de redacción, es decir, cómo hemos integrado a periodistas que nunca habían trabajado con hojas de

cálculo a la redacción para que comiencen a usar estos recursos. Y este manual también les muestra algunas experiencias de periodismo de datos en América Latina.

[00:09:36] Yo les puedo compartir lo que estamos haciendo en Infobae, yo soy la coordinadora de la unidad de datos de Infobae. Si ustedes buscan como Infobae Data, van a encontrar en Infobae Data un gran volumen de piezas periodísticas que fueron construidas a partir de volúmenes de grande..., volúmenes grandes de datos. Por ejemplo en estudios que miden la efectividad en el operativo de vacunación, como se hace en esta crónica que yo les estoy mostrando acá. En la que siempre hay una visualización interactiva, en este caso usamos un recurso, una plataforma denominada Flourish -- que también la vamos a ver en la última semana del curso--, y como característica de estos trabajos basados en datos, siempre al final del artículo -- ven aquí hay otra visualización interactiva, el tamaño de los datos era bastante grande e hicimos bastantes cruces de información con varias variables -- siempre al final del artículo lo que hacemos es no solamente compartir el análisis de esos datos, sino compartir el cómo se procesó la información. Es decir, contamos lo que hacemos, que descargamos y procesamos una base de datos, lo que hicimos y compartimos esa hoja de cálculo que era inicialmente un Excel, la compartimos a través de nuestra crónica mediante un link. Por ejemplo aquí en las columnas en amarillo para los datos originales que obtuvimos de esa base de datos y lo que están en otro color son datos que cruzamos, es decir que añadimos a nuestra base inicial. Por ejemplo, la población proyectada en este caso es un registro de sorteos de viviendas donde quisimos saber si el color político influía de alguna manera en la obtención de mayores beneficios para los beneficiarios. Y luego hicimos cálculos. Cálculos que tienen que ver con porcentajes o como este que hicimos en la columna G, que tiene que ver con el porcentaje de las personas beneficiadas con relación a las que se inscribieron. Y vimos que los gobiernos que estaban manejado por un determinado partido político tenían casi todos el 100 por 100 de positivos. Y después inscritos por cada 10 mil habitantes, que en este caso estamos analizando aquí los posibles beneficios a un plan de vivienda es una medida de proporciones, una tasa por cada 10 mil habitantes.

[00:12:28] Pero no se preocupen mucho por esto que puede parecer abrumador al principio, vamos a ir paso a paso y yo los quiero ir guiando en este proceso. Este curso justamente está pensado para todos aquellos que nunca han hecho periodismo de datos, que no tienen mucha experiencia de manejo de Excel. Por lo tanto, mi misión es ir guiándolos poco a poco para que puedan construir historias con datos. Nos vemos en el próximo video.