Periodismo de Datos, 2021-09-07

Anton-Bravo, Adolfo

Contents

1 Adolfo Antón Bravo

- Permitidme que me presente. Soy Adolfo Antón Bravo.

- Este es mi primer año como profesor asociado en la UC3M y espero que sea el inicio de muchos.

- He dado y doy clases en diversos másters. Para no hacer propaganda y por si alguien quiere conocer mi trabajo un poco más os muestro la página web infotics.es donde suelo poner todo lo que hago.

- También soy @adolflow en Twitter o @flowsta en Github.

- Mi recorrido por el Periodismo de datos comienza en 2013 cuando me encontraba realizando mi programa de doctorado y me tope con el periodismo de datos y Mar Cabra. A partir de ahí no os quiero aburrir y os remito a la web infotics.es.

2 ¿Qué es el Periodismo de datos?

Algunos comentarios sobre el periodismo de datos:

- Se dice que en Periodismo de Datos hay que aprender Excel. Me gustaría explicaros qué es Excel a continuación.

- El periodismo de datos moderno, del que bebemos actualmente, nace en 2006-2008 con una combinación de factores: abundancia de software de código abierto, HTML5 y Open Data. Hablaremos de ello más adelante.

- The Guardian es uno de los medios pioneros del periodismo de datos moderno.

- Como es una disciplina nueva también es una disciplina contenedor. Fundamentalmente hay tres áreas implicadas en el periodismo de datos:

    1. El periodismo, y solo puede haber periodismo si hay investigación.

    2. Los datos, es decir, registros electrónicos que han de ser tratados por ordenador.

    3. La visualización de datos: desde la Web hasta el papel pasando por la estadística, las distintas visualizaciones o la infografía.

2.1 Debate sobre el término

Esto se produjo en distintos momentos de los dos primeros días pero lo seguiremos recordando y ampliando.

- Aquí y ahora hablamos de periodismo de datos, y está bien.

- Cuando empezó en Europa y EE.UU. en 2008 se denominó periodismo guiado por datos. Suele ocurrir en inglés al menos con los saberes que, cuando se realizan de otra manera, se traslada ese proceso al nombre. Así por ejemplo en los 1990' yo hice cursos de Diseño gráfico asistido por ordenador (del *computer assisted graphic design*), ya que hasta entonces se hacía "diseño gráfico" de forma analógica.

- La irrupción de los portales de datos abiertos hizo que se denominara *Data Driven Journalism*, aunque luego se acortó a *Data Journalism*, pero ambos términos conviven.

- En EE.UU., por ejemplo, el precedente del periodismo de datos es el *precision journalism* o periodismo de precisión. José Luis Dader, catedrático de la Facultad de Ciencias de la Información de la UCM y que fue profesor durante el programa de doctorado, nos contaba cómo estuvo en EE.UU. para aprender esa disciplina y, al traerla a España y traducir el libro de Philip Meyer pensar si "periodismo de precisión" era un término apropiado. Pensó en "periodismo matemático", por su rigurosidad, pero pensaba que no se entendía bien; pensó en "periodismo informático", pero se iba a malinterpretar; o "periodismo científico", pero se iba a confundir con el periodismo que habla de ciencia. Al final se quedó con el original "periodismo de precisión" que al menos no inducía a errores y sí que hablaba de algo nuevo, sin por ello dejar de crear polémica ya que pareciera que el resto del periodismo no fuera preciso! Lo cierto es que se denominó así en EE.UU. para oponerse a un periodismo del estilo de Truman Capote.

- Ese periodismo de precisión no era la primera vez que en los EE.UU. se utilizaban ordenadores en periodismo. Veremos más adelante tanto el caso de Philip Meyer como el de la CBS de 1952. Este y otros usos de los ordenadores dieron en llamar a este periodismo como *Computer Assisted Reporting* o periodismo asistido por ordenador. Esta denominación pervive en la actualidad.

