Manual de Periodismo de Datos 1.0 en Español

How journalists can use data to improve the news

The Data Journalism Handbook

Edited by Jonathan Gray, Liliana Bounegru and Lucy Chambers

Manual de Periodismo de Datos 1.0

El Manual de Periodismo de Datos puede ser copiado libremente, redistribuido y reusado bajo los términos de la licencia Creative Commons Atribución-CompartirIgual. Los contribuyentes al Manual del Periodimo de Datos retienen el copyright sobre sus contribuciones respectivas y están de acuerdo en publicarlas bajo los términos de esta licencia.

Un trabajo en equipo

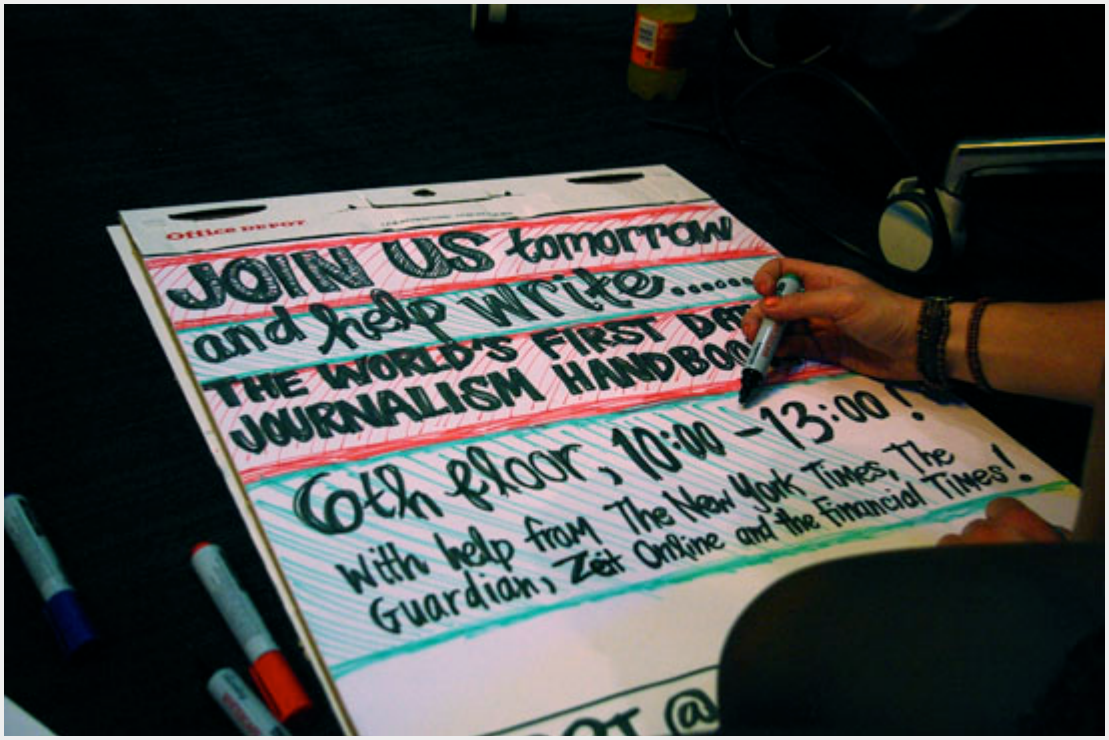


Figure 1. Cómo empezó todo

El Manual de Periodismo de Datos nació en un taller de 48 horas encabezado por European Journalism Centre y la Open Knowledge Foundation en la MozFest 2011 en Londres. Luego se amplió, convirtiéndose en un esfuerzo internacional en colaboración, que contó con la participación de docenas de los principales representantes del periodismo de datos y sus mejores exponentes.

En los 6 meses siguientes que pasaron entre el comienzo del libro y su primera presentación, cientos de personas contribuyeron de diversas maneras. Si bien hicimos nuestro mejor esfuerzo para reflejar a todos, hemos tenido una cantidad de anónimo, pseudónimos y editores imposibles de rastrear.

A todos los que aportaron y no aparecen en la lista, les decimos dos cosas. Primero, gracias. Segundo. Pueden por favor decirnos quiénes son de modo de poder darles el crédito que se merecen.

Contribuyentes

Las siguientes personas redactaron o contribuyeron directamente en la redacción de los textos en la actual versión del libro (y las ilustraciones son de la diseñadora gráfica Kate Hudson):

- Gregor Aisch, Open Knowledge Foundation
- Brigitte Alfter, Journalismfund.eu
- David Anderton, Periodista freelance
- James Ball, The Guardian
- Caelainn Barr, Citywire
- Mariana Berruezo, Hacks/Hackers Buenos Aires
- Michael Blastland, Periodista freelance
- Mariano Blejman, Hacks/Hackers Buenos Aires
- John Bones, Verdens Gang
- Marianne Bouchart, Bloomberg News
- Liliana Bounegru, European Journalism Centre
- Brian Boyer, Chicago Tribune
- Paul Bradshaw, Birmingham City University
- Wendy Carlisle, Australian Broadcasting Corporation
- Lucy Chambers, Open Knowledge Foundation
- Sarah Cohen, Duke University
- Alastair Dant, the Guardian
- Helen Darbishire, Access Info Europe
- Chase Davis, Center for Investigative Reporting
- Steve Doig, Walter Cronkite School of Journalism, Arizona State University
- Lisa Evans, The Guardian
- Tom Fries, Bertelsmann Stiftung
- Duncan Geere, Wired UK
- Jack Gillum, Associated Press

- Jonathan Gray, Open Knowledge Foundation
- Alex Howard, O'Reilly Media
- Bella Hurrell, BBC
- Nicolas Kayser-Bril, Journalism++
- John Keefe, WNYC
- Scott Klein, ProPublica
- Alexandre Léchenet, Le Monde
- Mark Lee Hunter, INSEAD
- Andrew Leimdorfer, BBC
- Friedrich Lindenberg, Open Knowledge Foundation
- Mike Linksvayer, Creative Commons
- Mirko Lorenz, Deutsche Welle
- Esa Mäkinen, Helsingin Sanomat
- Pedro Markun, Transparência Hacker
- Isao Matsunami, Tokyo Shimbun
- Lorenz Matzat, OpenDataCity
- Geoff McGhee, Stanford University
- Philip Meyer, Professor Emeritus, University of North Carolina at Chapel Hill
- Claire Miller, WalesOnline
- Cynthia O'Murchu, Financial Times
- Oluseun Onigbinde, BudgIT
- Djordje Padejski, Knight Journalism Fellow, Stanford University
- Jane Park, Creative Commons
- Angélica Peralta Ramos, La Nacion (Argentina)
- Cheryl Phillips, The Seattle Times
- Aron Pilhofer, New York Times
- Lulu Pinney, Diseñador infógrafo freelance
- Paul Radu, Organised Crime and Corruption Reporting Project
- Simon Rogers, The Guardian

- Martin Rosenbaum, BBC
- Amanda Rossi, Amigos de Januária
- Martin Sarsale, Hacks/Hackers Buenos Aires
- Fabrizio Scrollini, London School of Economics and Political Science
- Sarah Slobin, Wall Street Journal
- Sergio Sorin, Hacks/Hackers Buenos Aires
- Jonathan Stray, The Overview Project
- Brian Suda, (optional.is)
- Chris Taggart, OpenCorporates
- Jer Thorp, The New York Times R&D Group
- Andy Tow, Hacks/Hackers Buenos Aires
- Luk N. Van Wassenhove, INSEAD
- Sascha Venohr, Zeit Online
- Jerry Vermanen, NU.nl
- César Viana, University of Goiás
- Farida Vis, University of Leicester
- Pete Warden, Independent Data Analyst and Developer
- Chrys Wu, Hacks/Hackers

Lo que este libro es (y lo que no es)

Este libro busca ser un recurso útil para aquellos interesados en convertirse en periodistas de datos o que simplemente quieran tomarlo como un pasatiempo.

Muchas personas contribuyeron a su escritura, y a través de nuestra edición hemos tratado de hacer que se reflejen sus distintas voces y visiones. Esperamos que su lectura resulte una conversación rica e informativa respecto de lo que es el Periodismo de Datos, por qué es importante, y cómo hacerlo.

Lamentablemente, leer este libro no le proveerá un repertorio general de conocimientos y capacidades que necesitará para convertirse en periodista de datos. Esto requeriría una vasta biblioteca manejada por cientos de expertos capaces de responder preguntas sobre cientos de temas. Por suerte, tal biblioteca existe; se llama Internet. En cambio, esperamos que este libro lo oriente sobre cómo iniciarse y dónde mirar si quiere avanzar. Los ejemplos y tutoriales son ilustrativos más que exhaustivos.

Consideramos muy afortunado haber contado con tanto tiempo, energía y paciencia de todos nuestros contribuyentes y nos hemos esforzado por aprovecharlo de la mejor manera. Esperamos que –además de ser una fuente de referencia útil- el libro ayude a documentar la pasión y el entusiasmo, la visión y la energía de un movimiento en crecimiento. El libro intenta mostrar lo que sucede tras bambalinas, las historias detrás de los artículos.

El Manual de Periodismo de Datos es una obra en progreso. Si cree que algo necesita ser corregido o está notoriamente ausente, por favor indíquelo para su inclusión en la siguiente versión. También está disponible gratuitamente bajo una licencia Creative Commons de Atribución Compartir bajo la misma Licencia y lo alentamos fuertemente a que lo comparta con quien crea que puede interesarse en su lectura.

Liliana Bounegru (@bb_liliana)
Lucy Chambers (@lucyfedia)
Jonathan Gray (@jwyg)
March 2012

El Manual de un vistazo

Este manual a un vistazo: la infografista Lulu PInney creó este magnífico afiche, que da una visión general del contenido del Manual de periodismo de datos

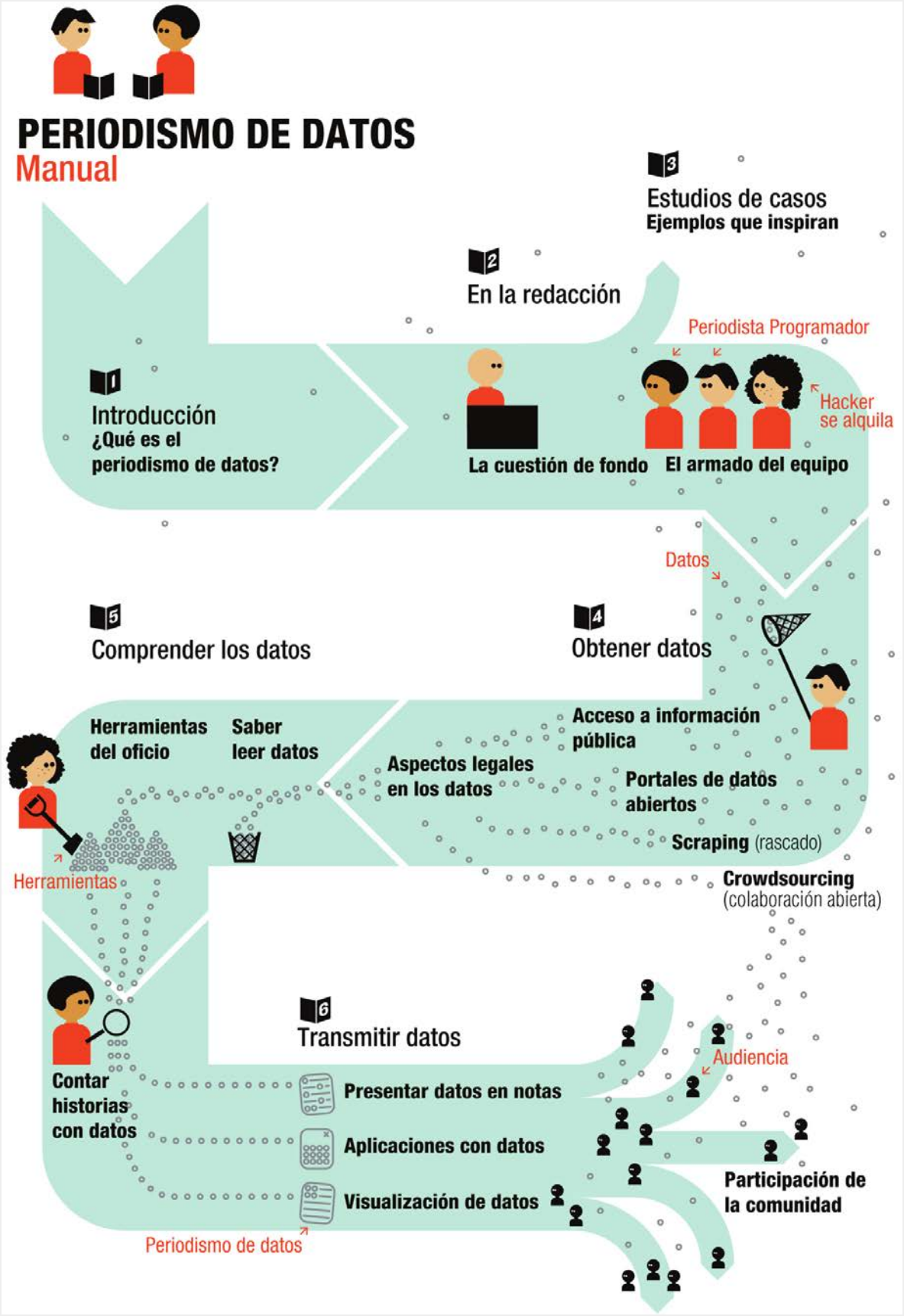


Figure 2. El Manual de un vistazo

Introducción

¿Qué es el periodismo de datos? ¿Qué potencial tiene? ¿Cuáles son sus límites? ¿De dónde viene? En esta sección analizamos qué es el periodismo de datos y lo que puede significar para las organizaciones de noticias. Paul Bradshaw (Birmingham City University) y Mirko Lorenz (Deutsche Welle) se refieren a la particular importancia de esta nueva disciplina. Destacados periodistas de datos comentan las claves a tener en cuenta y sus ejemplos favoritos. Finalmente Liliana Bounegru (European Journalism Centre) ubica al Periodismo de Datos en un contexto histórico más amplio.

Qué contiene este capítulo?

¿Qué es el periodismo de datos?

¿Qué es el periodismo de datos? Podría contestar, simplemente, que es periodismo que se hace con datos. Pero eso no es de gran ayuda.

Tanto "datos" como "periodismo" son términos problemáticos. Algunos creen que "datos" es cualquier colección de cifras, por lo general reunidas en una hoja de cálculo. Hace 20 años, esos eran prácticamente los únicos datos que manejaban los periodistas. Pero ahora vivimos en un mundo digital, un mundo en el que casi cualquier hecho puede ser (y casi todo es) descripto con números.

Su carrera profesional, 300.000 documentos confidenciales, las personas que componen su círculo de amigos; todo esto puede ser (y es) descripto con solo dos números: ceros y unos. Fotos, video, y audio; asesinatos, enfermedades, votos políticos, corrupción y mentiras, también descriptos con ceros y unos.

¿Qué es lo que hace que el periodismo de datos sea diferente del resto del periodismo? Quizás sean las nuevas posibilidades que aparecen, cuando se combina el tradicional "olfato para las noticias" y la capacidad de narrar una historia convincente, con la escala y alcance de la información digital disponible en la actualidad.

Y esas posibilidades pueden aparecer en cualquier momento del proceso periodístico: cuando contamos con la programación necesaria para automatizar el proceso de recoger y combinar información proveniente del gobierno municipal, la policía y otras fuentes civiles, como hizo Adrian Holovaty con ChicagoCrime y luego EveryBlock.

O usar software para encontrar relaciones entre cientos y miles de documentos, tal como hizo The Telegraph con los gastos de los parlamentarios.

Figure 3. Investigue los gastos de su representante parlamentario (The Guardian)

El periodismo de datos puede ayudar a un periodista a contar una historia convincente por medio de infografías atractivas. Por ejemplo, las conversaciones espectaculares de Hans Roslign sobre la visualización de la pobreza mundial con Gapminder (que se puede traducir como Recuerdabrecha, n. del t.) han atraído millones de visitas en todo el mundo. Y la obra popular de David McCandless al destilar grandes cifras –tales como poner en contexto el gasto público, o la polución generada por el volcán islandés- muestra la importancia de un diseño claro en Information is Beautiful.

O puede ayudar a explicar cómo se relaciona una historia con un individuo, como hacen ahora la BBC y el Financial Times habitualmente con sus interactivos sobre el presupuesto (donde usted puede averiguar cómo el presupuesto lo afecta en particular a usted en vez de a un genérico "Juan Pueblo"). Y puede abrir el proceso mismo de búsqueda de información, como hace The Guardian de modo tan exitoso al compartir datos, contexto y preguntas en su Datablog.

Los datos pueden ser la fuente del periodismo de datos, o pueden ser la herramienta con la que se narra la historia o ambas cosas. Como cualquier fuente, debe tratarse con escepticismo; y como cualquier herramienta, debemos ser conscientes de cómo puede modelar y limitar las historias que se crean con la misma.

— Paul Bradshaw, Birmingham City University

Por qué debieran usar datos los periodistas

El periodismo está sitiado. En el pasado, como sector, nos basábamos en ser los únicos que operábamos una tecnología para multiplicar y distribuir lo que había pasado de un día al otro. La imprenta servía como puerta de entrada. Cualquiera que quisiera llegar a la gente de una ciudad o una región a la mañana siguiente, recurría a los diarios. Esa era se acabó.

Hoy las noticias fluyen al mismo tiempo que suceden, a través de múltiples fuentes, testigos presenciales y blogs, y lo que ha sucedido es filtrado a través de una vasta red de conexiones sociales, se jerarquiza, se comenta y muy a menudo se ignora.

Por eso el periodismo de datos es tan importante. Reunir, filtrar y visualizar lo que sucede más allá de lo que nos muestran nuestros ojos tiene creciente valor. En la economía global de hoy el jugo de naranja que toma por la mañana, el café que prepara... hay relaciones invisibles entre estos productos, otra gente y usted. El lenguaje de esta red es el de los datos: pequeños puntos de información que a menudo son irrelevantes como instancia individual, pero enormemente importantes cuando se los ve desde el ángulo correcto.

En este momento, unos cuantos periodistas pioneros ya están demostrando cómo se puede usar datos para crear una visión más profunda de lo que sucede a nuestro alrededor y cómo puede afectarnos.

El análisis de datos puede revelar “la forma de una historia” (Sarah Cohen) o proveernos una “nueva cámara” (David McCandless). Usando datos, la tarea de los periodistas pasa de centrarse en ser los primeros en informar, a ser los que nos dicen lo que un proceso podría significar realmente. La gama de temas puede ser amplia. La próxima crisis financiera en ciernes. Los datos económicos detrás de los productos que usamos. El mal uso de fondos o errores políticos, presentados con una visualización convincente que deje poco margen para rebatirla.

Es por esto que los periodistas debieran ver los datos como una oportunidad. Es posible, por ejemplo, revelar cómo una amenaza abstracta (como el desempleo) afecta a la gente de acuerdo a su edad, su género o su nivel de educación. Usar datos transforma algo abstracto en algo que todos pueden entender y con lo que pueden relacionarse.

Pueden crear herramientas de cálculo personalizadas para ayudar a la gente a tomar decisiones, se trate de comprar un auto o una casa, decidir un rumbo educativo o profesional en su vida, o hacer un control de costos para no meterse en deudas.

Pueden analizar la dinámica de una situación compleja como disturbios o un debate político, mostrar falacias y ayudar a todos a encontrar posibles soluciones para problemas complejos.

Formarse en la búsqueda, depuración y visualización de datos es transformador para la profesión de reunir información también. Los periodistas que dominen esto descubrirán que apoyar sus artículos en datos y la visión que aportan es un alivio. Menos adivinar, menos buscar citas; en vez de ello, un periodista puede crear una posición fuerte apoyada en datos y esto puede afectar mucho el rol del periodismo.

Además, introducirse en el periodismo de datos ofrece una perspectiva para el futuro. Hoy, cuando las redacciones se reducen, la mayoría de los periodistas esperan cambiar el área de las relaciones públicas. Pero los periodistas de datos o los científicos de datos ya son un grupo de profesionales muy solicitados, no solo por los medios. Las empresas e instituciones de todo el mundo buscan “gente que encuentre sentido a las cosas”, y profesionales que sepan cómo revisar datos y convertirlos en algo tangible.

Los datos representan una promesa, y esto es lo que entusiasma a las redacciones, haciéndolas buscar un nuevo tipo de periodista. Para la gente que trabaja por su cuenta, manejar datos ofrece un camino para obtener nuevas oportunidades y un salario estable también. Véalo de este modo: en vez de contratar periodistas que llenen rápidamente páginas y sitios en la red con contenido de bajo valor, el uso de datos podría crear demanda para paquetes interactivos, que solo pueden crearse invirtiendo una semana entera en resolver una cuestión. Esto es un cambio positivo para muchos sectores de los medios.

Hay una barrera que impide a los periodistas usar este potencial: la necesidad de

capacitarse para trabajar con datos en todos los pasos, desde una primera pregunta hasta un gran impacto periodístico basado en datos.

Trabajar con datos es como introducirse en un territorio vasto y desconocido. A primera vista los datos crudos resultan inteligibles para los ojos y la mente. Tales datos son inmanejables. Es difícil ordenarlos correctamente para su visualización. Se necesita periodistas experimentados, que tengan la energía como para analizar datos crudos a menudo confusos o aburridos y "ver" las historias ocultas allí.

— Mirko Lorenz, Deutsche Welle

El estudio

El European Journalism Centre realizó una encuesta para saber más sobre las necesidades de capacitación de los periodistas. Descubrimos que hay una gran disposición de salir de la postura cómoda del periodismo tradicional, e invertir tiempo para dominar nuevas capacidades. Los resultados de la encuesta demuestran que los periodistas ven la oportunidad, pero necesitan un poco de apoyo para superar los problemas iniciales que les impiden trabajar con daos. Hay confianza de que si el periodismo de datos fuera adoptado de modo más universal, los flujos de trabajo, las herramientas y los resultados mejorarían rápidamente. Pioneros tales como The Guardian, The New York Times, The Texas Tribune, y Die Zeit siguen elevando el nivel con sus artículos basados en datos.

¿El periodismo de datos seguirá siendo el dominio de un pequeño puñado de pioneros o pronto toda organización de noticias tendrá su propio equipo de periodistas dedicados especialmente a los datos. Esperamos que este manual ayude a más periodistas y redacciones a aprovechar este campo emergente.

