Periodismo de datos. Capítulo 01

Clasificación

Según la Global Editors Network el periodismo de datos se puede dividir en varias categorías, las cuales no son necesariamente excluyentes ya que usan métodos y protocolos comunes y tienden a ser complementarias.

Estas son las categorías:

- *Data-driven investigative journalism*
- *Data-driven applications*
- *Storytelling with data*
- *Data journalism website*

Según Wikipedia, el periodismo investigativo de datos (lo que llamaremos **"periodismo de datos"** por simplicidad) es un proceso periodístico que se basa en el análisis y filtrado de grandes bases de datos con el propósito de crear una noticia periodística.

Me voy a concentrar en esta categoría ya que exiten varios reportajes de periodismo de datos en esta línea que han ganado concursos internacionales:

- El diario La Nación (Argentina): Luego que el senado argentino publicara los reportes de sus gastos desde el 2004 en forma de archivos PDF, La Nación consiguió extraer, transformar, normalizar, tabular y estructurar los datos. Esto permitió encontrar gastos sospechosos e inusuales que terminaron en una investigación judicial del vice-presidente de la república Amado Boudou.
- Los periodistas Giannina Segnini y Ernesto Rivera analizaron las declaraciones juradas de los ministros del gobierno de Laura Chinchilla, Costa Rica. Encontraron que los ministros habían subvalorado el valor de sus casas para pagar menos impuestos. Luego que la investigación viera la luz, 7 ministros se apresuraron en corregir el valor de sus inmuebles.

Obtención de datos lenta

Si bien una investigación periodística inicia con un dateo ya sea de un "garganta profunda" o un Edward Snowden, a veces es necesario obtener los datos de manera independiente. A veces los datos son facilitados por el datero pero muchas veces los datos están disponibles en los portales de instituciones estatales, registros públicos, etc.

A veces la obtención de datos es forzosamente lenta. Por ejemplo cuando uno pide datos en la SUNARP. Otras veces los datos se pueden obtener a por montones y rápidamente si se tiene la ayuda de un hacker (un hacker ético, claro está; si no tienes hacker, consíguete un geek). Por ejemplo el Ministerio de Justicia tiene en su web todas las resoluciones ministeriales emitidas en formato PDF. Bajarse cada PDF implicaría hacer una búsqueda en su portal, seleccionar la resolución de qué día te quieres bajar y finalmente hacer click en "download". Durante el segundo gobierno aprista se emitieron 2,184 resoluciones y bajar todos estos PDFs manualmente, uno por uno, demoraría una eternidad.

Figure 1: Yes, I am your typical geek

Obtención de datos veloz

Pero lo bueno es que los archivos están almacenados de manera consistente. El nombre del cada archivo PDF consiste en la fecha en que se emitió la resolución (ddmmyy, osea día, mes y año):

`http://spij.minjus.gob.pe/Normas/textos/ddmmyyT.pdf`

Este trabajo es demasiado facil para un hacker ético. Solo basta escribir un programita de 9 líneas de código para bajarse TODAS las resoluciones:

https://gist.github.com/aniversarioperu/7071796

Una vez que el programa empieza a correr irá bajando cada PDF, un por uno, ya que el nombre de los archivos se puede constuir usando las fechas de un calendario. Este programa terminará su trabajo en unas cuantas horas sin necesidad que el periodista y/o hacker realicen actividad manual alguna.

Los geeks tienen muchas herramientas *open-source* disponibles para realizar sus actividades. Una herramienta que se usa mucho para descargar contenido desde las web se conoce como `curl`. Si se usa correctamente, este programa puede aparentar ser un usuario humano, ya que puede suministar nombre de usuario y contraseña a las páginas que lo requieran, puede lidiar con cookies, usar certificados para autenticación, usar proxies y muchas cosas más. Es algo así como la navaja suiza de los geeks para asuntos de descarga de datos.

Como ves, es muy ventajoso para un periodista de investigación estar asociado a uno o más hackers, o geeks, o nerds. Uno de los aportes principales de los hackers éticos al periodismo de datos es la rapidez. Al aprovechar de sus habilidades tecnológicas es posible acelerar la investigación sobre todo cuando hay que realizar actividades repetitivas. En el mundo digital, las actividades repetitivas deben ser ejecutadas por computadoras, no por humanos. Las computadoras son buenas para ejecutar tareas repetitivas ya que son infinitamente más rápidas que un humano.

Esas tareas repetitivas pueden ser automatizadas y ejecutadas por los hackers y sus computadoras. La labor del periodista es otra, es analizar qué datos son importantes de ser cosechados, qué otro tipo de datos deben ser asociados con el fin de obtener una historia. La labor analítica y de pensamiento crítico debe ser realizada por el periodista. Para esto es de vital importancia la experiencia e intuición del periodista.

TL;DR

Amigo, amiga periodista, si necesitas bajar cientos, miles de PDFs, imágenes o páginas de un sitio web, contáctate con tu geek más cercano. Hay altas probabilidades que el geek pueda usar sus habilidades para bajar lo que necesitas en un santiamén.