- Hay otras denominaciones en EE.UU. como la que la comunidad de computer assisted reporters ha creado: *News Nerdery* o los frikis de las redacciones.

- En Inglaterra, por ejemplo, la comunidad de periodistas de datos se ha denominado *Journocoders* o perioprogramadores.

- Y en EE.UU. también una comunidad pionera fue la denominada *Hacks and Hackers*. *Hacks* significa "hachazos" y se refiere al martilletear de teclear en la máquina de escribir.

- En Argentina, Sandra Crucianelli sigue hablando de "periodismo de bases de datos", y no le falta razón pues finalmente, para trabajar con los datos, alguna "base de datos" has de tener... aunque el concepto de base de datos también ha evolucionado.

3 Qué conocimientos previos tenéis

Comentáis tres tecnologías:

3.1 Wordpress

- No lo vamos a usar pero me gusta que lo citéis para ver si sabemos lo que es y todo lo que implica.

- Wordpress es un *CMS* (Content Management System, sistema de gestión de contenidos).

- Tanto Wordpress como otros CMS funcionan con la arquitectura WAMP o LAMP, principalmente está última.

- *LAMP* responde a Linux, Apache, MySQL y PHP y es la combinación de las cuatro tecnologías implicadas en Wordpress.

- Cuando vemos una página en Wordpress estamos viendo el ordenador de alguien, sea una persona o una empresa. Ese ordenador va normalmente con el sistema operativo GNU/Linux.

- Para que ese ordenador sirva páginas web requiere un servidor web. Apache es el más utilizado en el mundo, es software libre también, como GNU/Linux, aunque con otra licencia. También se usa mucho NGINX.

- En el HTML de Wordpress se incluye el lenguaje de programación PHP (PHP Hypertext Preprocessor, un acrónimo recursivo, figura muy utilizada en el mundo del software) que es el que construye las páginas HTML haciendo las llamadas a la base de datos donde se almacenan los contenidos de Wordpress.

- MySQL (donde SQL significa *Structured Query Language* o lenguaje de consulta estructurada) es el servidor de la base de datos.

3.2 SEO

- *SEO* responde a *Search Engine Optimization* u optimización del motor de búsqueda.

- No tiene más relación con Wordpress que el hecho de que Wordpress se utiliza para hacer web y cualquier web suele querer ser encontrada por un motor de búsqueda. Lo veremos en otro momento.

3.3 Excel

- No vamos a usar Excel pero conviene que sepamos qué es y por qué.

- Hay quien confunde Excel con hojas de cálculo o tablas de datos pero Excel es, por un lado, un programa para visualizar datos tabulados; por otro, un programa donde se pueden utilizar funciones para trabajar con los datos tabulados; y finalmente, un formato de datos XLSX.

- Para trabajar con datos tabulados veremos más adelante distintas formas, preferiblemente libres y/o abiertas.

- La última X proviene de XML cuando en 2008 Microsoft Office convirtió sus formatos de archivo en formatos compatibles con XML.

- *XML* significa *eXtensible Markup Language*. Se utiliza mucho en la industria/administraciones públicas pero no vamos a utilizarlo, al menos de momento, a no ser que lleguemos a otra cosa que es el *XPath* para el *scrapping*.

- Básicamente, XML es como que cualquiera puede crear sus propios elementos HTML. Por eso se puede utilizar para organizar la información, los datos o los procesos y compartirlo, si se quisiera.

- Microsoft Office es software propietario y privativo, no permite el libre uso. Más información en la siguiente sesión.

4 HTML

Veamos un poco también de *HTML* ya que, aunque no vamos a trabajarlo específicamente, sí que se usará en determinados momentos.

- *HTML* es el acrónimo de *HyperText Markup Language* o lenguaje de marcado de hipertexto.

- Es uno de los lenguajes, el básico y principal, que se utiliza en la Web.

- Es un lenguaje estructurado de marcas expresado en los signos <>.