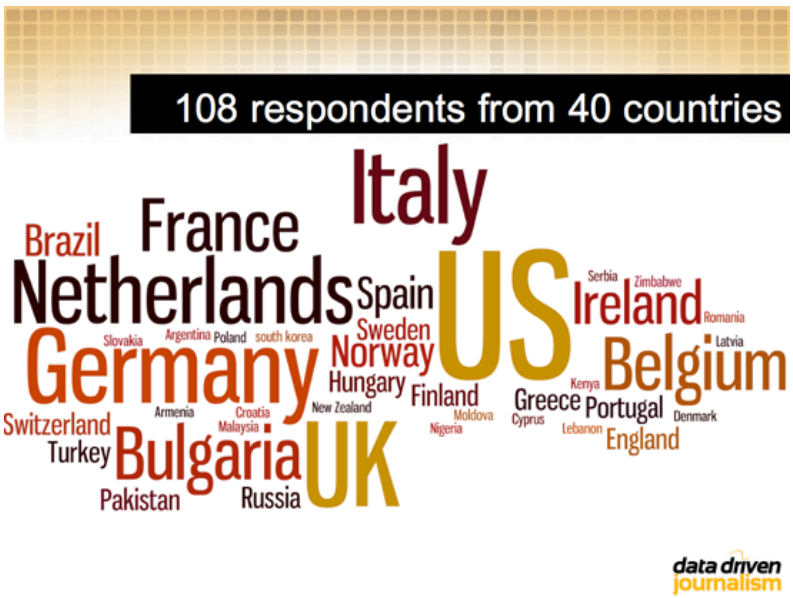


Figure 4. Encuesta del European Journalism Centre sobre necesidades de capacitación

¿Por qué es importante el periodismo de datos?

Preguntamos a algunos de los principales practicantes y partidarios del periodismo de datos por qué piensan que el periodismo de datos es un desarrollo importante. Los datos crudos resultan inteligibles para los ojos y la mente. Esto es lo que dijeron:

Filtrar el flujo de datos.

Cuando había escasez de información, la mayor parte de nuestros esfuerzos estaban dedicados a buscarla y reunirla. Ahora que la información es abundante, es más importante el procesamiento. El procesamiento tiene dos niveles: 1) análisis para encontrar sentido y estructura en el flujo sin fin de datos y 2) presentación de esa información para meter lo que es importante y relevante en la cabeza del consumidor. Al igual que la ciencia, el periodismo de datos da a conocer sus métodos y presenta sus descubrimientos de un modo que pueda ser verificado a través de su replicado.

— Philip Meyer, Professor Emeritus, University of North Carolina at Chapel Hill

Nuevos enfoques para narrar historias

El periodismo de datos es un término abarcativo que, para mí, incluye un conjunto de herramientas, técnicas y enfoques de la narrativa siempre crecientes. Puede incluir todo, desde el tradicional periodismo asistido por computadoras (usando datos como una "fuente") hasta la visualización más avanzada de datos y aplicaciones de noticias. El objetivo unificador es periodístico: proveer información y análisis para ayudar a informarnos todos sobre asuntos importantes de actualidad.

— Aron Pilhofer, New York Times

Como periodismo fotográfico con una laptop

"El periodismo de datos" difiere del "periodismo escrito" solo en que usamos un equipo diferente. Todos nos ganamos la vida olfateando, reportando y relacionando historias. Es como el "periodismo fotográfico"; solo hay que cambiar la cámara por una laptop.

— Brian Boyer, Chicago Tribune

El periodismo de datos es el futuro

El periodismo de datos es el futuro. Los periodistas tienen que saber manejar datos. Hace un tiempo uno descubría cosas hablando con gente en bares, y puede ser que esto siga

sucediendo a veces. Pero ahora también se trata de analizar datos, equiparse con herramientas, y analizarla y encontrar lo que es interesante. Tener todo en perspectiva, ayudando a la gente a ver cómo encajan las piezas (para no repetir todo), y qué pasa en el país.

— *Tim Berners-Lee, founder of the World Wide Web*

El procesamiento de cifras se une al pulido del lenguaje

El periodismo de datos es tender un puente para superar la brecha entre los técnicos estadísticos y los cinceladores de palabras. Ubicar cosas destacadas e identificar tendencias que no solo son significativas estadísticamente sino que también son relevantes para desentrañar el mundo de hoy, que es intrínsecamente complejo.

— *David Anderton, freelance journalist*

Actualizar sus capacidades

El periodismo de datos implica un nuevo conjunto de habilidades para buscar, comprender y visualizar fuentes digitales, en una época en que las capacidades básicas del periodismo tradicional ya no bastan. No lo reemplaza, le agrega cosas.

En un momento en que las fuentes se están volviendo digitales, los periodistas pueden y tienen que estar más en contacto con estas fuentes. Internet abrió posibilidades que van más allá de lo que podemos entender hoy. El periodismo de datos es solo el comienzo de la evolución de nuestras prácticas pasadas para adaptarse al online.

El periodismo de datos sirve a dos importantes propósitos para las organizaciones de noticiosas: encontrar historias únicas (no de los cables) y ejecutar la función de alerta. Especialmente en tiempos de crisis financieras, estos objetivos son importantes para los diarios.

Desde el punto de vista de un diario regional, el periodismo de datos es crucial. Existe el dicho: “una teja floja en su casa se considera más importante que disturbios en un país lejanos”. A uno lo golpea en la cara e impacta en su vida de modo más directo. Al mismo tiempo, la digitalización está en todas partes. Debido a que los diarios locales tienen este impacto directo en su vecindario y las fuentes se vuelven digitalizadas, un periodista debe saber cómo encontrar, analizar y visualizar una historia a partir de datos.

— *Jerry Vermanen, NU.nl*

Un remedio para la asimetría de la información

La asimetría de la información –no la falta de información sino la incapacidad de absorberla

y procesarla a la velocidad y con el volumen que nos llega- es uno de los problemas más significativos que enfrentan los ciudadanos al elegir cómo vivir sus vidas. La información tomada de medios impresos, visuales y radiales influye en las opciones y las acciones de los ciudadanos. El buen periodismo de datos ayuda a combatir la asimetría de la información.

— Tom Fries, Bertelsmann Foundation

Una respuesta a las relaciones públicas basadas en datos

La disponibilidad de herramientas de medición y sus precios decrecientes –en una combinación auto-sustentada que se concentra en el desempeño y la eficiencia en todos los aspectos de la sociedad- han llevado a quienes toman las decisiones a cuantificar los avances de sus políticas, monitorear tendencias e identificar oportunidades.

Las compañías continuamente encuentran nuevas mediciones que muestran su buen desempeño. A los políticos les encanta alardear de las cifras sobre reducción de desempleo y crecimiento del PBI. La falta de conocimientos por parte de los periodistas respecto de los escándalos de Enron, Worldcom, Madoff o Solyndra es prueba de la incapacidad de muchos profesionales de ver más allá de las cifras. Hay una tendencia a aceptar las cifras más que otros datos, ya que tienen un aura de seriedad, aunque sean completamente inventadas.

El saber manejar datos ayudará a los periodistas a aguzar su sentido crítico al enfrentar cifras, y ojalá que les sirva para avanzar un poco en su relación con los departamentos de RRPP.

— Nicolas Kayser-Bril, Journalism++

Proveer interpretaciones independientes de información oficial

Luego del terremoto devastador y el subsecuente desastre de la planta nuclear de Fukushima en 2011, la importancia del periodismo de datos se ha hecho claro para la gente de medios en Japón, país que en general va a la zaga en materia de periodismo digital.

Quedamos a la deriva cuando el gobierno y los expertos no tuvieron datos creíbles acerca de los daños. Cuando los funcionarios ocultaron al público los datos SPEEDI (predicción de difusión de materiales radioactivos), no estábamos en condiciones de decodificarlos aunque se hubiesen filtrado. Voluntarios comenzaron a reunir datos sobre radioactividad usando sus propios recursos, pero no estábamos armados con conocimientos estadísticos, de interpolación, de visualización y demás. Los periodistas tienen que tener acceso a los datos en crudo y aprender a no depender de las interpretaciones oficiales de los mismos.

— Isao Matsunami, Tokyo Shimbun

Manejar el diluvio de datos

Los desafíos y las oportunidades que presenta la revolución digital siguen complicando al periodismo. En una era de abundancia de información, los periodistas y los ciudadanos necesitan mejores herramientas, se trate de curar los samizdat del siglo XXI en Medio Oriente, procesar una avalancha de datos difundidos a medianoche, o encontrar la mejor manera de visualizar la calidad del agua en una nación. Al debatirnos con los desafíos del consumo que presenta este diluvio de datos, las nuevas plataformas de edición también están dando a todos el poder de reunir y compartir datos digitalmente, convirtiéndolos en información. Mientras los periodistas y editores han sido los vectores tradicionales de la colecta y diseminación de información, el ambiente horizontal de información ahora hace que las noticias se conozcan primero online y no en las redacciones.

En todo el planeta, de hecho, el vínculo entre los datos y el periodismo se está fortaleciendo. En una era de grandes cantidades de datos, la creciente importancia del periodismo de datos está en la capacidad de sus practicantes de dar contexto, claridad y –quizás lo más importante, encontrar la verdad en la cantidad en expansión de contenido digital en el mundo. Eso no significa que las organizaciones de medios integradas de hoy no tengan un rol crucial. Lejos de ello. En la era de la información, se necesita más que nunca a los periodistas para curar, verificar, analizar y sintetizar los datos. En ese contexto, el periodismo de datos tiene una profunda importancia para la sociedad.

Hoy, encontrarle sentido a los grandes volúmenes de datos, en particular los datos no estructurados, serán un objetivo central de los científicos de todo el mundo, trabajen en salas de redacción, Wall Street o Silicon Valley. Notoriamente esa meta se verá facilitada sustancialmente por un conjunto creciente de herramientas comunes, sean empleadas por tecnólogos del estado, tecnólogos de la salud o desarrolladores de las redacciones.

— Alex Howard, O'Reilly Media

Nuestras vidas son datos

El buen periodismo de datos es difícil, porque el buen periodismo es difícil. Significa cómo obtener los datos, cómo entenderlos, y cómo encontrar la historia. A veces hay callejones sin salida, y a veces no hay una gran historia. Al fin de cuentas, si solo fuera cuestión de apretar el botón indicado, no sería periodismo. Pero eso es lo que hace que valga la pena –en un mundo en el que nuestras vidas cada vez son más datos-, que sea esencial para una sociedad libre y justa.

— Chris Taggart, OpenCorporates

Una manera de ahorrar tiempo

Los periodistas no tienen tiempo para perder transcribiendo cosas a mano y complicarse tratando de obtener información de archivos PDF, por lo que aprender un poco de código (o saber dónde buscar gente que puede ayudar) es increíblemente valioso.

Un periodista de Folha do São Paulo estaba trabajando con el presupuesto local y me llamó para agradecernos por publicar online las cuentas de la municipalidad de São Paulo (2 días de trabajo para un solo hacker). Dijo que las había estado transcribiendo a mano los últimos 3 meses, tratando de encontrar una historia. También recuerdo haber resuelto un "problema de PDF" para *Contas Abertas*, una organización que monitorea noticias parlamentarias: 15 minutos y 15 líneas de código, en vez de un mes de trabajo.

— *Pedro Markun, Transparência Hacker*

Una parte esencial del herramental del periodista

Creo que es importante destacar el aspecto "periodístico" o de reportero del "periodismo de datos. El ejercicio no debiera ser analizar o visualizar datos por el gusto de hacerlo, sino utilizarlo como herramienta de modo de aproximarnos más a la verdad de lo que sucede en el mundo. Veo la capacidad de analizar e interpretar datos como parte esencial del set de herramientas actual de los periodistas, en vez de una disciplina por separado. Al fin de cuentas, todo tiene que ver con el buen periodismo y contar historias del modo más apropiado.

El periodismo de datos es otra manera de analizar el mundo y hacer que los poderes constituidos rindan cuentas. Con una creciente cantidad de datos disponible, ahora es más importante que nunca que los periodistas sean conscientes de las técnicas del periodismo de datos. Esta debe ser una herramienta que cualquier periodista debiera incorporar, se trate de aprender cómo trabajar directamente con datos, o a colaborar con alguien que lo pueda hacer.

Su verdadero potencial está en ayudarlo a obtener información que de otro modo sería muy difícil de encontrar o demostrar. Un buen ejemplo es la historia de Steve Doig que analizó patrones de daños del huracán Andrew. Unió dos conjuntos distintos de datos: uno que mapeaba el nivel de destrucción causado por el huracán, y otro que muestra las velocidades de los vientos. Esto le permitió señalar áreas en las cuales las malas prácticas en la construcción de edificios contribuyeron/intensificaron el impacto del desastre. Ganó por la historia un Pulitzer Prize en 1993 y sigue siendo un gran ejemplo de lo que es posible.

Idealmente se usan los datos para descubrir cosas destacadas, sorprendentes o áreas de interés. En este sentido, actúan como pistas. Si bien las cifras pueden ser interesantes, no

basta escribir solamente sobre datos. Hay que hacer el trabajo de periodista para explicar qué significan.

— Cynthia O'Murchu, Financial Times

Adaptarse a cambios en nuestro ambiente de información

Las nuevas tecnologías digitales generan nuevas maneras de producir y diseminar el conocimiento en la sociedad. El periodismo de datos puede entenderse como el intento de los medios de adaptarse y responder a los cambios en el ambiente de la información, incluyendo maneras de contar historias más interactivas y multidimensionales, que permite a los lectores explorar las fuentes que subyacen a las noticias, alentándolos a participar en el proceso de crear y evaluar historias.

— César Viana, University of Goiás

Una manera de ver cosas que de otro modo podría no ver

Algunas historias sólo pueden entenderse y explicarse analizando –y a veces visualizando- datos. Las relaciones entre personas o entes poderosos quedarían sin revelar, las muertes causadas por políticas farmacéuticas permanecerían ocultas, las políticas ambientales que dañan el medio continuarían sin límite. Pero cada una de estas situaciones han podido modificarse gracias a los datos obtenidos, analizados y aportados por los periodistas a los lectores. Los datos pueden ser simples como una planilla de cálculo, o un registro de llamadas telefónicas, o complejos como los resultados de pruebas escolares o datos de infecciones hospitalarias; como sea, allí hay historias que vale la pena contar.

— Cheryl Phillips, The Seattle Times

Una manera de enriquecer los artículos

Podemos pintar cuadros de nuestras vidas completas con nuestro rastro digital. Desde lo que consumimos y navegamos, hasta donde y cuando viajamos, nuestras preferencias musicales, nuestros primeros amores, los hitos de nuestros hijos, incluso nuestros últimos deseos, todo puede ser rastreado, digitalizado, almacenado en la nube y difundido.**Este universo de datos puede ser sacado a la superficie para narrar historias, responder preguntas e impartir una comprensión de la vida de maneras que actualmente superan incluso la más rigurosa y cuidadosa reconstrucción de anécdotas.

— Sarah Slobin, Wall Street Journal

No se necesitan nuevos datos para tener una primicia

A veces los datos ya son públicos y están disponibles, pero nadie los ha analizado atentamente. En el caso del informe de Associated Press sobre 4500 páginas de documentos desclasificados que describen las acciones de contratistas de seguridad privados durante la guerra de Irak, el material fue obtenido por un periodista independiente a lo largo de varios años, usando pedidos de Acceso a la Información

dirigidos al departamento de Estado de EE.UU. Escanearon los resultados impresos y los subieron a DocumentCloud, lo que nos permitió hacer nuestro análisis general.

— *Jonathan Stray, The Overview Project*

Algunos ejemplos favoritos

Le preguntamos a algunos de nuestros colaboradores acerca de sus ejemplos favoritos de periodismo de datos y qué les gusta de los mismos. Sus respuestas, a continuación:

No causar daño, en el Las Vegas Sun

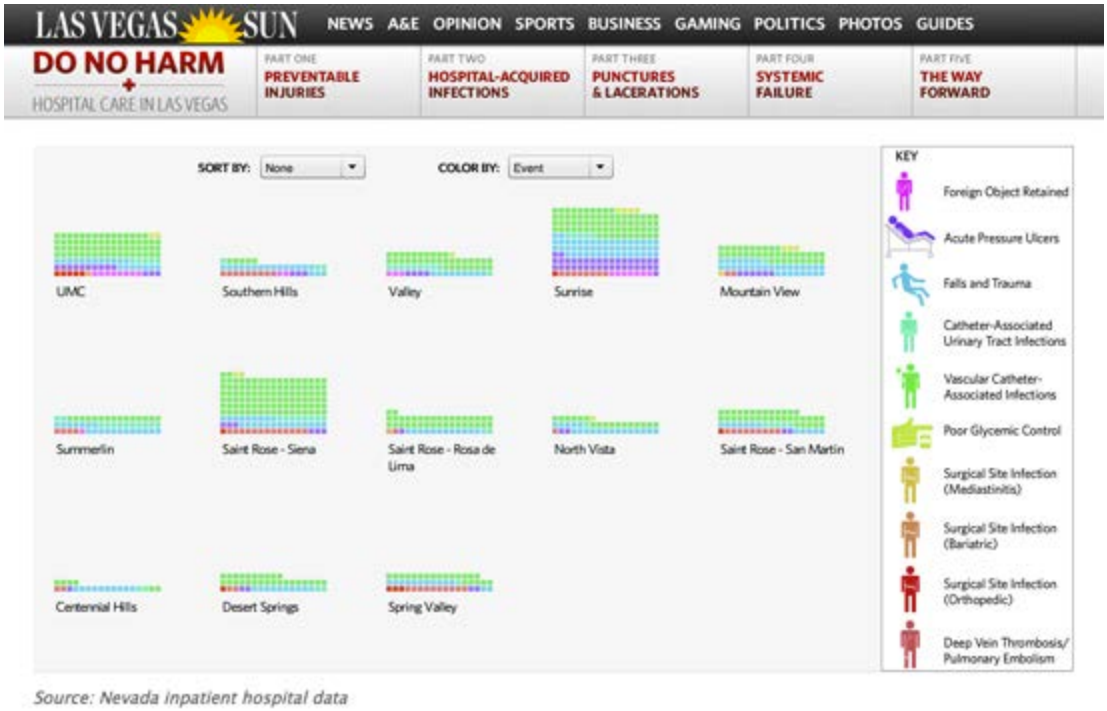


Figure 5. No causar daño (The Las Vegas Sun)

Mi ejemplo favorito es la serie No causar daño de 2010 en Las Vegas Sun, sobre la atención en los hospitales. El Sun analizó más de 2.900.000 de registros de aranceles hospitalarios, que revelaron más de 3600 lesiones, infecciones y errores quirúrgicos evitables. Obtuvieron

datos a través de un pedido de acceso a archivos públicos e identificaron más de 300 casos en que los pacientes murieron por errores que pudieron haberse prevenido. Contiene distintos elementos, incluyendo un gráfico interactivo que permite al lector ver (por hospital) donde se dieron lesiones quirúrgicas más a menudo de lo esperado; un mapa con un cronograma que muestra cómo se extienden las infecciones hospital por hospital; y un gráfico interactivo que permite a los usuarios ordenar los datos por lesiones evitables o por hospital, para ver dónde la gente se ve afectada. Me gusta porque es muy fácil de entender y navegar. Los usuarios pueden explorar los datos de manera muy intuitiva.

Además tuvo un impacto real: la legislatura de Nevada respondió con 6 legislaciones. Los periodistas involucrados trabajaron muy duro para obtener y desmenuzar los datos. Uno de los periodistas, Alex Richards, envió los datos a los hospitales y al Estado al menos una docena de veces para lograr que se corrigieran los errores.

— Angélica Peralta Ramos, La Nación (Argentina)

Base de datos de salarios de empleados del Estado

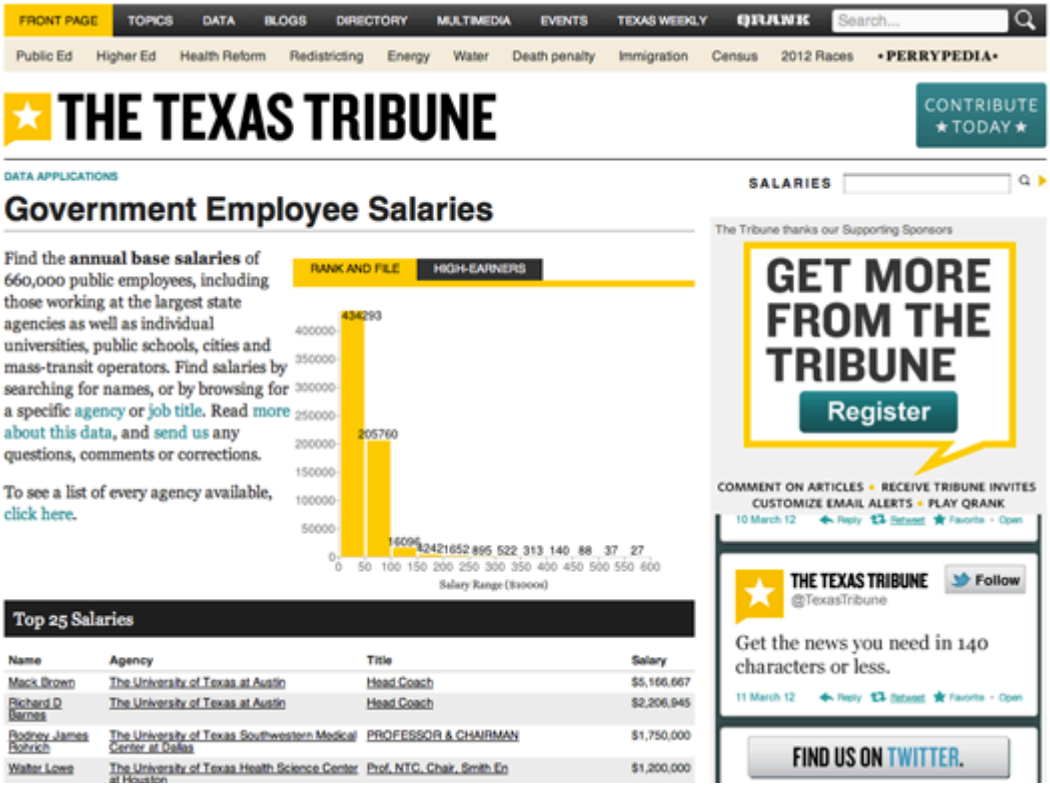


Figure 6. Salarios de Empleados del Estado (The Texas Tribune)

Me encanta el trabajo que pequeñas organizaciones independientes realizan todos los días tales como ProPublica o el Texas Tribune, que tiene a Ryan Murphy como gran periodista de datos. Si tuviera que elegir, optaría por el proyecto de base de datos de Salarios de

Empleados del Estado del Texas Tribune. Este proyecto reúne en una base de datos la información de los salarios de 660.000 empleados estatales para que los usuarios busquen y ayuden a generar historias. Se puede buscar por ente estatal, nombre o salario. Es simple, significativo y pone a disposición del público información hasta ahora inaccesible. Es fácil de usar y genera historias de manera automática. Es un gran ejemplo que muestra por qué el Texas Tribune concentra la mayor parte de su tráfico en sus páginas de datos.