- Dentro de las marcas se sitúan los **elementos** HTML como por ejemplo `p` de párrafo o `h1` de *header 1*, algo así como el título o encabezamiento principal.

- Conviene no confundir marcas con elementos.

- HTML es un lenguaje informático pero no es un lenguaje de programación.

- En un navegador (me refiero a Firefox, Chrome o derivados) si pulsamos el atajo de teclado `C-u` (Control + u) accedemos al código fuente de la página.

- Probad con una página de un medio internacional como The Guardian que además es referente como pionero del periodismo de datos moderno con el equipo del Datablog en 2008 compuesto por Simon Rogers y Paul Bradshaw (ya hablaremos de ellos en próximas clases).

- Lo que sale entre las marcas `<!--` y `-->` son comentarios de HTML, es decir, contenido que aunque está en el codigo fuente, en el HTML, no se visualiza. Esto es habitual en todos los lenguajes informáticos, unos caracteres reservados para que el navegador/visualizador del HTML, no interprete el contenido que está a continuación.

- Si no llegas aquí no puedes saber que quizás seas la próxima persona a la que contraten.

Figure 1: We are Hiring!

5 Lenguajes informáticos

Habéis comentado que algunas cosas os suenan a *lenguajes informáticos* o *lenguajes de programación* pero, ¿son lo mismo?

- Los lenguajes informáticos son todos los que entiende o puede entender el ordenador (a través de software, claro).

- Dentro de los lenguajes informáticos están los lenguajes estructurados, como puede ser HTML, que sirve para estructurar documentos.

- Y también están los lenguajes de programación que sirven para programar acciones que haga el ordenador. Entre estos se encuentran Java, C, C#, Python o R.

- En la web se utiliza mucho JavaScript, es el que aporta la interactividad. Tampoco lo vamos a abordar pero algo veremos... ¡al menos saber que existe y qué es lo que hace!

6 ¿Qué es la Web?

Ante esta preguntas algunas habéis respondido:

1. Un sistema que comunica información que se comparte globalmente.

2. Un espacio virtual de compartimentos que se relacionan entre sí.

3. Una base de servidores

4. ¿Qué pintan los dominios?

Son buenas respuestas para debatir. Un aspecto importante de la Web es que, a nivel técnico, es otro servicio de los que corren/se prestan en la red de redes que es Internet.

7 Qué es Internet

La Internet... :keycap_asterisk:

- Además de la mencionada definición concisa y precisa de "una red de redes", Internet funciona gracias a los protocolos TCP/IP.

- *TCP* responde a *Transmission Control Protocol* o protocolo de control de la transmisión.

- *IP* responde a *Internet Protocol*, os resultará más familiar porque al estar conectado a una red como es una Intranet el router nos tiene que dar una dirección de la red local, una IP.

- En una red TCP/IP hay unos 65500 puertos de escucha o de comunicación posible. La Web utiliza uno de ellos al menos, el 80.

- Hay otros servicios como el correo electrónico o la mensajería instantánea que utilizan otros puertos.

- Es cierto que se puede acceder a servicios de correo electrónico y mensajería instantánea por la Web, eso es porque las webs también pueden ser *webapps* o aplicaciones web. Esas aplicaciones conectan con los servicios de correo electrónico o mensajería instantánea que operan en los otros puertos y nos los muestran en un entorno web.

- Actualmente funcionamos con la versión original del protocolo, la denominada **IPv4**, que permitía hasta 4.300 millones de direcciones. Como se vio que se iba a quedar corto se empezó a trabajar en la versión **IPv6**, la cual actualmente funciona también pero conviven ambas. Este artículo de NordVPN lo explica muy bien.

□ (Queda pendiente explicar qué es una **VPN**.

8 HTTP

También os suena y sabéis la diferencia entre HTTP y HTTPS, la S es de "segura" y ahora es el estándar por defecto. No os fiéis de una página que no lo tenga.

- *HTTP* responde a *HyperText Transmission Protocol* o protocolo de control de la transmisión.

- Es como funciona la web, un protocolo muy simple pero no por ello limitado, al contrario.

- Cuenta con 4 acciones posibles:

    1. POST, publicar o crear. Es cuando se crea un documento nuevo.

    2. GET, obtener o bajarse. Es lo que hacemos cuando vemos una página web, solicitamos una copia de la web al servidor.

    3. DELETE, borrar el documento.

    4. PUT, actualiza un documento ya existente.

- Esta es una de las *APIs* más sencillas y conocidas. *API* significa *Access Programming Interface* o interfaz de programación de acceso, es algo así como los códigos para comunicarse con una web.

- HTTP es una API universal pero luego cada recurso puede tener la suya propia. Por ejemplo, Twitter tiene su propia API y así ocurre con muchos recursos que tienen muchos contenidos que ofrecen de maneras diversas.

- Si recordáis estas cuatro acciones que permite HTTP os aseguro que tenéis mucho ganado en relación con el uso que vais a hacer de la Web.

9 Dominios

Y también habéis comentado algo de los dominios.

- Los dominios están ahí para evitar tener que sabernos la dirección IP del servidor donde está alojada la página web que queremos visitar.

- Cuando escribimos el dominio en el navegador el ordenador consulta con el servidor de nombres de dominio *DNS* (/Domain Name System

) e indica en qué dirección IP se encuentra alojada la web. Esto es una petición GET de HTTP.

10 Github

Os presento a Github. Aunque lo ha comprado Micro$oft y ha perdido mucho de su glamour es una buena forma de habituarse a trabajar con un software de control de versiones que permite la colaboración y además, Github ofrece algunas posibilidades que hacen que lo vayamos a utilizar bastante.

- Se trata de una de las herramientas más usadas en periodismo de datos.

- Github es la suma de git, el software, y hub, el espacio montado por GitHub..

- Github es un espacio donde podemos alojar los repositorios o proyectos git.

- Empezamos dando por válido una analogía: es como un Wordpress donde vamos a poner nuestros contenidos web.

- Se pueden crear repositorios, algo así como una carpeta de nuestro sistema de ficheros del ordenador.

- Ahora también se pueden crear proyectos pero, de momento, creamos un repositorio.

- Hay que crear una cuenta :abc:

- En GitHub y en lo que escribamos no utilizaremos M$Word sino otros programas libres y/o abiertos y la sintaxis simple Markdown.

11 Las nubes

¡Cuidado con las nubes! :cloud-lightning:

- Se habla de la nube, cloud storage, *cloud computing* pero no hay nubes sino ordenadores de otras personas.

- En Periodismo de datos, dado que es periodismo de investigación, conviene no utilizar software del que no nos fiemos ni aplicaciones de terceros ni servicios en la nube. No solo nuestros datos o nuestra investigación pueden estar en peligro, también nosotrxs mismxs.

- Preferiblemente usaremos aplicaciones libres y/o abiertas y aplicaciones locales, aunque hay algunas excepciones que debieran circunscribirse al inicio y ser temporales.

12 Herramientas de visualización

- Depende del ritmo, veremos unas u otras. Hay muuuuchas.

Figure 2: There is no cloud, just other people's computers

- Seguro que veremos Datawrapper que aunque es un servicio de terceros, es gratuita y está basada en D3js que es libre.

- Hay otras similares como Infogram o Flourish.

- Hay librerías de visualización de datos de los lenguajes de programación Bash, Python y R que veremos si exploramos o no.

- Atlas o taxonomías de visulización de datos.

- Ejemplos, proyectos, compendios, newsletters...

13 Pruebas

- ¿Qué es el periodismo de datos? Aporta tus impresiones sobre el debate.

- Qué lenguajes informáticos conoces. Razona la respuesta.

- Cuál es la diferencia entre Internet y la Web. Razona la respuesta.

- ¿Qué fue determinante para el nacimiento del periodismo de datos moderno?