— *Simon Rogers, The Guardian*

Visualización de texto completo de los registros de la guerra de Irak, Associated Press

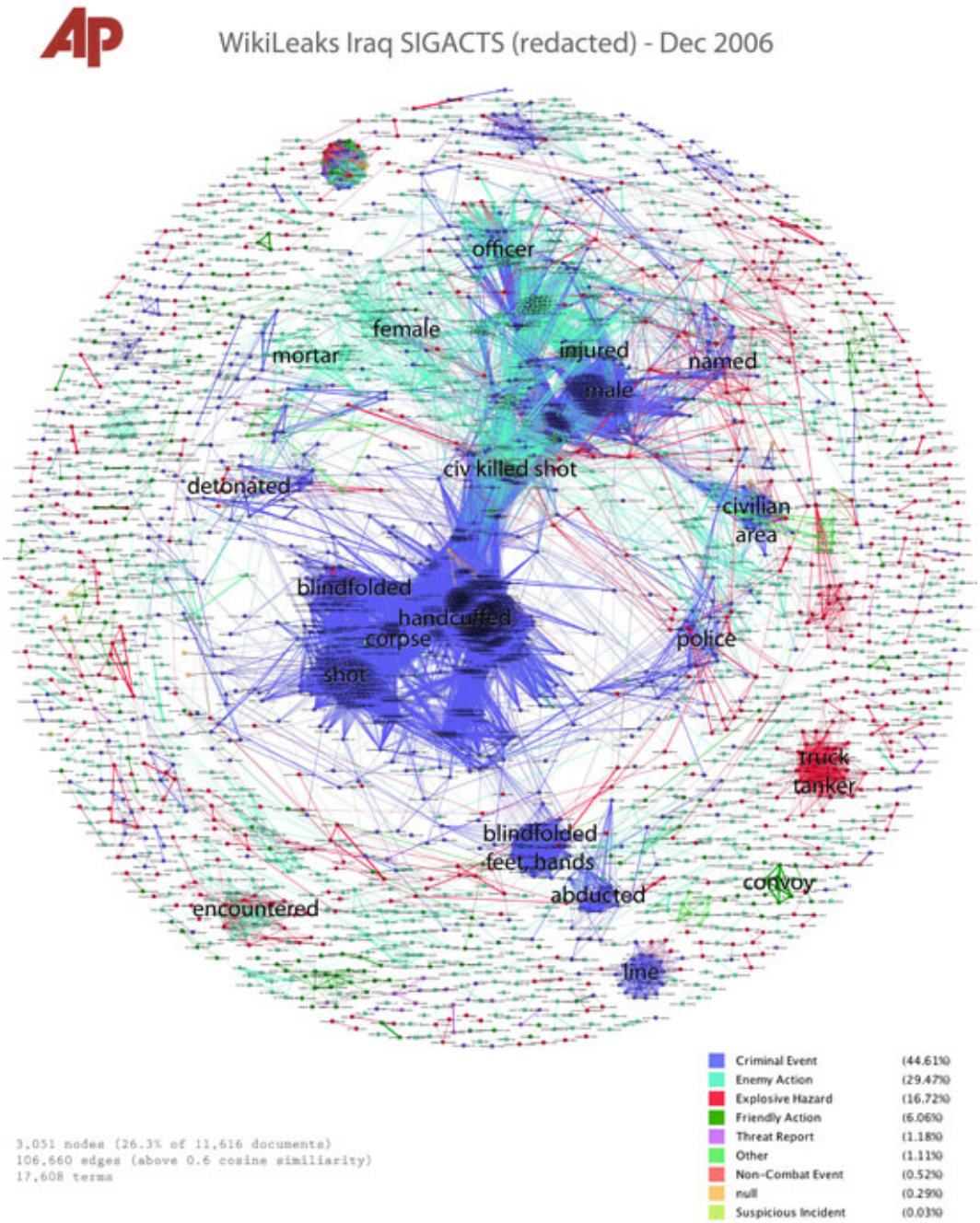


Figure 7. Análisis de los registros de guerra (Associated Press)

El trabajo de Jonathan Stray y Julian Burgess sobre los registros (logs) de la Guerra de Irak es una llamativa incursión en el análisis de texto y la visualización, utilizando técnicas experimentales para comprender temas que vale la pena explorar, dentro de un gran conjunto de datos en formato texto.

Por medio de técnicas y algoritmos de analítica de textos, Jonathan y Julian crearon un método que muestra concentraciones de palabras clave contenidas en miles de informes del gobierno de Estados Unidos sobre la guerra de Irak, difundido por WikiLeaks, en un formato visual.

Si bien este método tiene limitaciones y el trabajo es experimental, es un enfoque nuevo e innovador. En vez de tratar de leer todos los archivos o revisar los registros de guerra con una noción preconcebida de lo que puede encontrarse ingresando palabras claves y revisando el resultado, esta técnica calcula y visualiza temas/palabras clave de particular relevancia.

Con crecientes cantidades de datos en formato texto (emails, informes, etc.) y numérico llegando al dominio público, encontrar maneras de determinar áreas de interés clave se volverá cada vez más importante. Es un sub-campo interesante del periodismo de datos.

— Cynthia O'Murchu, Financial Times

Misterios de Asesinatos

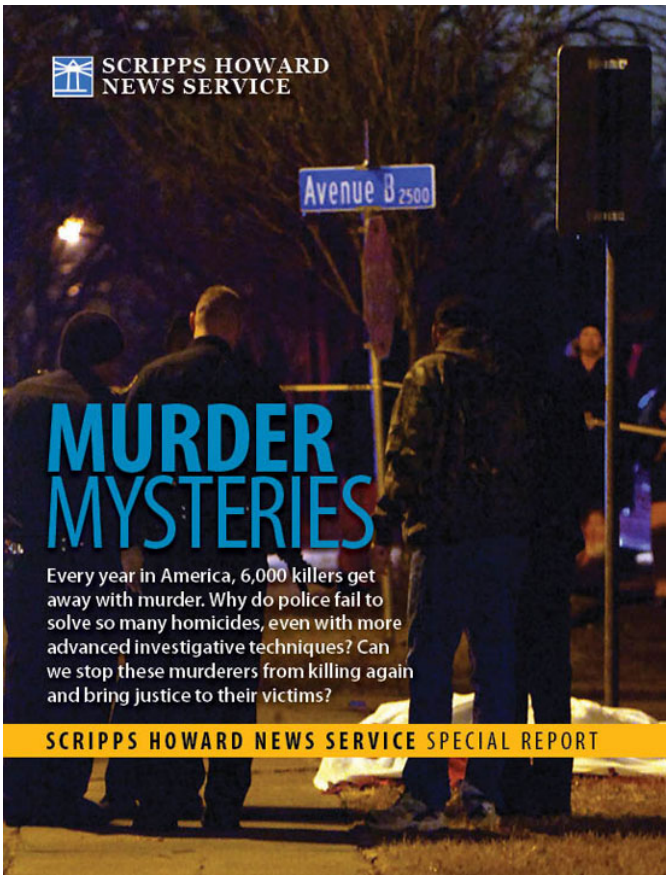


Figure 8. Misterios de asesinatos (Scripps Howard News Service)

Una de mis piezas favoritas de periodismo de datos es el proyecto de Misterios de Asesinatos, por Tom Hargrove del Scripss Howard News Service. A partir de datos oficiales y pedidos de acceso a registros públicos, creó una base de datos, que incluye el detalle demográfico de más de 185.000 asesinatos no resueltos, y luego diseñó un algoritmo para buscar patrones que sugieran la posible presencia de asesinos seriales.

Este proyecto tiene todo: un gran trabajo, una base de datos mejor que la del estado, análisis inteligente usando técnicas de ciencias sociales, y una presentación interactiva de datos online de modo que los lectores puedan explorar por su cuenta.

— Steve Doig, Walter Cronkite School of Journalism, Arizona State University

Máquina de Mensajes

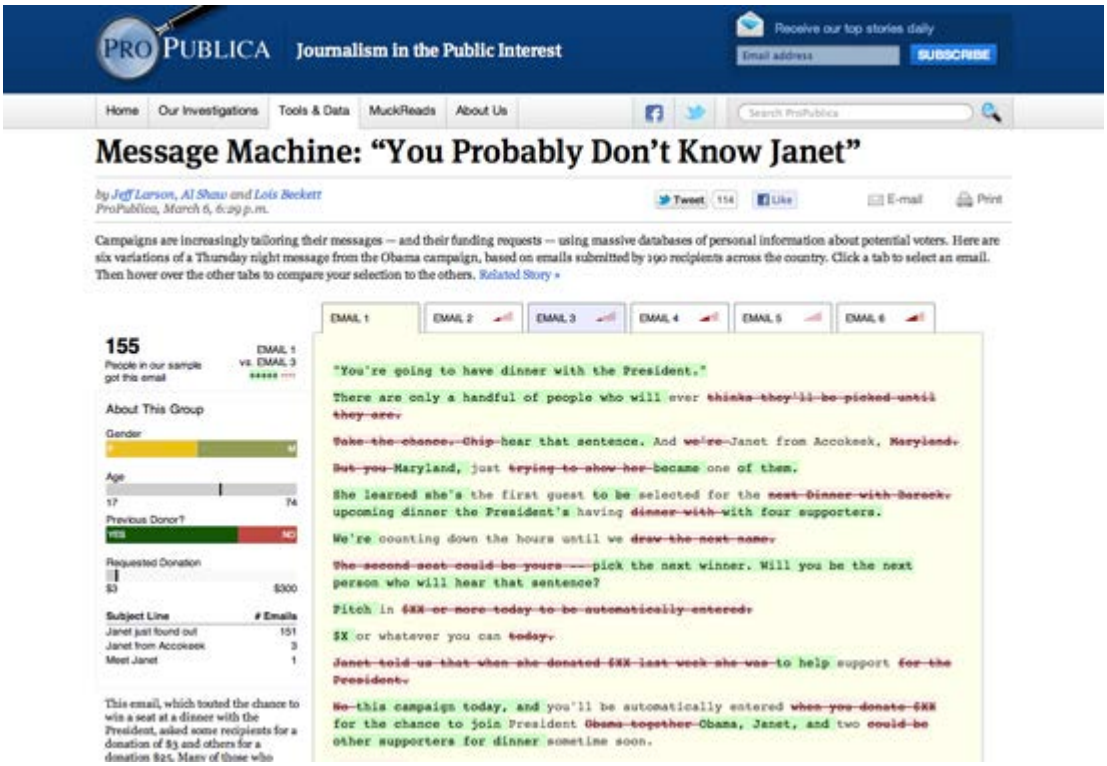


Figure 9. Máquina de Mensajes (ProPublica)

Me encanta la historia de Máquina de Mensajes de ProPublica y su blog nerd. Todo comenzó cuando un grupo de tuiteros expresó curiosidad por haber recibido correos electrónicos diferentes de la campaña de Barack Obama. La gente de ProPublica tomó nota y pidió a su público que reenviaran los correos que recibieran de la campaña. La presentación es elegante, un análisis diferencial visual de varios correos diferentes que fueron enviados esa noche. Es admirable porque recogieron sus propios datos (una pequeña muestra, pero lo suficiente como para contar la historia). Pero es aún más admirable porque cuenta la historia de un fenómeno en curso: gran cantidad de datos utilizados en campañas

políticas para dirigir mensajes a individuos específicos. Es sólo un anticipo de cosas por venir.

— *Brian Boyer, Chicago Tribune*

Chartball

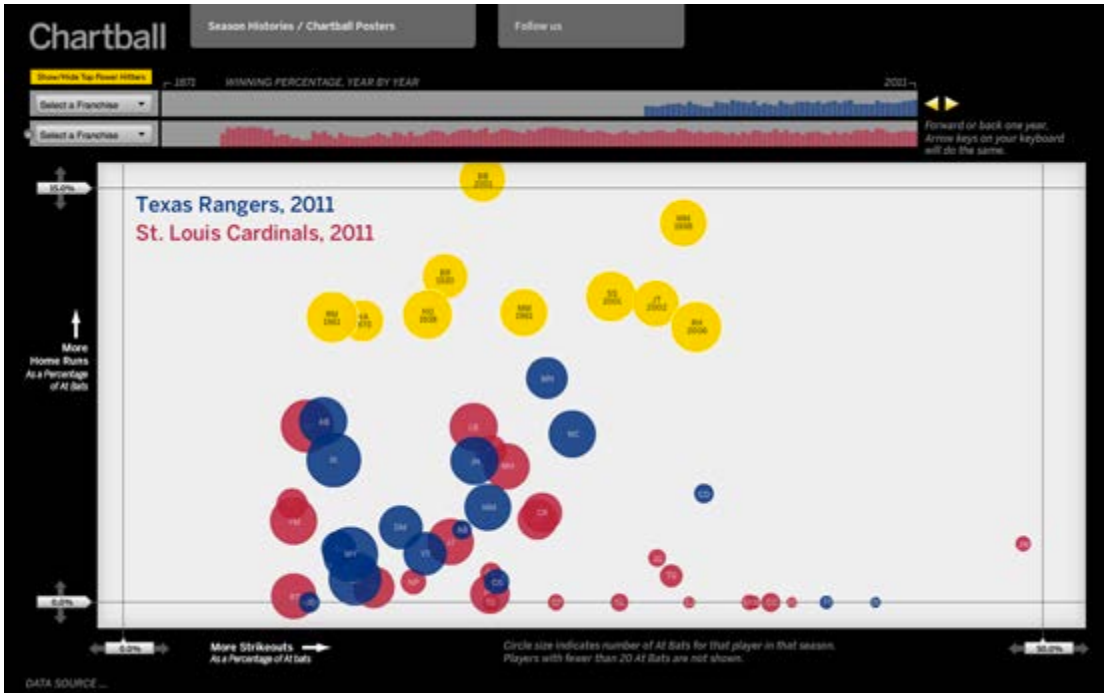


Figure 10. Gráfico de victorias y derrotas (Chartball)

Uno de mis proyectos de periodismo de datos favoritos es el trabajo de Andrew García Phillips sobre http://www.chartball.com/Chartball. Andrew es fanático de los deportes con un voraz apetito de datos, un ojo tremendo para el diseño y la capacidad de escribir código. En Chartball no solo visualiza el conjunto de la historia, sino que detalla los éxitos y fracasos de jugadores individuales y equipos. Ofrece contexto, un gráfico atractivo y su trabajo es profundo, divertido e interesante, y a mí ni siquiera me interesan demasiado los deportes.

— *Sarah Slobin, Wall Street Journal*

El periodismo de datos en perspectiva

En agosto de 2010 algunos colegas del European Journalism Centre y yo organizamos lo que creemos que fue una de las primeras conferencias internacionales de periodismo de datos, que se realizó en Ámsterdam, Holanda. En aquel momento no había mucha discusión respecto del tema, y solo había un par de organizaciones conocidas ampliamente por su labor en esta área.

La manera en que organizaciones de noticias como The Guardian y el New York Times

manejaron las grandes cantidades de datos difundidos por WikiLeaks, es uno de los grandes casos que impulsaron el término. En aquel momento el concepto comenzó a tener un uso más amplio (junto con "el periodismo asistido por computadora") para describir cómo los periodistas utilizaban datos para mejorar su cobertura y amplificar investigaciones profundas de un tema dado.

Hablando con periodistas de datos y estudiosos del periodismo on Twitter, parecería que una de las formulaciones más tempranas de lo que ahora reconocemos como periodismo de datos, en 2006 por Adrian Holovaty, fundador de EveryBlock, un servicio de información que permite a los usuarios saber lo que ha estado sucediendo en su área, en su manzana. En su breve ensayo "Un modo fundamental en que los sitios de diarios tienen que cambiar", sostiene que los periodistas deben publicar datos estructurados procesables por la computadora, junto con el "gran bodoque de texto" tradicional:

Por ejemplo, digamos que un diario ha escrito una historia sobre un incendio local. Poder leer la historia en un celular está bien. Viva la tecnología. Pero lo que realmente quiero poder hacer es explorar los datos en crudo de esa historia, uno por uno, con capas de atribuciones, y una infraestructura para comparar detalles del incendio con incendios anteriores: fecha, momento, lugar, victimas, número de la central de bomberos, distancia de la central de bomberos, nombres y años de experiencia de los bomberos que actuaron, el tiempo que les llevó a los bomberos llegar, e incendios posteriores, cuando sea que sucedan.

¿Pero qué es lo que distingue esto de otras formas de periodismo que usan bases de datos o computadoras? ¿Cómo y en qué medida el periodismo de datos es distinto de otras formas de periodismo del pasado?

Periodismo Asistido por Computadora y Periodismo de Precisión

Usar datos para mejorar los informes y presentar información estructurada (aunque no sea legible por la computadora) al público tiene una larga historia. Quizás lo más relevante en forma inmediata para lo que ahora llamamos periodismo de datos es el periodismo asistido por computadora, conocido por la sigla CAR, que fue el primer abordaje organizado y sistemático del uso de computadoras para recoger y analizar datos de modo de mejorar las noticias.

El CAR fue usado por primera vez en 1952 por CBS para predecir los resultados de la elección presidencial. Desde la década del '60 periodistas (en su mayoría de investigación y de Estados Unidos) han buscado controlar el poder de modo independiente analizado bases de datos de registros públicos con métodos científicos. También conocido como "periodismo de servicio público" los partidarios de estas técnicas con ayuda de computadoras han buscado revelar tendencias, demostrar la falsedad de creencias

populares y revelar injusticias perpetradas por autoridades y corporaciones privadas. Por ejemplo, Philip Meyer trató de demostrar la falsedad de los informes de los disturbios en Detroit de 1967, para reflejar que no eran solo sureños poco educados los que participaban. Las historias de Bill Dedman sobre "El Color del Dinero" en la década del '80 reveló prejuicios raciales sistémicos en las políticas de crédito de las principales instituciones financieras. En su artículo "Lo Que Salió Mal" Steve Doig buscó analizar los patrones de daños del huracán Andrew a comienzos de la década del '90, para comprender el efecto de las políticas y prácticas de desarrollo urbanas fallidas. Los reportes basados en datos han generado valiosos servicios al público y permitido a los periodistas ganar importantes premios.

A comienzos de la década del '70 el término *periodismo de precisión* fue acuñado para describir este tipo de recolección de noticias: "la aplicación de métodos de investigación de las ciencias sociales y de la conducta a la práctica del periodismo" (de "The New Precision Journalism", por Philip Meyer). Se creó el periodismo de precisión para que fuera practicado en las principales instituciones de medios por profesionales formados en periodismo y ciencias sociales. Nació en respuesta al "nuevo periodismo", una forma de periodismo en el que las técnicas del periodismo se aplican a las noticias. Meyer sugiere que lo que se necesita son técnicas científicas de recolección y análisis de datos, en vez de técnicas literarias, para que el periodismo pueda cumplir con su cometido de objetividad y verdad.

Se puede entender el periodismo de precisión como una reacción frente a algunas de las fallas y debilidades comúnmente citadas: la dependencia de informes de prensa (lo que se describió luego como "churnalismo"), el prejuicio en favor de fuentes autorizadas, etc. Meyer ve que estas debilidades derivan de la falta de aplicación de técnicas científicas de información y métodos científicos tales como encuestas y registros públicos. En los '60, el periodismo de precisión fue utilizado para representar a grupos marginales y sus historias.

Según Meyer:

El periodismo de precisión era una manera de expandir el herramental del periodista para hacer que temas antes inaccesibles o sòlo accesibles de modo tosco, estuvieran abiertos a la investigación periodística. Fue especialmente útil para dar voz a grupos minoritarios y disidentes que luchaban por lograr representación.

Un artículo influyente publicado en la década del '80 respecto de la relación entre el periodismo y las ciencias sociales se hace eco del discurso sobre el periodismo de datos. Los autores, dos profesores de periodismo estadounidenses, sugieren que en las décadas de los años '70 y '80, la comprensión del público de lo que son las noticias se amplía, de una concepción más estrecha de "eventos noticiosos" al "reporte situacional" (o informes sobre

tendencias sociales). Por ejemplo, al usar bases de datos de censos o encuestas, los periodistas logran “ir más allá de la información de eventos específicos, aislados, para proveer contexto que les da significado”.

Como era de esperar, la práctica de usar datos para mejorar el periodismo existe desde que hay datos. Como señala Simon Rogers, el primer ejemplo de periodismo de datos en The Guardian data de 1821. Es una tabla de escuelas en Manchester que da la cantidad de estudiantes que asisten a clases y los costos por escuela, Según Rogers, esto ayudó a mostrar el número real de estudiantes que recibían educación gratuita, que era mucho mayor de lo que mostraban las cifras oficiales.

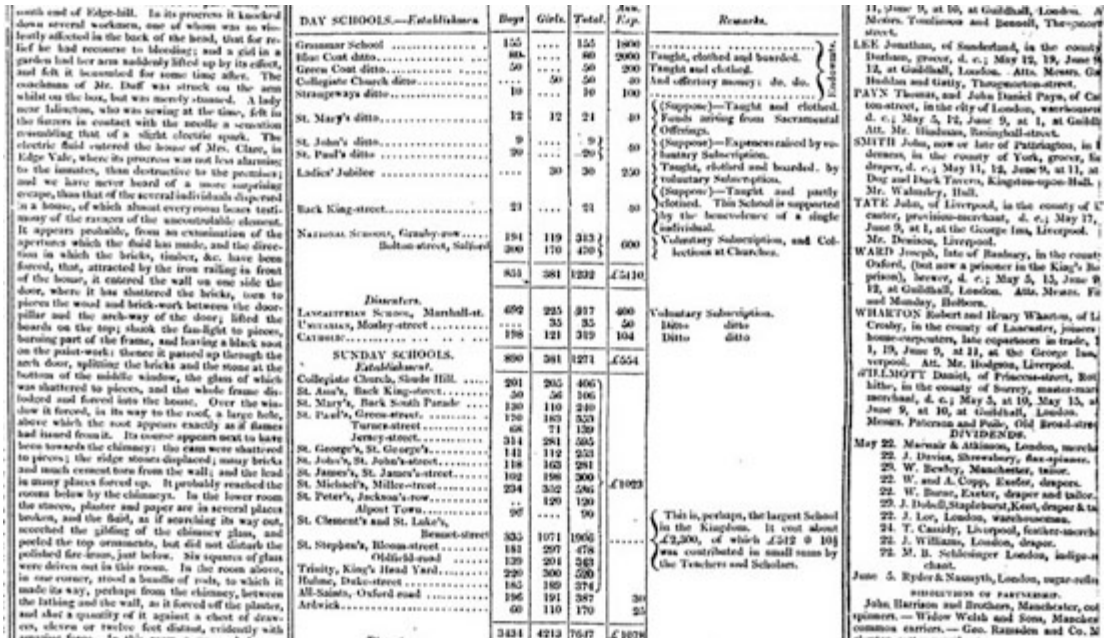


Figure 11. Periodismo de datos en The Guardian en 1821 (The Guardian)

Otro ejemplo temprano en Europa es de Florence Nightingale y su informe clave, "Mortalidad del Ejército Británico", publicado en 1858. En su informe al parlamento usó gráficos para promover mejoras en los servicios de salud para el ejército británico. El más famoso de ellos es su “coxcomb”, una espiral de secciones que representan muertes por mes, en el que se destaca que la gran mayoría de las muertes eran por enfermedades prevenibles, en vez de balas.

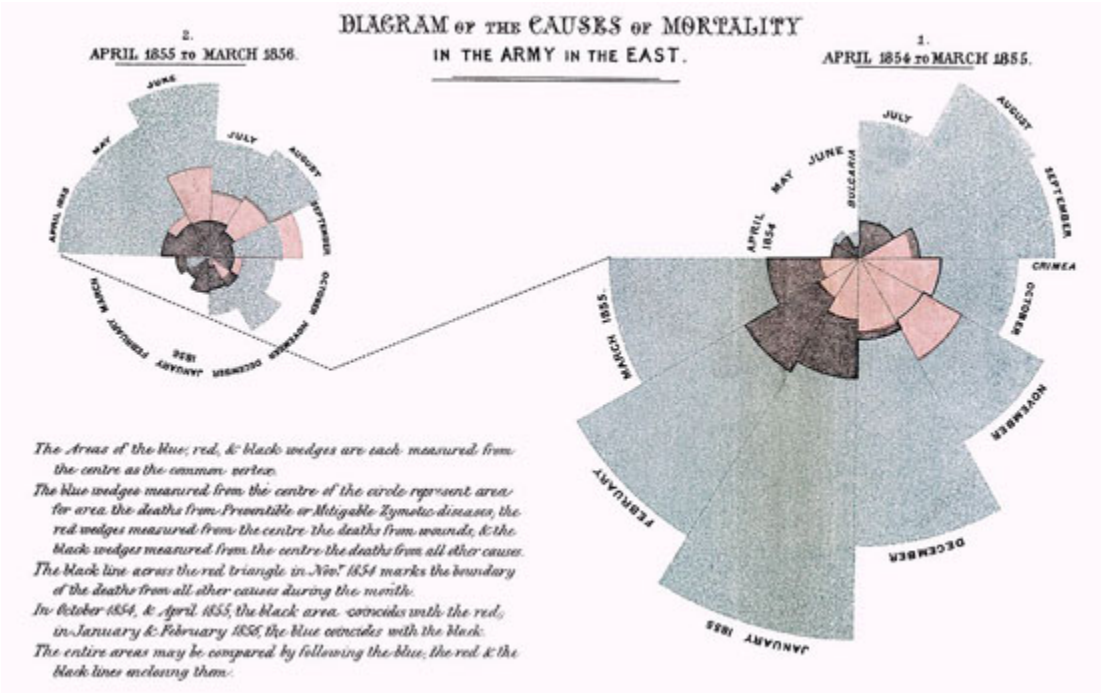


Figure 12. Mortalidad de la armada británica por Florence Nightingale (imagen de Wikipedia)

Periodismo de Datos y Periodismo Asistido por Computadora

En este momento hay un debate sobre “continuidad y cambio” en torno de la etiqueta de “periodismo de datos” y su relación con prácticas previas periodísticas que emplean técnicas computacionales para analizar conjuntos de datos.

Algunos sostienen que hay una diferencia entre CAR y el periodismo de datos. Dicen que CAR es una técnica para recoger y analizar datos como una manera de fortalecer el periodismo (generalmente de investigación), mientras que el periodismo de datos presta atención a la manera en que los datos se ubican en el conjunto del flujo de trabajo periodístico. En este sentido el periodismo de datos presta tanta –y a veces más- atención a los datos mismos, en vez de usarlos simplemente como un medio para encontrar o dar más fuerza a determinadas historias. De allí que encontremos el Datablog de The Guardian o que el Texas Tribune publica juegos de datos junto con los artículos –o incluso solo juegos de datos por sí mismos- para que la gente analice y explore.

Otra diferencia es que en el pasado los periodistas de investigación se encontraban faltos de información respecto de una pregunta que trataban de contestar, o una cuestión que trataban de abordar. Si bien esto sigue sucediendo, también existe una abundancia abrumadora de información con la que los periodistas a veces no saben qué hacer. No saben cómo obtener valor de los datos. Un ejemplo reciente es el Sistema de Información Online Combinada, la mayor base de datos del Reino Unido de información sobre gasto público. Esta base de datos fue durante mucho tiempo un reclamo de los partidarios de la transparencia, pero dejó confundidos y sin respuesta a muchos periodistas cuando se publicó. Como me escribió recientemente Philip Meyer: “Cuando la información era escasa,

la mayor parte de nuestros esfuerzos estaban dedicados a buscarla y recogerla. Ahora que hay información abundante, el procesamiento es más importante".

Por otro lado, algunos sostienen que no hay ninguna diferencia significativa entre el periodismo de datos y el periodismo asistido por computadoras. A esta altura resulta claro que incluso las prácticas más recientes de los medios más novedosos combinan cosas conocidas desde hace tiempo con algo nuevo. Antes que debatir si el periodismo de datos es completamente nuevo, una postura más fructífera sería considerarlo como parte de una tradición más longeva, pero que responde a nuevas circunstancias y condiciones. Aunque no haya una diferencia en cuanto a metas y técnicas, el surgimiento de la etiqueta "periodismo de datos" al comienzo del siglo indica una nueva fase en la que el mero volumen de los datos libremente disponibles online –combinado con herramientas sofisticadas centradas en el usuario, la auto edición y las herramientas de colaboración abierta (crowdsourcing)- permite a más gente trabajar con más datos de modo más fácil que nunca.

El periodismo de datos tiene que ver con la alfabetización masiva en el manejo de datos.

Las tecnologías digitales y la red están cambiando de modo fundamental la manera en que se edita la información. El periodismo de datos es una parte del ecosistema de herramientas y prácticas que han surgido en torno a los sitios y servicios de datos. El citado y el compartir materiales de distintas fuentes es parte de la naturaleza de la estructura de hipervínculos de la red, y la manera en que estamos acostumbrados a navegar la información hoy. Yendo más hacia atrás, el principio que está en la base de la estructura de hipervínculos de la red es el principio de la cita usado en los trabajos académicos. Citar y compartir materiales y sus fuentes y los datos detrás de la historia es una de las maneras básicas en las que el periodismo de datos puede mejorar le periodismo, lo que el fundador de WikiLeaks Julian Assange, llama el "periodismo científico".

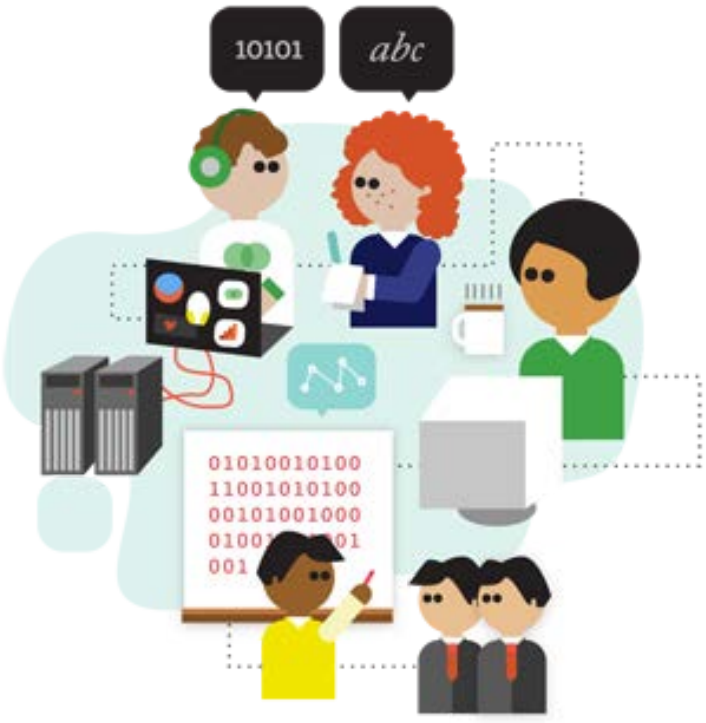
Al permitir a cualquier persona recurrir a fuentes de datos y encontrar información que es relevante, así como verificar afirmaciones y cuestionar los supuestos comunes, el periodismo de datos representa efectivamente la democratización masiva de recursos, herramientas, técnicas y metodologías que antes eran usadas por especialistas, fueran estos periodistas de investigación, científicos sociales, estadísticos, analistas u otros expertos. Si bien actualmente citar vínculos con fuentes de datos es algo específico del periodismo de datos, avanzamos hacia un mundo en el que los datos estarán integrados sin fisuras en el tejido de los medios. Los periodistas de datos tienen un rol importante en cuanto a ayudar a bajar las barreras a la comprensión y el manejo de datos, e incrementar la alfabetización en datos de sus lectores a escala masiva.

En este momento la comunidad creciente de personas que se llaman periodistas de datos es

en gran medida diferente de la comunidad CAR más madura. Esperemos que en el futuro veamos vínculos más fuertes entre estas dos comunidades, del mismo modo que vemos a ONG y organizaciones de medios sociales como ProPublica y el Bureau of Investigative Journalism trabajando junto con medios tradicionales en investigaciones. Mientras la comunidad de periodismo de datos puede tener formas más innovadoras de difundir datos y presentar historias, el enfoque profundamente analítico y crítico de la comunidad CAR es algo de lo que el periodismo de datos podría aprender.

— *Liliana Bounegru, European Journalism Centre*

En la redacción



¿Cómo se ubica el periodismo de datos en las redacciones del mundo? ¿Cómo convencieron importantes periodistas a sus colegas de que es buena idea publicar bases de datos o lanzar aplicaciones de noticias basadas en datos? ¿Los periodistas debieran aprender a escribir código o trabajar en tándem con programadores talentosos? En esta sección analizamos el rol de los datos y el periodismo de datos en la Australian Broadcasting Corporation, la BBC, el Chicago Tribune, The Guardian, el Texas Tribune, y el Zeit Online. Aprendemos cómo descubrir y contratar buenos programadores, cómo atraer a la gente con un tema a través de hackatones y otros eventos, cómo colaborar a nivel internacional y cómo configurar modelos de negocios para periodismo de datos.

Qué contiene este capítulo?

La iniciativa de periodismo de datos de ABC

La Australian Broadcasting Corporation es la difusora pública nacional de Australia. Sus fondos anuales son de alrededor de 1.000 millones de dólares australianos, lo que permite sostener 7 cadenas radiales, 60 estaciones locales de radio, 3 servicios de televisión digital, un nuevo servicio de televisión internacional y una plataforma online con esta oferta siempre en expansión de contenido digital y generada por los usuarios. La última cifra disponible indica que tiene más de 4500 empleados de tiempo completo, y casi el 70% produce contenido.

Somos una difusora nacional muy orgullosa de nuestra independencia, aunque con fondos del estado, por ley estamos claramente separados. Nuestra tradición es de periodismo de servicio público independiente. La ABC es considerada la organización de noticias más confiable del país.

Estos son tiempos que entusiasman; bajo el mando de un director ejecutivo (el ex ejecutivo del diario Mark Scott), se ha alentado a los productores de contenido de ABC a ser “ágiles”, como dice el mantra corporativo.

Por supuesto que es más fácil decirlo que hacerlo.

Pero la iniciativa con la que se buscaba alentar esto, ha derivado en una competencia entre el personal por fondos para desarrollar proyectos multi-plataforma. Así se concibió el primer proyecto de periodismo de datos de la ABC.

En algún momento de comienzos de 2010 me metí en una sesión de propuestas para enfrentar a 3 jefes de “ideas” con mi proyecto.

Lo había estado masticando por un tiempo, llenándome con el periodismo de datos que ofrecía el ya legendario Datablog de The Guardian, y eso solo para empezar.

Mi argumento fue que no había duda de que en 5 años la ABC tendría su propia unidad de periodismo de datos. Era inevitable opiné. Pero la cuestión era cómo llegaríamos a eso y quién iniciaría la tarea.

Aquellos lectores que no conocen la ABC deben pensar en una vasta burocracia construida a lo largo de 70 años. Su oferta primaria siempre fue radio y televisión. Con el advenimiento de un sitio en la red, en la última década esta oferta de contenido se extendió a textos, imágenes fijas y un grado de interactividad hasta entonces inimaginada. El sitio web estaba forzando a la ABC a repensar cómo distribuía la torta (sus fondos) y qué tipo de torta estaba cocinando (contenido).

Por supuesto que es una obra en curso.

Pero otra cosa estaba pasando con el periodismo de datos. Gobierno 2.0: (que como descubrimos se cumple habitualmente en la difusión de datos en Australia) comenzaba a ofrecer nuevas maneras de narrar historias que hasta entonces estaban escondidas en ceros y unos.
Comenté todo esto a las personas que me escuchaban. También dije que necesitábamos identificar nuevos conjuntos de capacidades y formar a periodistas en el manejo de nuevas herramientas. Necesitábamos un proyecto para comenzar a andar.

Y me dieron el dinero.

El 24 de noviembre de 2011, el proyecto multiplataforma de la ABC y ABC News Online salió en vivo con "Coal Seam Gas by the Numbers" (Las cifras de gas metano de carbón).

Figure 1. Coal Seam Gas en números (ABC News Online)

Se componía de 5 páginas de mapas interactivos, visualizaciones de datos y texto.

No era exclusivamente periodismo de datos, sino un híbrido de periodismo que nació de la mezcla de gente del equipo y la historia, que ahora es uno de los temas más calientes en Australia.

La joya era un mapa interactivo que muestra yacimientos y concesiones de gas metano de carbón en Australia. Los usuarios podían buscar por lugar y escoger entre distintos modos para ver concesiones o yacimientos. Usando el zoom los usuarios podían ver quién estaba a cargo de la exploración, la situación del yacimiento y la fecha de perforación. Otro mapa mostraba la ubicación de la actividad en gas metano de carbono con relación a sistemas de aguas subterráneas en Australia.

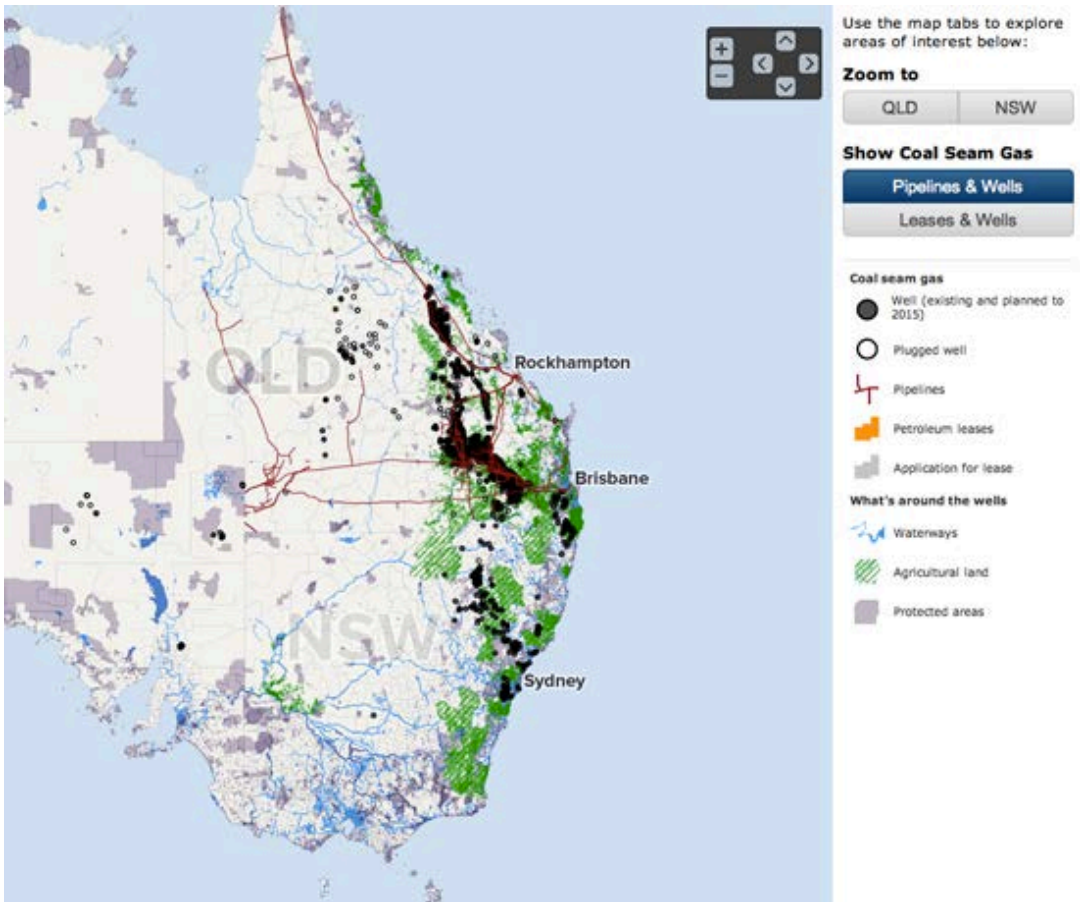


Figure 2. Mapa interactivo de yacimientos y concesiones petroleras en Australia (ABC News Online)

Teníamos visualizaciones de datos que específicamente abordaban la cuestión de la producción de desechos de sal y de agua que serían generados de acuerdo al escenario que se diera.

Otra sección del proyecto investigó el derrame de productos químicos en una cuenca fluvial local.

Nuestro equipo

- Un desarrollador y diseñador de sitios en la red
- Un periodista a cargo
- Un investigador part-time con experiencia en extracción de datos, planillas de cálculos Excel y depuración de datos.
- Un periodista part-time
- Un productor ejecutivo de consultor
- Un consultor académico con conocimientos de búsqueda de datos, visualización de gráficos y capacidades avanzadas de investigación.
- Los servicios de un gerente de proyecto y la asistencia administrativa de la unidad multiplataforma de ABC.
- Algo importante, también teníamos un grupo de referencia de periodistas y otros a los que consultamos según las necesidades.

¿De dónde obtuvimos los datos?

Los datos para los mapas interactivos fueron obtenidos de shapefiles (un tipo común de archivo para datos geo-espaciales) descargados de sitios web del Estado.

Otros datos sobre sal y agua fueron tomados de una variedad de informes.

Los datos sobre deshechos químicos se tomaron de permisos ambientales emitidos por el Estado.

¿Qué descubrimos?

"Las cifras de gas metano de carbón" era ambicioso en contenido y en escala. Lo más importante para mí era determinar qué habíamos aprendido y qué debíamos hacer de modo diferente la próxima vez.

El proyecto de periodismo de datos incorporó a mucha gente que normalmente no se encuentra en ABC: en términos del vulgo, hackers. Muchos no hablábamos el mismo idioma ni entendíamos lo que el otro grupo hace. ¡El periodismo de datos revoluciona las cosas!

Las cosas prácticas:

- La ubicación del equipo en un mismo lugar. Nuestro programador y diseñador estaban fuera del lugar y venían para reuniones. ¡Esto definitivamente no era óptimo! Hay que poner a todos en el mismo cuarto que los periodistas.
- Nuestro consultor de producción ejecutiva también estaba en otro nivel del edificio. Necesitábamos estar mucho más cerca, simplemente por la cuestión de poder pasar a verlo en cualquier momento.
- Escoger una historia que solo se basara en datos.

El cuadro grande: algunas ideas

Las grandes organizaciones de medios tienen que crear capacidad para responder a los desafíos del periodismo de datos. Mi intuición es que hay muchos técnicos y hackers ocultos en los departamentos técnicos de los medios desesperados por salir a la luz. Por lo que necesitamos “reuniones de periodistas y hackers”, talleres donde los geeks secretos (como en agente secreto – se refiere a gente que domina la técnica informática pero no tiene como trabajo las cuestiones técnicas sino que son periodistas, etc.), los periodistas más jóvenes, los programadores, y los diseñadores salen a jugar con periodistas más experimentados para compartir conocimientos y tener orientación. Tarea: ¡descargar este conjunto de datos y atacar!

Ipso facto, el periodismo de datos es interdisciplinario. Los equipos de periodismo de datos

se componen de gente que en el pasado no habría trabajado junta. El espacio digital ha desdibujado las fronteras.

Vivimos en una comunidad política fracturada, desconfiada. El modelo de negocios que antes generaba periodismo independiente profesional –por imperfecto que sea- está al borde del colapso. Debemos preguntarnos, como muchos ya lo hacemos, cómo sería el mundo sin un cuarto poder viable. El periodista e intelectual estadounidense Walter Lippman comentó en la década de '20 que "se reconoce que no puede existir una opinión pública sana sin acceso a las noticias". Esa afirmación es igualmente válida ahora. En el siglo XXI todo el mundo está en la blogósfera. Es difícil diferenciar a los periodistas profesionales del cuentero, el mentiroso, el simulador y quién defiende intereses creados. Cualquier sitio o fuente puede hacerse pasar por creíble, bien presentado y honesto. Las referencias confiables se mueren junto al camino. Y en este nuevo espacio de periodismo basura, los hipervínculos pueden llevar a los lectores interminablemente a otras fuentes más inútiles pero de aspecto brillante que no hacen más que llevar de un hipervínculo a otro en el salón digital de los espejos. El término técnico para esto es que el "macaneo" atonta el cerebro.

En el espacio digital todo el mundo es un narrador, ¿verdad? No. Si el periodismo profesional –y con ello me refiero a aquellos que se dedican a la narración de historias de modo ético, equilibrado, valiente en la búsqueda de la verdad- ha de sobrevivir, entonces el oficio debe reafirmarse en el espacio digital. El periodismo de datos es otra herramienta con la que navegaremos el espacio digital. Es donde mapearemos, daremos vuelta, separaremos, filtraremos, extraeremos y veremos la historia en medio de tantos ceros y unos. En el futuro trabajaremos junto a los hackers, los programadores, los diseñadores. Es una transición que requiere una seria acumulación de capacidades. Necesitamos gerentes de noticias que entiendan la conexión entre lo digital y el periodismo para empezar a invertir en esa construcción.

— *Wendy Carlisle, Australian Broadcasting Corporation*

Periodismo de datos en la BBC

El término "periodismo de datos" puede cubrir una gama de disciplinas y se usa de modos variados en las organizaciones de noticias, por lo que puede ser útil definir lo que queremos decir por "periodismo de datos en la BBC. En general el término cubre proyectos que usan datos para hacer una o más de las siguientes cosas:

- Permitir al lector descubrir información que es relevante para sí mismo.
- Revelar una historia que es llamativa y antes se desconocía
- Ayudar al lector a entender mejor una cuestión compleja.

Estas categorías pueden superponerse, y en un medio online a menudo pueden beneficiarse de algún nivel de visualización.

Que sea personal

En el sitio de BBC News hemos estado usando datos para ofrecer servicios y herramientas para nuestros usuarios desde hace más de una década.

El ejemplo más consistente, que se publicó por primera vez en 1999, es el de nuestras tablas de liga escolar, que usan los datos publicados anualmente por el Estado. Los lectores pueden encontrar las escuelas locales ingresando el código postal, y compararlas con una cantidad de indicadores. Periodistas de educación también trabajan con el equipo de programadores rastreando las historias antes de su publicación.

Cuando empezamos a hacer esto, no existía un sitio oficial que ofreciera al público la posibilidad de explorar datos. Pero ahora que el Departamento de Educación tiene su propio servicio nuestra tarea se concentra más en las historias que surgen de los datos.

El desafío en este área debe ser dar acceso a datos en los que hay un claro interés público. Un ejemplo reciente de un proyecto en el que expusimos un gran conjunto de datos no disponible normalmente para el público en general, fue el informe especial "Todas las muertes en todos los caminos". Ofrecimos una búsqueda por código postal, permitiendo a los usuarios encontrar la locación de todos los accidentes fatales en caminos en el Reino Unido en la última década.

Visualizamos algunos de los datos y cifras principales que surgen de los datos policiales y, para dar al proyecto más dinámica y un rostro humano, hicimos equipo con la London Ambulance Association y BBC London radio y TV para rastrear choques en la capital cuando sucedían. Esto se reportó en vivo online, así como vía Twitter usando el hashtag #crash24, y las colisiones fueron incorporadas al mapa a medida que se informaban.

Herramientas simples

Además de proveer maneras de explorar grandes conjuntos de datos, también hemos tenido éxito en crear herramientas simples, que proveen relevantes recortes de información para los usuarios. Estas herramientas apelan a los que tienen poco tiempo y pueden no querer explorar análisis extensos. La capacidad de compartir fácilmente un dato personal es algo que hemos comenzado a incorporar como estándar.

Un ejemplo simple de este enfoque es nuestro servicio “El mundo en 7000 millones: cuál es su número” publicado coincidentemente con la fecha oficial en la que la población mundial superó los 7000 millones. Ingresando su fecha de nacimiento, el usuario podía saber que

“número” fue en términos de la población global cuando nació y luego compartir ese número vía Twitter o Facebook. La aplicación usa datos aportados por el fondo de desarrollo de la población de la ONU. Fue muy popular y se convirtió en el vínculo más usado en Facebook en el Reino Unido en 2011.

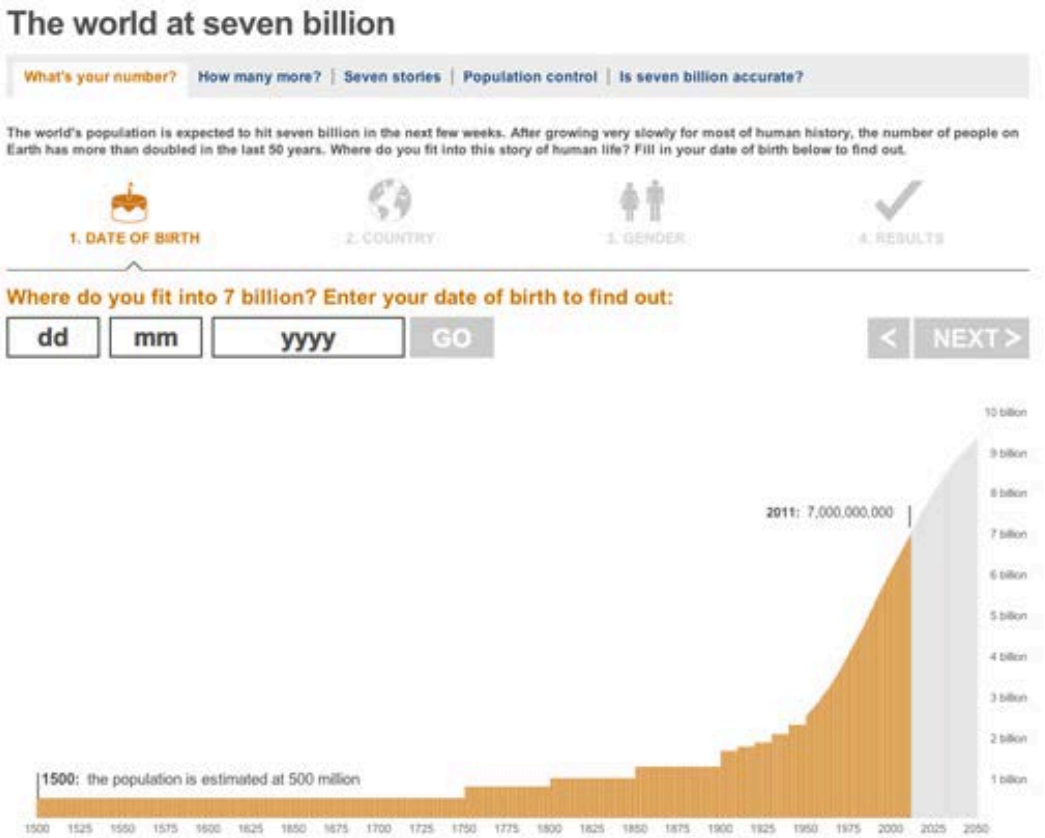


Figure 3. El mundo en 7000 millones (BBC)

Otro ejemplo reciente es el calculador de presupuesto de la BBC, que permitió a los usuarios descubrir en qué medida estarán mejor o peor cuando entre en vigor el presupuesto del Tesoro, y luego compartir la cifra. Hicimos equipo con la firma contable KPMG LLP, que nos dio los cálculos basados en el presupuesto anual y entonces trabajamos duro para crear una interface atractiva que alentara a los usuarios a completar la tarea.

Explotar los datos

¿Pero dónde está el periodismo en todo esto? Encontrar historias en los datos es una definición más tradicional de periodismo de datos. ¿Hay una exclusiva enterrada en la base de datos? ¿Son precisas las cifras? ¿Prueban o no que existe el problema? Estas son todas preguntas que un periodista de datos o un periodista asistido por computadora debe hacerse. Pero puede llevar mucho tiempo estudiar un conjunto de datos muy grande con la

esperanza de encontrar algo llamativo.

En esta área nos ha resultado más productivo asociarnos con programas o equipos de investigación que cuentan con el conocimiento experto y tiempo para investigar una historia. El programa Panorama de actualidad de la BBC pasó meses trabajando con el Centre for Investigative Journalism, recogiendo datos sobre la paga en el sector público. El resultado fue un documental de TV y un informe especial online, "La paga del sector público: las cifras", donde se publicó todos los datos y se visualizaba con análisis sector por sector.

Además de asociarnos con periodistas de investigación, tener acceso a numerosos profesionales con conocimiento especializado es esencial. Cuando un colega del equipo del sector de economía analizó los datos de recortes del gasto publicados por el gobierno, llegó a la conclusión de que los hacía aparecer mayores de lo que eran en realidad. El resultado fue una historia exclusiva, "Encontrar sentido a los datos", complementada por una clara visualización, que ganó un premio de la Royal Statistical Society.

Comprender una cuestión

Pero el periodismo de datos no tiene por qué producir una exclusiva que nadie más ha descubierto. La tarea del equipo de visualización de datos es combinar un gran diseño con una narración editorial clara para ofrecer una experiencia convincente al usuario. Visualizaciones atractivas de los datos apropiados pueden ser usadas para lograr una mejor comprensión de una cuestión o historia, y frecuentemente usamos este enfoque en nuestras narraciones en la BBC. Una técnica utilizada en nuestro rastreador de demandantes es mostrar en un mapa de calor la evolución de los datos con el paso del tiempo para dar una clara visión del cambio.

El servicio de datos "Red de deuda de la Eurozona" explora la red interconectada de créditos entre países. Ayuda a explicar una cuestión complicada de un modo visual, usando color y flechas proporcionales combinadas con un texto claro. Una consideración importante es alentar al usuario a explorar el servicio o seguir una narrativa, sin hacerlo sentir abrumado por las cifras.

Visión general de un equipo

El equipo que produce periodismo de datos para el sitio de BBC News se compone de alrededor de 20 periodistas, diseñadores y programadores.

Además de proyectos y visualizaciones de datos, el equipo produce todas las infografías y recursos multimedia interactivos en el sitio de noticias. En conjunto todo esto forma parte

de una colección de técnicas de narración que llamamos periodismo visual. No tenemos gente identificada específicamente como periodistas de datos, pero todo el personal de redacción del equipo tiene saber usar aplicaciones básicas de planillas de cálculo tales como Excel y Google Docs para analizar datos.

Son centrales a cualquier proyecto de datos las capacidades técnicas, el asesoramiento de nuestros programadores y las capacidades de visualización de nuestros diseñadores. Si bien todos somos en primer lugar periodistas, diseñadores o programadores, seguimos trabajando fuerte para aumentar nuestra comprensión y dominio de todas las áreas de conocimiento.

Los productos centrales para explorar datos son Excel, Google Docs y Fusion Tables. El equipo también ha usado, aunque en menor medida, MySQL, bases de datos Access y Solr para explorar conjuntos de datos mayores; y usó RDF y SPARQL para comenzar a analizar maneras en las que podemos modelar eventos usando tecnologías Linked Data. Los programadores también usan su lenguaje de programación preferido, sea ActionScript, Python, o Perl, para reunir, analizar o desmenuzar en general un conjunto de datos en los que podemos estar trabajando. Perl se utiliza para parte de la edición.

Usamos Google, Bing Maps y Google Earth, junto con ArcMAP de Esri, para explorar y visualizar datos geográficos.

Para gráficos utilizamos la Suite de Adobe incluyendo After Effects, Illustrator, Photoshop y Flash, aunque en estos tiempos rara vez publicamos archivos Flash en el sitio, dado que JavaScript –en particular JQuery y otras bibliotecas de JavaScript tales como Highcharts, Raphael y D3- cada vez más cubren nuestros requisitos de visualización.

— Bella Hurrell and Andrew Leimdorfer, BBC

El equipo de aplicaciones de noticias del Chicago Tribune

El equipo de aplicaciones de noticias del Chicago Tribune es una banda de alegres hackers incrustada en la redacción. Trabajamos en estrecha relación con editores y periodistas para ayudar a: 1) investigar y contar historias; 2) ilustrar historias online y 3) crear recursos de la red siempre actualizados para la buena gente de Chicagolandia.

Es importante que estemos en la redacción. A menudo nuestro trabajo se define en conversaciones cara a cara con periodistas. Saben que nos gusta ayudar a escribir algo que dé vida a un aburrido sitio oficial, desentrañar una parva de PDF, o convertir de otras maneras lo que no es datos, en algo que se pueda analizar. Es lo que ayuda a nuestro equipo a encontrar su rumbo; de este modo nos enteramos de potenciales proyectos de datos cuando se están gestando.

A diferencia de muchos equipos en este campo, el nuestro fue fundado por tecnólogos para quienes el periodismo fue un cambio de carrera. Algunos adquirimos un título de Maestría en Periodismo después de pasar varios años programando para empresas, y otros vinieron de la comunidad de gobierno abierto.

Trabajamos de modo ágil. Para asegurarnos de estar siempre sincronizados, cada mañana comienza con una reunión de 5 minutos de a pie. Frecuentemente programamos de a pares; 2 programadores en un teclado a menudo son más productivos que 2 programadores en 2 teclados. La mayoría de los proyectos no requieren más que una semana de producción, pero en proyectos más largos trabajamos en iteraciones de una semana y mostramos nuestros trabajo a los interesados (periodistas y editores por lo general) toda las semanas. El mantra es “fracasar rápido”. Si uno está haciendo las cosas mal tiene que saberlo lo antes posible, especialmente cuando se programa con un plazo fijo.

Esto de _hackear_de modo iterativo y con plazos tiene un aspecto tremendamente positivo: siempre estamos actualizando nuestro set de herramientas. Cada semana producimos 1 o 2 aplicaciones y luego, a diferencia de los talleres normales de software, podemos sacarlo de nuestra mente y pasar al siguiente proyecto. Es una alegría que compartimos con los periodistas, y cada semana podemos aprender algo nuevo.

Figure 4. El equipo de aplicaciones del The Chicago Tribune (foto por Heather Billings)

Todas las ideas de aplicaciones provienen de periodistas y editores en la redacción. Creo que esto nos diferencia de equipos de aplicaciones de otras redacciones, que frecuentemente producen sus propias ideas. Hemos establecido fuertes relaciones personales y profesionales en la redacción, y la gente sabe que cuando tiene datos viene a nosotros.

Gran parte de nuestro trabajo en la redacción es de apoyo a los periodistas. Los ayudamos a trabajar datos, reconvertir PDF en planillas de cálculo, investigamos en sitios de la red, etc.

Es un servicio que nos gusta dar porque nos permite conocer desde sus inicios el trabajo de datos que se da en la redacción. Parte de ese trabajo se convierte en una aplicación de noticias: un mapa, una tabla o a veces sitios de mayor escala.

Antes vinculábamos la aplicación a la historia escrita, pero eso no resultaba en demasiado tráfico. Actualmente, las aplicaciones aparecen cerca de la parte superior de nuestro sitio y la aplicación tiene un link con la historia, lo que funciona bien tanto para la aplicación como para la historia. Hay una sección del sitio que es para nuestro trabajo, pero no tiene mucho tráfico. Eso no es sorprendente. "Oigan, hoy quiero unos datos" no es algo que se escuche muy seguido.

Nos encanta la cuenta de visitas del sitio y nos encantan las alabanzas de nuestros pares, pero eso no es lo importante. La motivación siempre debe ser el impacto; en la vida de la gente, en las leyes, en hacer que los políticos rindan cuentas y así siguiendo. La pieza escrita habla de la tendencia y la humaniza con unas cuantas anécdotas. ¿Pero qué hace el lector cuando terminó de leer la historia? ¿Está segura su familia? ¿Sus hijos están siendo educados adecuadamente? Nuestro trabajo da sus frutos cuando ayuda a un lector a encontrar su propia historia en los datos. Entre los ejemplos de trabajos impactantes y personalizados que hemos hecho se incluyen las aplicaciones de Informes de seguridad en geriátricos y el Boletín de Calificaciones de Escuelas.

— Brian Boyer, Chicago Tribune

El detrás de escena del Datablog de The Guardian

Cuando lanzamos el Datablog, no teníamos idea a quién podrían interesarle los datos en crudo, las estadísticas y visualizaciones. Como dijo un jefe en mi oficina: "¿Por qué alguien querría eso?".

El Datablog, que yo edito, debía ser un pequeño blog que ofreciera los conjuntos de datos completos que respaldan nuestras historias periodísticas. Ahora consiste en una primera página; búsquedas de datos de gobiernos y desarrollo global; visualizaciones de datos realizadas por artistas gráficos de The Guardian y de toda la red, y herramientas para explorar datos de gasto público. Todos los días usamos planillas de cálculos de Google para compartir los datos completos que respaldan nuestro trabajo; visualizamos y analizamos esos datos y luego los usamos para proveer historias para el diario y el sitio.

Como editor de noticias y periodista trabajando con gráficos, era una extensión lógica del trabajo que ya venía haciendo, acumulando nuevos conjuntos de datos y batallando con ellos para tratar de encontrar sentido a las historias de noticias del día.

La pregunta que me hicieron fue respondida. Han sido unos años increíbles para los datos públicos. Obama abrió los archivos de datos del gobierno de EE.UU. como primer acto

legislativo, y su ejemplo pronto fue seguido por sitios de datos gubernamentales en todo el mundo: Australia, Nueva Zelanda y el sitio del gobierno británico, Data.gov.uk.

Hemos tenido el escándalo de los gastos de los parlamentarios, la pieza más inesperada de periodismo de datos de Gran Bretaña, con el resultado de que Westminster ahora está comprometido a difundir cantidades inmensas de datos todos los años.

Tuvimos una elección general en la que cada uno de los partidos políticos más importantes se comprometió a la transparencia de datos, abriendo nuestros propios archivos de datos al mundo. Los diarios han dedicado valioso centimetraje a la apertura de la base de datos COINS del Tesoro.

Al mismo tiempo, a media que la red produce más y más datos, los lectores de todo el mundo están más interesados que nunca en los datos en crudo detrás de las noticias. Cuando lanzamos el Datablog, creíamos que el público serían programadores buscando crear aplicaciones. De hecho es gente que quiere saber más sobre las emisiones de carbono, inmigración de Europa oriental, el desglose de las muertes en Afganistán, o incluso la cantidad de veces que los Beatles usaron la palabra “amor” en sus canciones (613).

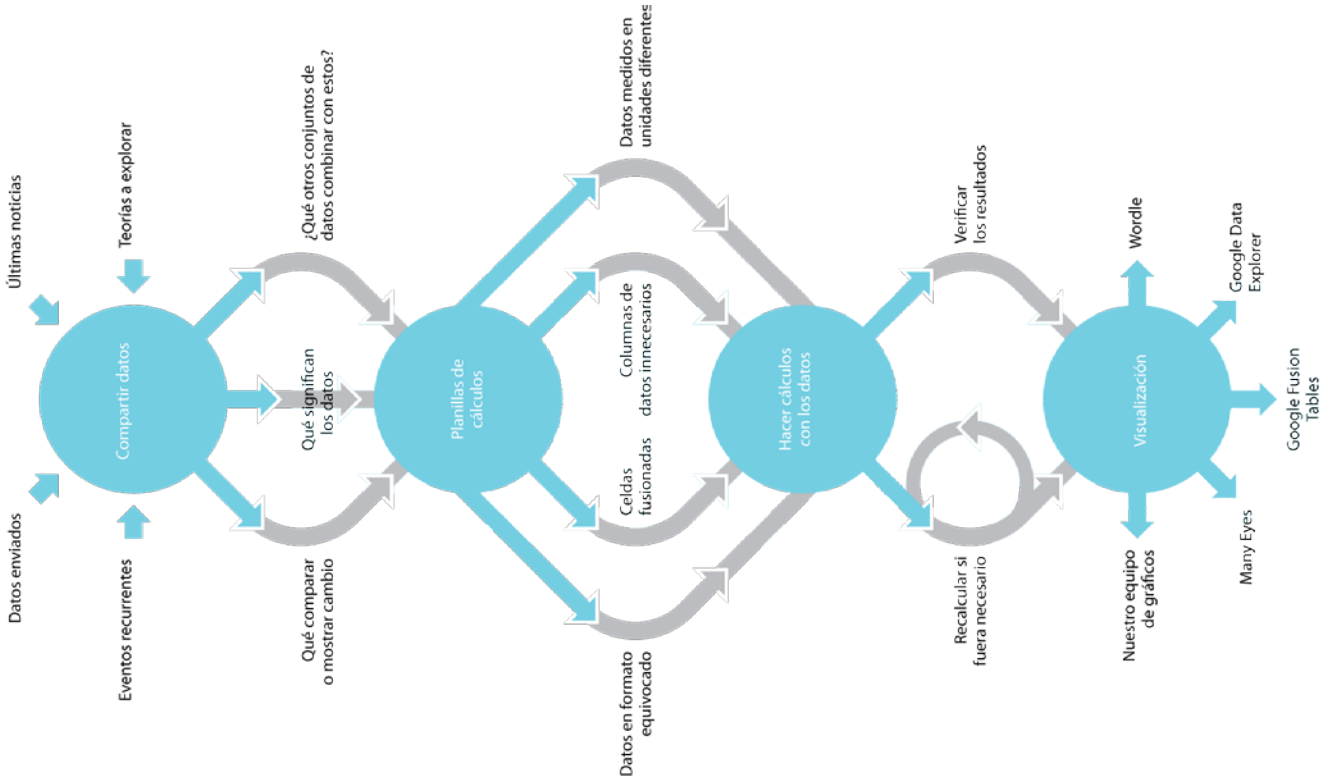


Figure 5. Visualización del proceso de producción de The Guardian Datablog (The Guardian)

Gradualmente el trabajo del Datablog ha reflejado y enriquecido las historias a las que nos enfrentamos. Recurrimos a la opinión del público sobre 458.000 documentos relacionados con los gastos de los parlamentarios y analizamos los datos detallados de lo que sostuvieron los parlamentarios. Ayudamos a nuestros usuarios a explorar bases de datos de gastos detalladas del Tesoro y publicamos los datos detrás de las noticias.

Pero el evento que cambió el juego para el periodismo de datos se dio en la primavera de 2010, comenzando por una hoja de cálculo: 92.201 filas de datos, cada una conteniendo un desglose detallado de un evento militar en Afganistán. Estos fueron los registros de guerra de WikiLeaks. Es decir, la primera parte. Seguirían dos episodios más: Irak y los cables. El término oficial para las primeras dos partes fue SIGACTS: La Base de Datos de Acciones Significativas de las fuerzas armadas de Estados Unidos.

Las organizaciones de noticias dependen mucho de la ubicación y la proximidad con la mesa de noticias. Si uno está cerca, es fácil sugerir historias y ser parte del proceso: inversamente, no estar cerca es literalmente ser ignorado. Antes de WikiLeaks estábamos en un piso diferente, junto con el equipo de Gráficos. Desde WikiLeaks estamos en el mismo piso, junto a la mesa de noticias. Significa que es más fácil para nosotros sugerir ideas a la mesa de noticias, y para los periodistas al otro lado de la redacción pensar en nosotros para que los ayudemos con historias.

No hace mucho los periodistas eran quienes controlaban el acceso a los datos oficiales. Escribíamos historias sobre las cifras y se las transmitíamos a un público agradecido que no estaba interesado en las estadísticas en bruto. La idea de incorporar información en crudo a nuestros diarios era anatema.

Ahora la dinámica ha cambiado hasta el punto de resultar irreconocible. Nuestro rol se está convirtiendo en el de intérpretes; ayudar a la gente a comprender los datos e incluso publicarlos porque son interesantes por sí mismos.

Pero las cifras sin análisis son sólo números, que es donde encajamos nosotros. Cuando el primer ministro de Gran Bretaña sostuvo que los desmanes de agosto de 2011 no tuvieron nada que ver con la pobreza, pudimos cruzar el lugar de residencia de los que hicieron los desmanes con los indicadores de pobreza para mostrar la verdad.

Detrás de todas nuestras historias de periodismo de datos hay un proceso. Está cambiando permanentemente, a medida que vamos usando nuevas herramientas y técnicas. Alguna gente dice que la respuesta es convertirse en una especie de súper hacker, escribir programas y sumergirse en SQL. Uno puede tener esa postura. Pero gran parte del trabajo que hacemos es con Excel.

Primero ubicamos los datos o los recibimos de una variedad de fuentes, de historias de noticias nuevas, datos oficiales, investigaciones de periodistas y así en más. Entonces comenzamos a ver qué podemos hacer con los datos; ¿necesitamos combinarlos con otro conjunto de datos? ¿Cómo podemos mostrar cambios a lo largo del tiempo? Esas planillas de cálculo a menudo tienen que ser muy depuradas, porque todas esas columnas extrañas y celdas fusionadas de modos raros no ayudan a comprender la información. Y eso suponiendo que no es un PDF, el peor formato de datos conocido por la humanidad.

A menudo los datos oficiales vienen con códigos oficiales agregados; cada escuela, hospital, sector, y municipalidad tiene un código de identificación.

Los países también los tienen (el código del Reino Unido por ejemplo es GB). Son útiles porque uno podría querer cruzar conjuntos de datos, y es sorprendente la cantidad de formas de escribir las cosas y arreglos de palabras que pueden trabar eso. Está Birmania y Myanmar, por ejemplo, o Fayette County en Estados Unidos (hay 11 de ellas en estados que van de Georgia a Virginia Occidental). Los códigos nos permiten comparar las cosas comparables.

Al final del proceso está el producto: ¿será una historia o un gráfico o una visualización y qué herramientas usaremos? Nuestras principales herramientas son las gratuitas con las que podemos producir algo rápidamente. Los gráficos más sofisticados son producidos por nuestro equipo de desarrollo.

Esto significa que comúnmente usamos los Google Charts para pequeños gráficos y tortas lineales, o Google Fusion Tables para crear mapas de modo rápido y fácil.

Puede parecer algo nuevo pero no lo es.

En la primera edición del Manchester Guardian (el sábado 5 de mayo de 1821), las noticias estaban en la página trasera, como en todos los diarios de aquellos tiempos. El primer ítem en la primera plana era un aviso de un perro labrador perdido.

Entre las historias y las citas de poemas, un tercio de la contratapa está ocupado con datos. Una tabla completa de los costos de escuelas en la zona nunca antes “presentados al público”, escribe “NH”.

NH quería que se publicaran sus datos porque de otro modo eclesiásticos sin formación informarían sobre los mismos. Su motivación era que “la información que contiene es valiosa; porque si no se sabe en qué medida se extiende la educación... las opiniones que puedan formarse sobre la condición y el progreso futuro de la sociedad serán necesariamente incorrectas”. Dicho de otro modo, si la gente no sabe lo que pasa, ¿cómo puede mejorar la sociedad?

No se me ocurre mejor justificación de lo que estamos tratando de hacer ahora. Lo que hace un tiempo era una historia para la última página, ahora puede estar en primera plana.

— Simon Rogers, the Guardian

Periodismo de datos en el Zeit Online

El proyecto PISA based Wealth Comparison (Comparación de Riqueza basada en PISA, es una visualización interactiva que permite la comparación de niveles de vida en diferentes

países. Utiliza datos del informe de calificación de educación en el mundo, PISA 2009, publicado en diciembre de 2010. El informe se basa en un cuestionario que interroga a estudiantes de 15 años sobre su situación de vida en el hogar.

La idea era analizar y visualizar estos datos para ofrecer una manera única de comparar los estándares de vida en distintos países.

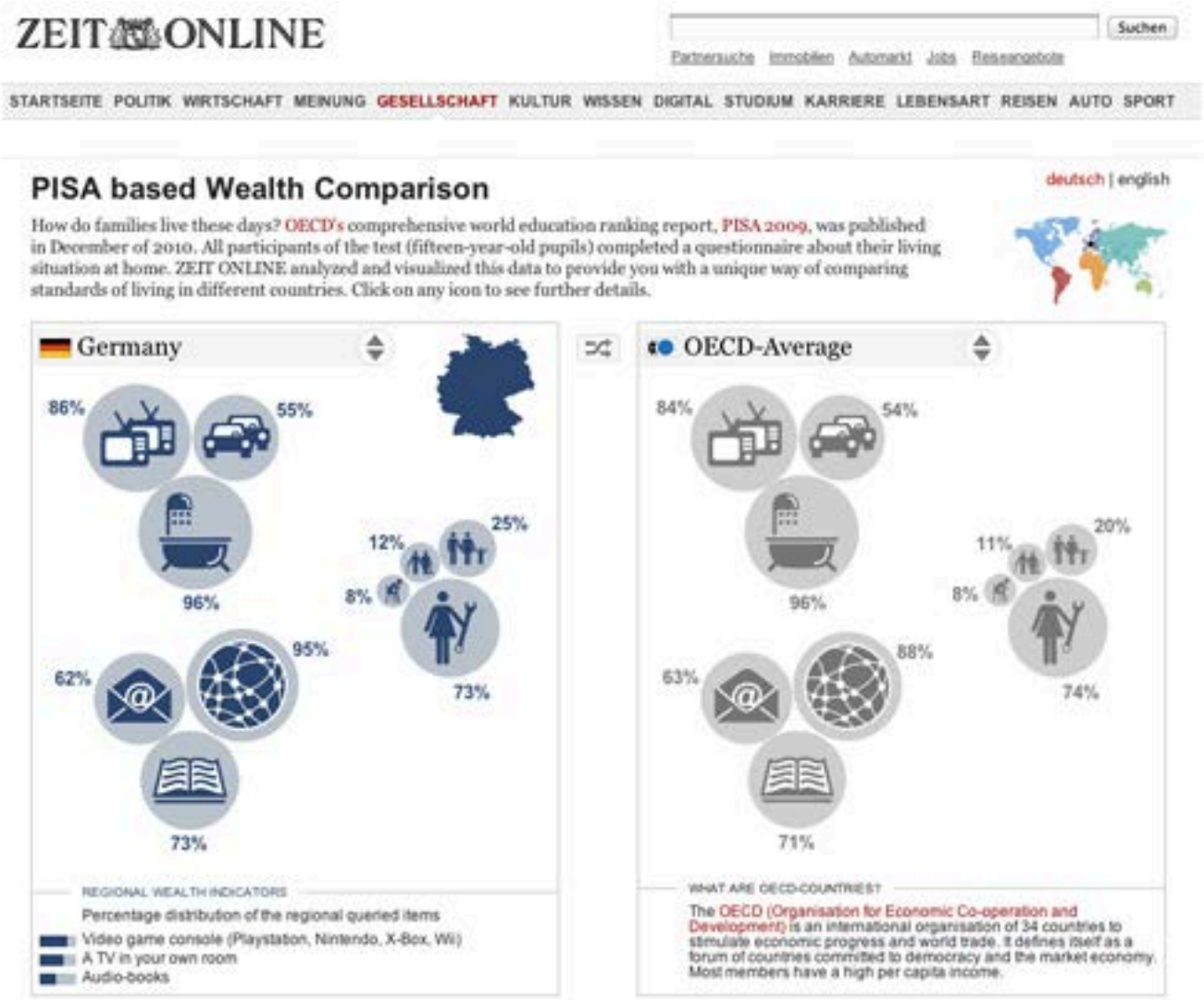


Figure 6. Comparación de riqueza basada en PISA (Zeit Online)

Primero, nuestro equipo editorial decidió qué datos parecían útiles para hacer comparables los niveles de vida y debían visualizarse, incluyendo:

- Riqueza (cantidad de TV, autos y baños disponibles en el hogar)
- Situación familiar (si hay abuelos viviendo con la familia, porcentaje de familias con solo un hijo, desempleo de los padres y el status laboral de la madre)
- Acceso a fuentes de conocimiento (Internet en el hogar, frecuencia de uso de correo electrónico y cantidad de libros que son de propiedad de la familia)
- 3 indicadores adicionales sobre el nivel de desarrollo de cada país.

Con la ayuda del equipo de diseño, estos datos fueron traducidos a íconos explícitos. Se creó un diseño de *front-end* para hacer posible la comparación entre los distintos países viéndolos como si fueran cartas de juego.

A continuación contactamos gente de la Open Data Network de Alemania para encontrar programadores dispuestos a ayudar con el proyecto. Esta comunidad de gente altamente motivada sugirió a Gregor Aisch, como diseñador de información muy talentoso, para que programara la aplicación que haría realidad nuestros sueños (sin usar Flash, lo que era muy importante para nosotros).

Gregor creó una visualización de muy alta calidad e interactiva, con un hermoso estilo de burbuja basado en la Raphaël-Javascript Library.

El resultado de nuestra colaboración fue un interactivo muy exitoso que tuvo mucho tráfico. Es fácil comparar dos países cualesquiera, lo que lo hace útil como herramienta de referencia. Eso significa que podemos volver a utilizarlo en nuestra tarea editorial diaria. Por ejemplo, si estamos cubriendo algo relacionado con las condiciones de vida en Indonesia, podemos rápida y fácilmente incrustar un gráfico comparando las condiciones de vida en Indonesia y Alemania). El conocimiento transferido a nuestro equipo fue una gran inversión para proyectos futuros.

En el Zeit Online encontramos que nuestros proyectos de periodismo de datos nos han traído mucho tráfico y han ayudado a atraer al público de nuevas maneras. Por ejemplo, hubo mucha cobertura de la situación de la planta nuclear en Fukushima luego del tsunami en Japón. Luego de que material radioactivo escapara de la usina, todos fueron evacuados en un radio de 30 kilómetros de la planta. La gente pudo leer y ver muchas cosas sobre la evacuación. Zeit Online encontró una manera innovadora de explicar el impacto de ésto para nuestro público alemán. Preguntamos: ¿Cuánta gente vive cerca de una planta nuclear en Alemania? ¿Cuánta gente vive dentro de un radio de 30 kilómetros? Un mapa muestra cuanta gente tendría que ser evacuada en una situación similar en Alemania. El resultado: mucho tráfico; de hecho el proyecto se expandió como un virus en los medios sociales. Los proyectos de periodismo de datos pueden ser adaptados con relativa facilidad a otros idiomas. Creamos una versión en idioma inglés respecto de la proximidad de las plantas nucleares en Estados Unidos, lo que fue un gran generador de tráfico. Las organizaciones de noticias quieren ser reconocidas como fuentes confiables y autorizadas entre sus lectores. Encontramos que los proyectos de periodismo de datos combinados con permitir a nuestros lectores ver y volver a utilizar los datos en bruto nos da un alto grado de credibilidad.

Por dos años el departamento de Investigación y Desarrollo yel Editor en Jefe del Zeit Online, Wolfgang Blau, han estado promoviendo el periodismo de datos como una manera importante de narrar historias. La transparencia, la credibilidad y la atracción de los usuarios son partes importantes de nuestra filosofía. Por eso el periodismo de datos es una

parte natural de nuestro trabajo actual y futuro. Las visualizaciones de datos pueden aportar valor a la recepción de una historia, y son un modo atractivo de que todo el equipo editorial presente su contenido.

Por ejemplo, el 9 de noviembre de 2011 el Deutsche Bank se comprometió a dejar de financiar a los fabricantes de bombas de racimo. Pero según un estudio de una organización sin fines de lucro Facing Finance, el banco siguió aprobando créditos a productores de bombas de racimo después de hacer esa promesa. Otras visualizaciones basadas en los datos muestran a nuestros lectores los diferentes flujos de dinero. Las distintas partes del Deutsche Bank están ordenadas en la parte de arriba, y las compañías acusadas de estar involucradas en la fabricación de bombas de racimo, abajo. En el medio se representan los créditos individuales siguiendo la línea temporal. Cuando se giran los círculos se ven los detalles de cada transacción. Por supuesto que se hubiera podido contar la historia por medio de un artículo escrito. Pero la visualización permite a nuestros lectores comprender y explorar las dependencias financieras de modo más intuitivo.

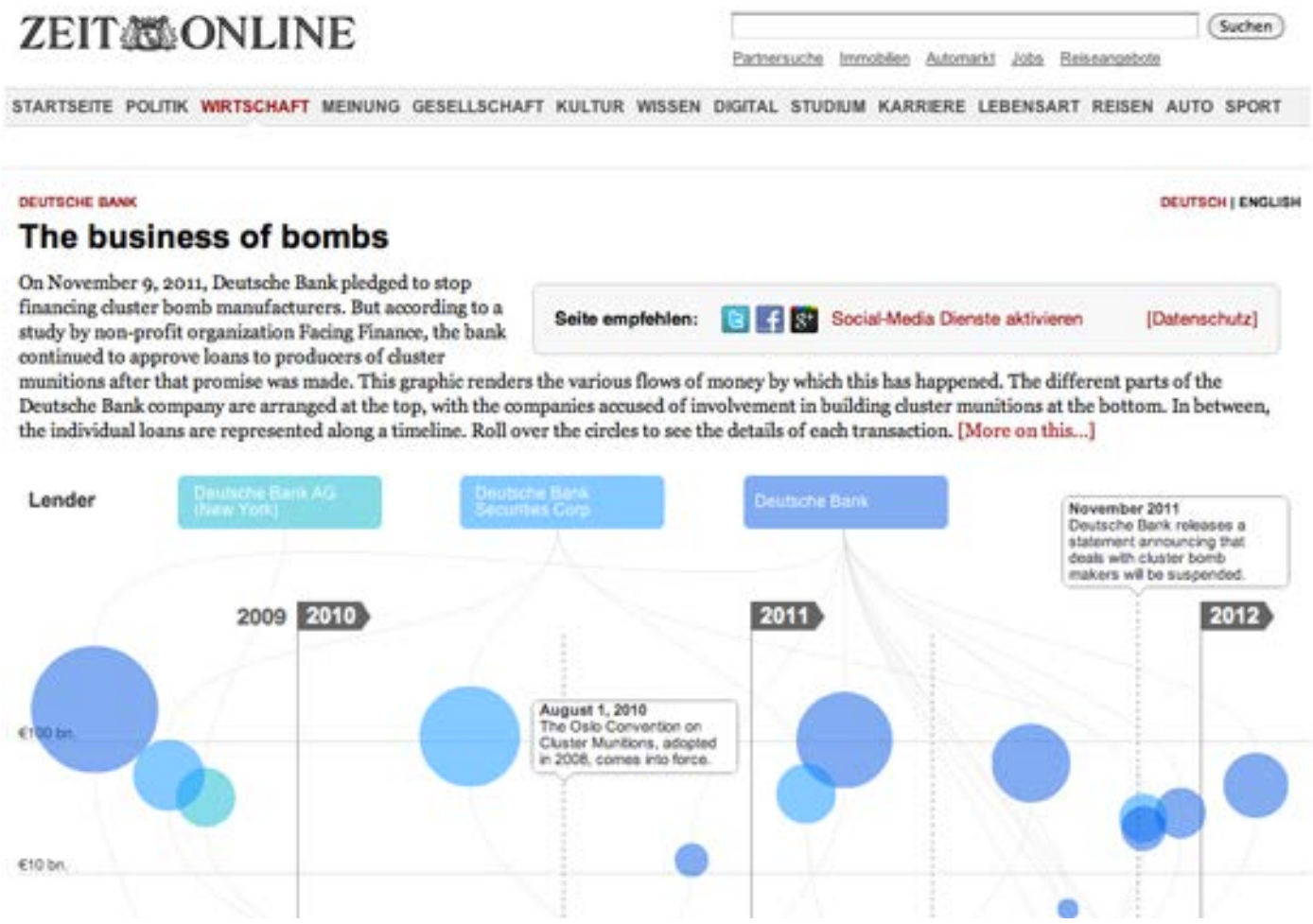


Figure 7. El negocio de las bombas (Zeit Online)

Otro ejemplo: la Oficina Alemana Federal de Estadísticas ha publicado un gran conjunto de datos sobre estadísticas vitales para Alemania, incluyendo el modelado de varios escenarios demográficos hasta 2060. La manera típica de representar esto es una pirámide poblacional, tal como la de la Agencia Federal de Estadísticas.

Con nuestros colegas del departamento de Ciencias, intentamos dar a nuestros lectores una mejor manera de explorar los datos demográficos proyectados, respecto de nuestra sociedad futura. Con nuestra visualización presentamos un grupo estadísticamente representativo de 40 personas de distintas edades desde los años 1950 hasta 2060. Están organizadas en 8 grupos diferentes. Se ve como una foto grupal de la sociedad alemana en distintos momentos. Los mismos datos visualizados en una pirámide poblacional tradicional da solo una visión muy abstracta de la situación, pero un grupo con chicos, gente joven, adultos, y gente mayor significa que los lectores pueden relacionarse más fácilmente con los datos. Basta tocar el botón de play para iniciar un viaje a través de once décadas. También puede ingresar su propio año de nacimiento y su género para convertirse en parte de la foto grupal: podrá ver su propio viaje demográfico a través de las décadas y su propia expectativa de vida.

— *Sascha Venohr, Zeit Online*

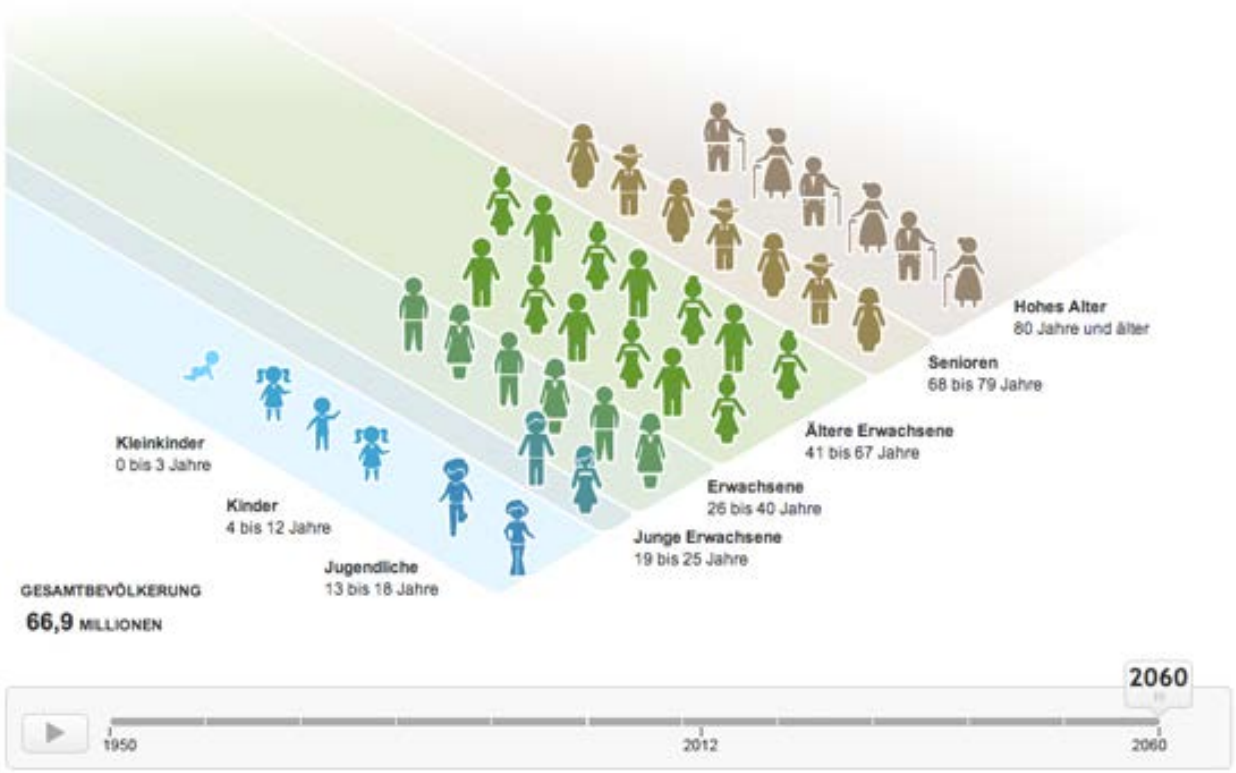


Figure 8. Visualización de datos demográficos (Zeit Online)

Cómo contratar un hacker

Una de las cosas que me preguntan regularmente los periodistas es cómo consigo un programador que me ayude con mis proyectos. No se engañe pensando que este es un proceso en una sola dirección; los hackers con preocupaciones sociales y los analistas de datos a menudo están igualmente interesados en tomar contacto con periodistas.

Los periodistas son usuarios poderosos de herramientas y servicios basados en datos. Desde la perspectiva de los programadores, los periodistas piensan sin esquemas para usar herramientas de datos en contextos que los programadores no siempre tomaron en cuenta (la retroalimentación es invalorable). También ayudan a crear contexto e interés por proyectos, y ayudan a hacer que sean relevantes. Es una relación simbiótica.

Por fortuna, esto significa que, esté pensando en contratar un hacker, o solo busque colaboración con un presupuesto muy restringido, es más que probable que haya alguien interesado en ayudarlo.

¿Entonces cómo se los encuentra? Dice Aron Pilhofer de The New York Times:

Quizás se encuentre con que su organización ya tiene gente con las capacidades que necesita, pero no necesariamente se encuentran en su sala de redacción. Visite los departamentos de tecnología y TI, y es probable que encuentre oro. También es importante apreciar la cultura de los programadores: si encuentra a alguien que tiene una computadora que se ve como la de la Figura 2-9 probablemente tenga un ganador.

Figure 9. Marca del honor: los hackers a menudo son fáciles de descubrir (foto por Lucy Chambers)

Algunas ideas más:

Coloque avisos en sitios de la red que ofrecen puestos de trabajo

Identifique y coloque avisos en sitios que apuntan a programadores que trabajan en distintos lenguajes. Por ejemplo, el Python Job Board.

Listas de correo relevantes para contactos

Por ejemplo las listas de correo NICAR-L y Data Driven Journalism.

Organizaciones relevantes para contactos

Por ejemplo, si quiere buscar datos en la red, puede contactar una organización como Scraperwiki que tienen un gran directorio de programadores confiables y dispuestos.

Súmese a grupos/redes relevantes

Esté atento a iniciativas tales como HACKS/HACKERS que reúnen a periodistas y técnicos. Ahora están surgiendo grupos de Hacks/Hackers en todo el mundo. También podría intentar publicar algo en su newsletter de búesqueda de empleo.

Comunidades de intereses locales

Puede intentar hacer una búsqueda rápida de expertos en determinada cuestión en su zona (por ejemplo “java-script” + “London”). Sitios tales como Meetup.com también pueden ser un gran punto de partida.

Hackatones y competencias

Haya o no dinero de premio involucrado, competencias de aplicaciones y visualizaciones, y días de programación a menudo son un terreno fértil para colaboraciones y lograr contactos.

Pregunte a un técnico

Los técnicos se juntan con otros técnicos. El boca a boca es siempre una buena manera de encontrar buena gente para trabajar.

— Lucy Chambers, Open Knowledge Foundation

Las capacidades de los hackers

Una vez que se encuentra un hacker, ¿cómo se sabe si es bueno? Le preguntamos a Alastair Dant de The Guardian cómo descubrir uno bueno:

Hacen de todo

Cuando hay que cumplir un plazo de entrega es mejor contar con alguien que maneja todas las alternativas, antes que con un maestro especializado en un recurso. Las aplicaciones de noticias requieren manejo de datos, gráficos dinámicos y audacia.

Ven todo el cuadro

El pensamiento holístico le da prioridad al valor narrativo por sobre el detalle técnico. Prefiero escuchar una nota tocada con sentimiento que el virtuosismo sin fin en escalas oscuras. Averigüe si a la persona la hace feliz trabajar junto a un diagramador.

Saben contar una historia

> La presentación narrativa requiere ordenar las cosas en el espacio y el tiempo. Averigüe cual es el proyecto del que se siente más orgulloso, y pídale que le diga cómo fue creado; esto revelará tanto su capacidad de comunicación como su manejo técnico.

Hablan de las cosas que van a hacer

> Crear cosas rápidamente requiere de equipos mixtos trabajando hacia metas comunes. Cada participante debe respetar a sus compañeros y estar dispuesto a negociar. Los obstáculos no previstos a menudo requieren ajustes de planes rápidos y concesiones colectivas.

Se autoeducan

> La tecnología evoluciona rápidamente. Es una lucha mantenerse al día. Habiendo conocido programadores con todo tipo de antecedentes, el rasgo más común es la disposición a aprender cosas nuevas cuando se necesita.

— Lucy Chambers, Open Knowledge Foundation, entrevista con Alastair Dant, Lead Interactive Technologist, the Guardian

Cómo encontrar el programador de sus sueños

La diferencia de productividad entre un programador bueno y uno extraordinario no es lineal, es exponencial. Contratar bien es extremadamente importante. Desgraciadamente, contratar bien es muy difícil. Es bastante difícil evaluar candidatos si uno no es un gerente técnico con experiencia. A eso hay que agregar los sueldos que las organizaciones de noticias pueden pagar y entonces es todo un desafío.

En el Tribune, reclutamos con dos ángulos: el atractivo emotivo y el atractivo técnico. El atractivo emocional es que el periodismo es esencial para que una democracia funcione. Si trabaja aquí puede cambiar el mundo. Técnicamente, promocionamos todo lo que puede aprender. Nuestros proyectos son pequeños, rápidos e iterativos. Cada proyecto tiene un nuevo conjunto de herramientas, un nuevo idioma, un nuevo tópico (seguridad anti incendios, el sistema jubilatorio) que debe aprender. La redacción es un cruce de caminos. Nunca dirigí un equipo que haya aprendido tanto y tan rápido como nuestro equipo.

En cuanto a donde buscar, hemos tenido mucha suerte encontrando grandes hackers en la comunidad de gobierno abierto. La lista de correo Sunlight Labs es donde los locos por la tecnología que quieren hacer el bien, y tienen empleos aburridos, se encuentran por la noche. Otro recurso potencial es Code for America. Cada año un grupo de individuos emergen de CfA buscando su siguiente gran proyecto. A esto se agrega que CfA tienen un proceso de selección riguroso; ya han evaluado los candidatos por usted. Hoy en día los periodistas interesados en programación también salen de las escuelas de periodismo. Están verdes tienen toneladas de potencial.

Por último, no basta con contratar programadores. Se necesita gerencia técnica. Un programador solitario (especialmente si acaba de salir de la escuela de periodismo sin experiencia de trabajo) va a tomar muchas malas decisiones. Incluso el mejor programador, si se lo deja solo, optará por trabajo técnicamente interesante, por sobre lo que es más importante para su público.

Llame a este puesto Editor de aplicaciones de noticias, gerente de proyecto, lo que sea. Al igual que los redactores, los programadores necesitan editores, mentores, y alguien que los apure para que tengan el software listo a tiempo.

— *Brian Boyer, Chicago Tribune*

Ayuda externa de expertos a través de hackatones

En marzo de 2010, la organización SETUP de cultura digital con sede en Utrecht organizó un evento llamado Hacking Journalism. El evento fue organizado para alentar una mayor colaboración entre programadores y periodistas.

“Organizamos hackatones para producir aplicaciones atractivas, pero no podemos reconocer historias interesantes en los datos. Lo que creamos no tiene relevancia social” dijeron los programadores. “Reconocemos la importancia de periodismo de datos, pero no tenemos las capacidades técnicas para crear las cosas que queremos”, dijeron los periodistas.

Figure 10. Periodistas y programadores en RegioHack (foto de Heinze Havinga)

Trabajando para un diario regional no había dinero o incentivos para contratar un programador para la redacción. El periodismo de datos seguía siendo algo desconocido para los diarios holandeses en aquel tiempo.

El modelo de hackatones era perfecto; un ambiente relajado para colaboración, con abundante pizza y bebidas energizantes. RegioHack fue un hackatón organizado por mi empleador, el diario regional De Stentor, nuestra publicación hermana TC Tubantia y Saxion Hogescholen Enschede que ofreció el lugar para el evento.

La organización era así: todos podían anotarse para un hackatón de 30 horas. Nosotros dábamos la comida y las bebidas. Apuntábamos a 30 participantes, que dividimos en 6 grupos. Estos grupos se concentrarían en distintos tópicos, tales como crimen, salud, transporte, seguridad, envejecimiento y poder. Para nosotros, los 3 objetivos principales para este evento eran los siguientes:

Encontrar historias

: Para nosotros el periodismo de datos es algo nuevo y desconocido. La única manera que podemos demostrar su utilidad es a través de historias bien armadas. Planeamos producir al menos 3 historias de datos.

Conectar gente

: Nosotros los periodistas no sabemos cómo se hace periodismo de datos y no pretendemos saberlo. Al colocar periodistas, estudiantes y programadores en un cuarto por 30 horas, queremos que compartan conocimientos y visiones.

Organizar un evento social

: Los diarios no organizan muchos eventos sociales, ni hablemos de hackatones. Queríamos experimentar cómo un evento de esas características puede dar resultados. De hecho hubiera podido ser incómodo: 30 horas con extraños, mucha jerga, golpearse la cabeza contra preguntas básicas y encontrar el terreno en el que cada uno se siente cómodo. Al convertirlo en un evento social (la pizza y las bebidas energizantes), queríamos crear un ambiente en el que periodistas y programadores pudieran sentirse cómodos y colaborar efectivamente.

Antes del evento, TC Tubantia realizó una entrevista con la viuda de un policía que escribió un libro sobre los años de servicio de su marido. También tenía un documento con todos los asesinatos registrados en la parte este de Holanda, mantenido por su marido desde 1945. Normalmente, publicaríamos este documento en nuestro sitio. Esta vez hicimos un tablero usando el software Tableau. También escribimos en el blog acerca de cómo se juntó todo esto en nuestro sitio RegioHack.

Durante el hackatón, un grupo de proyecto abordó el tema del desarrollo de escuelas y el envejecimiento de nuestra región. Al hacer una visualización de proyecciones futuras vimos qué ciudades estarían en problemas luego de unos años de caída de la matrícula. Teniendo

esto presente, hicimos un artículo sobre la manera en que esto afectaría las escuelas en nuestra región.

También iniciamos un proyecto muy ambicioso llamado De Tweehondererd van twente (en español Los Doscientos de Twente) para determinar quién tenía más poder en nuestra región y crear una base de datos de la gente más influyente. A través de un cálculo al estilo Google –quien tiene la mayor cantidad de vínculos con organizaciones poderosas- se compondrá una lista de gente influyente. Esto podría llevar a una serie de artículos, pero también es una herramienta poderosa para periodistas. ¿Quién tiene vínculos con quién? Se puede hacer preguntas a esta base de datos y usarla en la rutina diaria. Además, esta base de datos tiene valor cultural. Los artistas ya preguntaban si podían usar esta base de datos cuando estuviera terminada, para hacer instalaciones de arte interactivo.

Figure 11. Nuevas comunidades en torno al periodismo de datos (foto por Heinze Havinga)

Luego de RegioHack, advertimos que los periodistas consideraban al periodismo de datos como una adición viable al periodismo tradicional. Mis colegas siguieron usando y creando en base a las técnicas aprendidas ese día para generar proyectos más ambiciosos y técnicos, tales como una base de datos de los costos administrativos de la construcción de viviendas. Con estos datos, hice un mapa interactivo en Fusion Tables. Pedimos a nuestros lectores que jugaran con los datos y obtuvimos los resultados con la colaboración de la audiencia (crowdsourcing), por ejemplo. Luego de recibir muchas preguntas respecto de cómo se hace un mapa en Fusion Tables, también grabé un video tutorial.

¿Qué aprendimos? Aprendimos mucho, pero también encontramos muchos obstáculos. Reconocimos estos 4:

¿Por dónde comenzar, pregunta o datos?

> Casi todos los proyectos se trababan en la búsqueda de información. En la mayoría de los casos comenzaban con una pregunta periodística. ¿Y entonces? ¿Qué datos hay disponibles? ¿Dónde pueden encontrarse? ¿Y cuando encuentre estos datos podré responder su pregunta? Los periodistas por lo general saben dónde pueden encontrar información cuando investigan para un artículo. En el periodismo de datos, la mayoría de los periodistas no saben qué información está disponible.

Poco conocimiento técnico

> El periodismo de datos es una disciplina bastante técnica. A veces hay que filtrar, otras veces hay que hacer algo de programación para ver los resultados. Para hacer periodismo de datos se necesitan dos cosas: la visión periodística de un periodista experimentado y el conocimiento técnico de alguien que maneje todas las técnicas digitales. Durante RegioHack esta no era una presencia común.

¿Es noticia?

> Los participantes usaron principalmente un conjunto de datos para descubrir noticias, en vez de buscar interconexiones entre distintas fuentes. El motivo de esto es que se necesita algo de conocimiento estadístico para verificar noticias del periodismo de datos.

¿Cómo es la rutina?

> Todo lo anterior se resume en que no hay rutina. Los participantes tienen algunas capacidades pero no saben cómo, ni cuándo usarlas. Uno de los periodistas lo comparó con hacer una torta. “Tenemos los ingredientes: harina, huevos, leche, etcétera. Lo tiramos en una bolsa, la sacudimos y esperamos que salga una torta”. Tenemos todos los ingredientes, pero no conocemos la receta.

¿Y ahora qué hacemos? Nuestras primeras experiencias con el periodismo de datos podrían ayudar a otros periodistas o programadores que aspiren a ingresar en el mismo campo de trabajo, y estamos trabajando para producir un informe.

También estamos considerando cómo continuar RegioHack en forma de hackatón. Nos resultó divertido, educativo y productivo, y una gran introducción al periodismo de datos.

Pero para que el periodismo de datos funcione tenemos que integrarlo en la redacción. Los periodistas tienen que pensar en datos, además de citas, declaraciones de prensa, reuniones de consejos, etc. Al hacer RegioHack demostramos a nuestro público que el periodismo de datos no son solo palabras. Podemos escribir artículos mejor informados y más claros, presentando a los lectores artículos diferentes impresos y online.

— *Jerry Vermanen, NU.nl*

Seguir el rastro del dinero: colaboración internacional

Los periodistas de investigación y los ciudadanos interesados en descubrir el crimen organizado y la corrupción que afecta las vidas de miles de millones en todo el mundo cada día que pasa tienen acceso sin precedentes a información. Gobiernos y otras organizaciones colocan inmensos volúmenes de datos online y parece que la tan necesaria información está cada vez más al alcance de todos. Pero, al mismo tiempo, funcionarios corruptos en gobiernos y grupos del crimen organizado están haciendo todo lo que pueden para ocultar información para que no se conozcan sus crímenes. Se esfuerzan por mantener a la gente a oscuras mientras concretan negocios sucios que causan problemas a la sociedad, en todos sus niveles, y llevan a conflictos, hambrunas u otras crisis.

Es el deber de los periodistas investigadores exponer tales faltas y, al hacerlo, trabar los mecanismos corruptos y criminales.

Figure 12. El Tablero Investigativo (OCCRP)

Hay 3 guías principales que, si se siguen, pueden llevar a un buen periodismo cuando se investigan grandes actos de corrupción y criminales, incluso en los medios más austeros:

Piense en buscar fuera de su país

> En muchas instancias es mucho más fácil obtener información del extranjero que dentro del país en el que opera el periodista de investigación. La información

obtenida del extranjero vía bases de datos de otros países o usando las leyes de acceso a la información de otras naciones puede ser justo lo que necesita para armar el rompecabezas investigativo. Además, los criminales y los funcionarios corruptos no guardan su dinero en el lugar de donde lo robaron. Prefieren depositarlo en bancos extranjeros o invertirlo en otros países. El crimen es global. Las bases de datos que ayudan al periodista de investigación a rastrear dinero en todo el mundo pueden encontrarse en muchos lugares en Internet. Por ejemplo, el Investigative Dashboard permite a los periodistas seguir el rastro del dinero entre países.

Haga uso de redes periodísticas de investigación existentes

Periodistas de investigación de todo el mundo se agrupan en organizaciones tales como The Organized Crime and Corruption Reporting Project, The African Forum for Investigative Reporting, The Arab Reporters for Investigative Journalism y The Global investigative Journalism Network. Los periodistas también pueden usar plataformas de periodismo profesional tales como IJNet, donde se intercambia información global relacionada con periodismo todos los días. Muchos de los periodistas agrupados en redes trabajan en cuestiones similares y enfrentan situaciones similares, por lo que tiene mucho sentido intercambiar información y métodos. Hay listas de correo electrónico o grupos de redes sociales vinculados a estas redes, por lo que es fácil tomar contacto con colegas periodistas para pedir información o consejos. También pueden obtener ideas para historias a investigar en tales foros y listas de correo electrónico.

Hacer uso de la tecnología y colaborar con hackers

El software ayuda a los periodistas de investigación a acceder y procesar información. Varios tipos de software ayudan al investigador a no dejarse distraer por el ruido, a buscar y encontrar sentido a grandes volúmenes de datos y a encontrar los documentos indicados para descubrir la historia. Hay muchos programas de software que pueden usarse como herramientas para analizar, recoger o interpretar información y, lo que es más importante, los periodistas de investigación tienen que ser conscientes de que hay cantidades de programadores dispuestos a ayudar si se les pide. Estos programadores o hackers saben cómo obtener y manejar información y pueden ayudar mucho con el esfuerzo investigativo. Estos programadores, algunos de ellos miembros de movimientos globales en favor de la apertura de los datos, pueden convertirse en aliados invalorables en la lucha contra el crimen y la corrupción, son capaces de asistir a los periodistas en la recolección y análisis de la información.

Un buen ejemplo de una interfaz entre programadores y ciudadanos es ScraperWiki, un sitio en el que los periodistas pueden pedir ayuda a programadores en la extracción de datos

de sitios en la red. Investigative Dashboard tiene una lista de herramientas listas para usar que podrían ayudar a los periodistas a recoger, dar forma y analizar datos.

La utilidad de las guías mencionadas se ha hecho visible en muchas instancias. Un buen ejemplo es el trabajo de Khadija Ismayilova, una reportera de investigación Azerí muy experimentada que trabaja en un medio austero en lo que se refiere a acceso a información. Ismayilova ha superado obstáculos diariamente para ofrecer al público azerí información buena y confiable. En junio de 2001, Khadija Ismayilova, que trabajaba en la oficina de Baku Radio Europa Libre/Radio Libertad (conocida por las siglas RFE/RL), informó que las hijas del presidente Azerí, Ilham Aliyev, manejaban secretamente una compañía de telecomunicaciones en rápido ascenso, Azerfon, a través de compañías offshore con sede en Panamá. La compañía tiene casi 1.700.000 de suscriptores, cubre el 80 por ciento del territorio del país, y (en aquel tiempo) era el único proveedor de servicios 3G para Azerbaiján. Ismayilova pasó 3 años tratando de descubrir quienes eran los dueños de la compañía de telecomunicaciones, pero el gobierno se negaba a dar información sobre los accionistas y mintió numerosas veces sobre sus dueños. Incluso llegaron a decir que la compañía era propiedad de Siemens AG con sede en Alemania, cosa que ha sido negada directamente por esa corporación. La reportera azerí logró descubrir que Azerfon era propiedad de unas cuantas compañías privadas con sede en Panamá. Esto pareció ser una vía muerta para su informe hasta que recibió ayuda del exterior. A comienzos de 2011 Ismayilova supo a través del Investigative Dashboard que las compañías con sede en Panamá pueden ser rastreadas a través de una aplicación desarrollada por el programador y activista Dan O'Huiginn. Con esta herramienta finalmente logró sacar a luz el hecho de que las dos hijas del presidente estaban involucradas en la compañía de telecomunicaciones a través de las empresas con sede en Panamá.

O'Huiginn creó una herramienta que ayudó a periodistas de todo el mundo a informar sobre corrupción: Panamá, un paraíso offshore bien conocido, ha sido ampliamente utilizado por varios funcionarios corruptos como un lugar para ocultar dinero robado (desde compinches del ex presidente egipcio Hosni Mubarak hasta funcionarios sucios de los Balcanes o en América Latina). Lo que el programador-activista ha hecho se conoce como *scraping* (literalmente raspado y que se traduce como extraer datos, n. del t.) de la red: un método que permite la extracción y el reordenado de información para que pueda ser usada por investigadores. O'Huiginn extrajo información del registro de compañías de Panamá porque este registro, aunque abierto solo permite búsquedas si el periodista de investigación conoce el nombre de la compañía comercial que busca. Esto limitaba las posibilidades de investigaciones, ya que los periodistas generalmente buscan nombres de personas para rastrear sus activos. Extrajo los datos y creó un nuevo sitio donde también son posibles búsquedas basadas en nombres. El nuevo sitio permitió a periodistas de investigación de muchos países buscar información, tomando como referencia nombres de funcionarios en

gobiernos y parlamentos, y verificar si poseían en secreto corporaciones en Panamá (tal como sucedía con la familia del presidente de Azerbaiján).

Hay otras ventajas del uso de las guías destacadas más arriba, además de tener mejor acceso a información. Una de ellas tiene que ver con minimizar el daño y asegurar mejor protección para los investigadores que trabajan en ambientes hostiles. Esto se debe al hecho que cuando se trabaja en una red, el periodista no está solo; el periodista de investigación trabaja con colegas en otros países, por lo que es más difícil para los criminales descubrir quién es responsable de que se vean expuestos sus crímenes. Como resultado de ello a los gobiernos y funcionarios corruptos les resulta mucho más difícil atacarlos.

Otra cosa a tener en cuenta es que la información que no parece muy valiosa en una zona geográfica puede ser crucial en otra. El intercambio de información a través de redes de investigación puede llevar a sacar a luz historias muy importantes. Por ejemplo, la información de que un rumano fue atrapado en Colombia con 1 kilogramo de cocaína probablemente no sea una noticia de primera plana en Bogotá, pero podría ser muy importante para el público rumano si un periodista local logra descubrir que la persona que fue atrapada con el narcótico trabaja para el gobierno de Bucarest.

El periodismo de investigación eficiente es el resultado de la cooperación entre periodistas de investigación, programadores y otros que quieren usar datos para contribuir a crear una sociedad global más limpia y más justa.

— Paul Radu, Organized Crime and Corruption Reporting Project

Nuestras historias aparecen en forma de código

OpenDataCity fue fundado hacia fines de 2010. Por entonces no pasaba nada con lo que uno podría llamar periodismo de datos en Alemania.

¿Por qué lo hicimos? Muchas veces habíamos escuchado a gente trabajando para diarios y a gente de radio y televisión decir: “No estamos listos para crear una unidad de periodismo de datos en nuestra redacción. Pero con gusto tercerizaríamos esto a otros”.

Hasta donde sabemos somos la única compañía que se especializa exclusivamente en periodismo de datos en Alemania. Actualmente somos 3: dos somos periodistas y uno tiene un profundo conocimiento de la programación y la visualización. Contamos con un puñado de hackers, diseñadores y periodistas que trabajan por cuenta propia.

En los últimos 12 meses hemos encarado 4 proyectos de periodismo de datos con diarios y hemos ofrecido capacitación y consultoría a trabajadores de medios, científicos y escuelas de periodismo. La primera aplicación que hicimos fue TAZ, una herramienta interactiva sobre ruido en aeropuertos referida al nuevo aeropuerto de Berlín. Nuestro siguiente

proyecto notable fue una aplicación sobre retención de datos de uso de teléfonos móviles de un político alemán con ZEIT online. Por esto ganamos un premio Grimme Online, un premio Lead en Alemania, y un premio de Periodismo Online de la Online Journalism Association en Estados Unidos. En momentos que escribimos estas líneas tenemos varios proyectos encaminados, que van desde infográficos interactivos más simples hasta el diseño y el desarrollo de un programa de periodismo de datos intermedio.

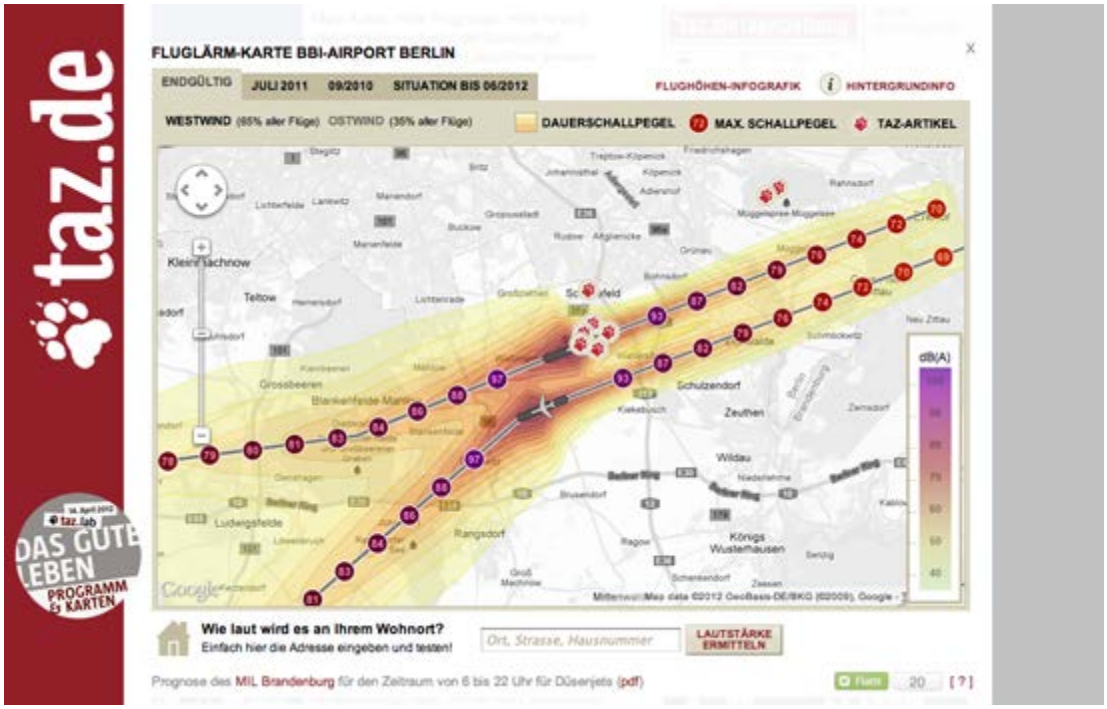


Figure 13. Mapa de ruido en aeropuerto (Taz.de)

Por supuesto que ganar premios ayuda a la reputación. Pero cuando hablamos con los editores, que tienen que aprobar los proyectos, nuestro argumento a favor de invertir en periodismo de datos no tiene que ver con ganar premios. Más bien es ganar audiencia en períodos más prolongados de modo sustentable. Es decir, crear cosas por su impacto de largo plazo, no por el golpe periodístico del momento, que a menudo se olvida en pocos días.

A continuación presentamos 3 argumentos que hemos usado para alentar a editores a abordar proyectos de más largo plazo:

Los proyectos de datos no envejecen

> De acuerdo a su diseño, se puede agregar nuevo material a las aplicaciones de periodismo de datos. Y no son solo para los usuarios, sino que pueden ser usados internamente para hacer informes y análisis. Si le preocupa que esto signifique que sus competidores también se beneficien de su inversión, puede resguardar algunos recursos o datos para uso interno solamente.

Puede apoyarse en su trabajo pasado

> Cuando aborda un proyecto de datos a menudo crea tramos de programas que pueden ser reutilizados o actualizados. El siguiente proyecto podría llevar la mitad del tiempo, porque sabe mucho mejor qué hacer (y qué no) y tiene tramos que puede reutilizar.

El periodismo de datos se pago solo

> Los proyectos basados en datos son más baratos que las campañas de marketing tradicionales. Las medios online a menudo invierten en cosas como Optimización de Motores de Búsqueda (OMB) y Marketing de Motores de Búsqueda (MMB). Un proyecto de datos ejecutado normalmente generará muchos clics y comentarios y puede extenderse como un virus en la red. Los editores comúnmente pagan menos por esto que por tratar de generar la misma atención a través del MMB.

Nuestro trabajo no es muy distinto del de otras agencias de nuevos medios: proveer aplicaciones o servicios para medios informativos. Pero quizás difiramos en que nos vemos en primer lugar como periodistas. A nuestros ojos los productos que entregamos son artículos o historias, aunque no se transmitan con palabras, imágenes, audio o video, sino en código. Cuando hablamos de periodismo de datos, tenemos que hablar de tecnología, software, dispositivos y cómo contar una historia con ellos.

Para dar un ejemplo, acabamos de trabajar en una aplicación que obtiene datos en tiempo real a través de un programa que extrae (*scrapea*) información del sitio del ferrocarril alemán, lo que nos permite desarrollar un monitor ferroviario interactivo para Süddeutsche Zeitung que muestra las demoras de trenes de larga distancia en tiempo real. Los datos de la aplicación son actualizados cada minuto aproximadamente y también proveemos un API. Empezamos a hacer esto hace varios meses y hasta ahora hemos acumulado un inmenso conjunto de datos que se agranda a cada hora. A esta altura incluye cientos de miles de filas de datos. El proyecto permite al usuario explorar estos datos en tiempo real, e investigar en el archivo de meses anteriores. Al final la historia que narramos será definida de modo significativo por la acción individual de los usuarios.

En el periodismo tradicional, debido al carácter lineal de los medios escritos o de difusión, tenemos que pensar en un comienzo, el fin, el desarrollo de la historia y el largo y el ángulo de nuestra pieza. Con el periodismo de datos las cosas son diferentes. Sí hay un comienzo. La gente llega al sitio y tiene una primera impresión de la interfaz. Pero a partir de allí se las tienen que arreglar solos. Pueden quedarse un minuto o media hora.

Nuestro trabajo como periodistas de datos es proveer el marco o el medio para esto. Junto con escribir código y manejar datos, tenemos que pensar en maneras ingeniosas de diseñar experiencias. La experiencia del usuario (UX) deriva principalmente de la Interfaz de Usuario (gráfica – GUI). Al final, esta es la parte que definirá el éxito de un proyecto. Se

puede tener el mejor código trabajando en el trasfondo, manejando un conjunto de datos interesante. Pero si la presentación es mala, no le importará a nadie.

Aún hay mucho por aprender y experimentar. Pero por suerte está la industria de los juegos, que ha estado innovando al respecto de las narrativas, los ecosistemas y las interfaces digitales desde hace varias décadas. Por lo que cuando desarrollamos aplicaciones de periodismo de datos, debemos estar atentos a cómo funciona el diseño de juegos y cómo se narran historias en los juegos. ¿Por qué juegos como Tetris son tan divertidos? ¿Y qué es lo que define los mundos abiertos de juegos como Grand Theft Auto o Skyrim rock?

Creemos que el periodismo de datos ha llegado para quedarse. En pocos años, los flujos de trabajo del periodismo de datos estarán incrustados naturalmente en las redacciones porque los sitios de noticias tendrán que cambiar. La cantidad de información disponible al público seguirá creciendo. Pero por suerte nuevas tecnologías seguirán permitiéndonos encontrar nuevas maneras de narrar historias. Algunas de las historias se basarán en datos y muchas aplicaciones y servicios tendrán carácter periodístico. La cuestión interesante es qué estrategia desarrollarán las redacciones para promover este proceso. ¿Crearán equipos de periodistas de datos integrados en sus redacciones? ¿Habrá departamentos de investigación y desarrollo, un poco como los departamentos internos de empresas que se tratan como si fueran independientes? ¿O habrá tercerización de partes del trabajo a compañías especializadas? Estamos recién en el comienzo y el tiempo dirá.

— Lorenz Matzat, OpenDataCity

Kaas & Mulvad: Contenido Semi-Terminado para Grupos con Intereses Específicos.

Los medios de grupos con intereses específicos constituyen un sector emergente, en gran medida ignorado por los teóricos de los medios, que potencialmente podría tener un tremendo impacto a través de redes online o proveyendo contenido a medios de noticias. Pueden definirse como medios (por lo general online), controlados por sectores de organizaciones o instituciones, utilizados para defender ciertos intereses y a ciertas comunidades. Las ONG comúnmente crean tales medios; lo mismo hacen los grupos de consumidores, las asociaciones profesionales, los sindicatos y así en más. La limitación clave de su capacidad de influir en la opinión pública u otras partes interesadas es a menudo que no cuentan con la capacidad para descubrir información importante, con más limitaciones incluso que los medios de noticias que han reducido su capacidad. Kaas & Muvlad, una corporación danesa con fines de lucro, es una de las primeras empresas de

medios de investigación que ofrece capacidad experta a estos medios de grupos con determinados intereses.

La firma se originó en 2007 al separarse del Instituto Danés de Periodismo Asistido por Computadora (Dicar) sin fines de lucro, que vendía informes a medios y capacitaba a periodistas en análisis de datos. Sus fundadores, Tommy Kaas y Nils Mulvad, fueron previamente periodistas en la industria de noticias. Su nueva firma ofrece lo que llaman "datos más visión periodística" (contenido que queda semi terminado, requiriendo edición o reescritura) principalmente a medios con determinados intereses, que utilizan el contenido para informes de prensa o artículos y los distribuyen a través de medios de noticias y sus propios medios (tales como sitios en la red). Entre los clientes directos se incluyen instituciones gubernamentales, firmas de Relaciones Públicas, sindicatos y ONG tales como EU Transparency y World Wildlife Fund. El trabajo para ONG incluye el seguimiento de subsidios agrícolas y de pesca y actualizaciones regulares sobre actividades de lobbistas de la UE generadas a través de "*scraping*" de sitios pertinentes. Entre los clientes indirectos se incluyen fundaciones que financian proyectos de ONG. La firma también trabaja con la industria de noticias; por ejemplo, un diario sensacionalista compró su servicio de seguimiento de celebridades.

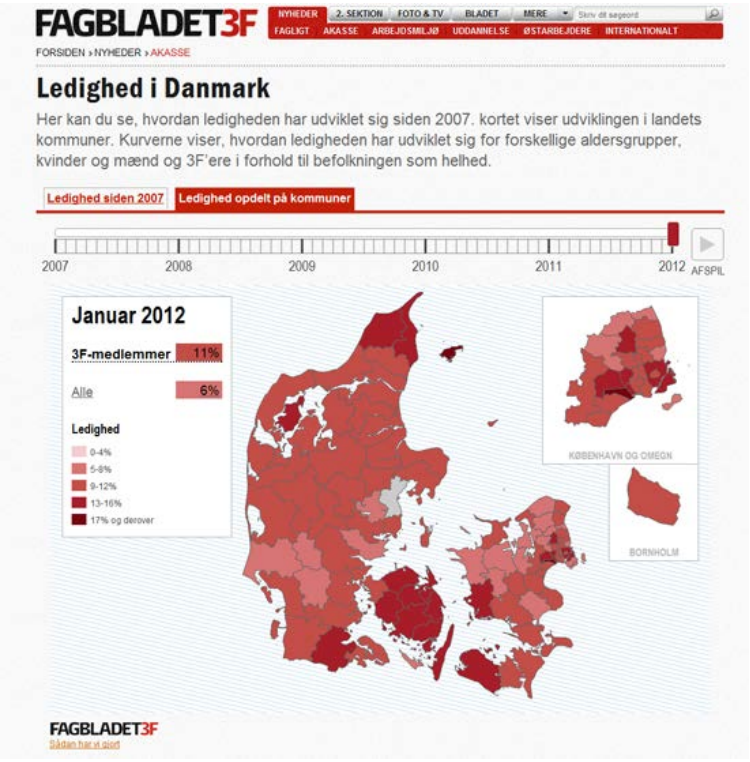


Figure 14. Grupos de interés en medios (Fagblaget3F)

Los proyectos de periodismo de datos en su portfolio incluyen:

Mapa de desempleo para 3F

> Una visualización de datos con cifras claves sobre desempleo en Dinamarca para 3F, que es el sindicato de los trabajadores sin capacitación en Dinamarca.

Condiciones de Vida para 3F

Otro proyecto para 3F muestra como son las distintas condiciones de vida en distintas partes de Dinamarca. El mapa usa 24 indicadores diferentes.

Deuda para “Ugebrevet A4”

Un proyecto que calcula un “índice de deuda” y visualiza las diferencias de economías privadas.

Instalaciones peligrosas en Dinamarca

Un proyecto que hace el mapa de y analiza la proximidad de instalaciones peligrosas con jardines de infantes y otras instituciones realizado por “Born&Unge”, una revista publicada por BUPL, El Sindicato Danés de Educadores de la Primera Infancia y Jóvenes.

Datos de Responsabilidad Corporativa para Vestas

Visualización de datos de 5 áreas de RC para la compañía de turbinas de viento danesa, Vestas, que con texto autogenerado. Actualizado automáticamente quincenalmente con 400 páginas web incluyendo desde datos a escala mundial hasta unidades individuales de producción.

Mapa de Nombres para Experian

Tipee su apellido y mire la distribución de este nombre en distintas áreas geográficas de Dinamarca.

Mapa de Smiley para Ekstra Bladet

Todos los días Kaas & Mulvad extraen todas las inspecciones por alimentos en mal estado y hacen el mapa de lo más actual para el diario danés Ekstra Bladet (a la mitad del sitio está el mapa).

Kaas & Mulvad no son los primeros periodistas en trabajar con grupos de determinados intereses en medios. Greenpeace, por ejemplo, habitualmente recurre a periodistas como colaboradores para sus informes. Pero no conocemos ninguna otra firma cuyas ofertas a medios de este tipo se basen en datos; es mucho más común que los periodistas trabajen con ONG como periodistas, editores o escritores.

Actualmente los medios informativos con ayuda de computadoras se concentran en la búsqueda y el descubrimiento (por ejemplo, WikiLeaks); en esto también Kaas & Mulvad son innovadores al concentrarse en análisis de datos. Su enfoque requiere no solo capacidades de programación sino también comprensión de qué tipo de información puede producir una historia con impacto. Se puede decir con certeza que cualquiera que desee imitar su servicio probablemente tendría que adquirir esos dos conjuntos de capacidades a través de asociaciones, porque raramente los individuos poseen ambos.

Procesos: TI innovadora más análisis

La firma encara alrededor de 100 proyectos al año, que duran entre pocas horas y pocos meses. También invierte continuamente en proyectos que expanden su capacidad y ofertas. El servicio de seguimiento de celebridades fue uno de tales experimentos. Otro involucró recorrer Internet en busca de noticias sobre ejecución de hipotecas y crear mapas de los casos. Los socios dicen que su primer criterio para encarar proyectos es disfrutar del trabajo y aprender del mismo; se buscan mercados luego de que se defina un nuevo servicio. Aseguran en que el sector de noticias les resultó difícil desarrollar nuevos métodos y nuevos negocios.

No tenemos editores o jefes que decidan qué proyectos podemos hacer, qué software o hardware podemos comprar. Podemos comprar las herramientas de acuerdo a las necesidades de los proyectos, como las mejores soluciones para búsqueda y extracción de texto. Nuestra meta es estar a la vanguardia en estas áreas. Tratamos de conseguir clientes que estén dispuestos a pagar o, si el proyecto es divertido, lo hacemos por menos dinero.

Valor creado: marcas e ingresos personales y de firmas

El giro en 2009 fue aproximadamente de 2.500.000 de coronas danesas (€ 336.000). La compañía también sostiene la reputación de los socios como periodistas de vanguardia, lo que mantiene la demanda para sus servicios educativos y de conferencias. Sus apariciones públicas al mismo tiempo sostienen la marca de la firma.

Percepciones claves de este ejemplo

- La crisis de capacidad en baja del sector informativo es también una crisis de sub utilización de capacidad. Kaas y Mulvad tuvieron que dejar el sector para hacer trabajo que valoran y eso da resultados. Nada impide a las organizaciones de noticias captar ese valor.
- Al menos en algunos sectores, existe un mercado rentable para “contenido semi-acabado” que puede servir a los grupos de interesados.
- Sin embargo, esta oportunidad plantea la cuestión de cuánto control pueden ejercer los periodistas sobre la presentación y uso de su trabajo por terceros. Recordamos que esta cuestión ya existe dentro del sector de las noticias (donde los editores pueden imponer cambios al producto de un periodista) y ha existido en otros sectores de medios (tales como el cine, donde no son precisamente raros los conflictos ente directores y estudios por el “corte final”). No es un riesgo moral particular de los medios de interesados, y tampoco va a desaparecer. Se necesita prestar más atención a la ética de esta realidad y mercado en crecimiento.
- Desde el punto de vista de los ingresos, un producto o servicio solo no basta. Las compañías exitosas que realizan periodismo de denuncia debieran tener un enfoque de cartera en el que la consultoría, la enseñanza, las conferencias y otros servicios aportan ingresos extra y sostienen la marca.

— Extracto editado de ``Disruptive News Technologies: Stakeholder Media and The Future of Watchdog Journalism Business Models "de Mark Lee Hunter y Luk N. Van Wassenhove, INSEAD Working Paper, 2010

Modelos de negocios para periodismo de datos

En medio de todo el interés y las expectativas respecto del periodismo basado en datos, hay una cuestión sobre la que siempre hay curiosidad en las redacciones: ¿cómo son los modelos de negocios?

Si bien debemos ser cuidadosos respecto de hacer predicciones, un análisis de la historia reciente y el estado actual del sector de los medios nos puede dar una visión. Hoy hay muchas organizaciones de noticias que se han beneficiado al adoptar nuevos enfoques.

Los términos como “periodismo de datos” y la nueva expresión de moda, “ciencia de datos”, pueden sonar como que describen algo nuevo, pero no es estrictamente cierto. En cambio estas nuevas etiquetas son solo maneras de caracterizar un cambio que ha estado cobrando fuerza a lo largo de décadas.

Muchos periodistas parecen inconscientes de la magnitud de los ingresos que ya se generan a través de la recolección de datos, su análisis y visualización. Este es el negocio de la refinación de la información. Con herramientas y tecnologías para procesar datos, es cada vez más posible echar luz sobre asuntos muy complejos, se trate de finanzas internacionales, deuda, demografía, educación, y así en más. El término “inteligencia de negocios” describe una variedad de conceptos de TI que apuntan a aportar una visión clara de lo que sucede en corporaciones comerciales. Las compañías grandes y rentables de nuestro tiempo, incluyendo McDonalds, Zara y H&M, dependen del seguimiento constante de datos para obtener ganancias. Y para ellos funciona bastante bien.

Lo que está cambiando es que las herramientas desarrolladas para este espacio ahora están disponibles para otros dominios, incluyendo los medios. Y hay periodistas que lo entienden. Está por caso Tableau, una compañía que provee un conjunto de herramientas de visualización. O el movimiento “Big Data” (Grandes Datos), en el que compañías de tecnología usan paquetes de software (a menudo de código abierto) para analizar pilas de datos, extrayendo conclusiones en milisegundos.

Estas tecnologías ahora se pueden aplicar al periodismo. Equipos de The Guardian y The New York Times están constantemente ampliando los límites de este campo naciente. Y lo que vemos actualmente es solo la punta del iceberg.

¿Pero cómo genera esto dinero para periodismo? El gran mercado mundial que actualmente se está abriendo tiene que ver con la transformación de datos de disponibilidad pública en

algo que podamos procesar: haciendo que los datos resulten visibles y humanos. Queremos poder relacionarnos con las grandes cifras que escuchamos todos los días en las noticias, lo que significan los millones y miles de millones para cada uno de nosotros.

Hay una cantidad de compañías de medios basadas en datos, muy rentables, que simplemente han aplicado este principio antes que otras. Disfrutan de tasas de crecimiento saludables y a veces ganancias que impresionan. Un ejemplo es Bloomberg. La compañía opera alrededor de 300.000 terminales y entrega datos financieros a sus usuarios. Si usted está en el negocio del dinero, esta es una herramienta poderosa. Cada terminal viene con un teclado con códigos de colores y hasta 30.000 opciones para mirar, comparar, analizar y ayudarlo a decidir que hacer a continuación. Este negocio central genera según se estima US$ 6300 millones al año, al menos según un artículo publicado en 2008 en The New York Times. Como resultado de ello Bloomberg ha estado contratando periodistas por todas partes, compraron la venerable pero perdidosa “Business Week”, y así siguiendo.

Otro ejemplo es el conglomerado de medios canadiense conocido hoy como Thomson Reuters. Comenzaron con un diario, compraron una cantidad de títulos conocidos en el Reino Unido y luego decidieron hace dos décadas dejar el negocio de los diarios. En vez de ello, han crecido en base a servicios de información, apuntando a proveer una perspectiva más profunda para clientes en una cantidad de sectores. Si le preocupa cómo ganar dinero con información especializada, mi consejo sería que simplemente lea la historia de la compañía en Wikipedia.

Y vea The Economist. La revista ha creado una marca excelente e influyente por el lado de los medios. Al mismo tiempo la “Economist Intelligence Unit” ahora es más como una consultora, informando sobre tendencias y pronósticos relevantes para casi todos los países del mundo. Emplean cientos de periodistas y sostienen que sirven a 1.500.000 de clientes en todo el mundo.

Y hay muchos servicios de nicho basados en datos que podrían servir como inspiración: eMarketer en Estados Unidos, que ofrece comparaciones, cuadros y consejos para cualquiera interesado en marketing en Internet; Stiftung Warentest en Alemania, institución que analiza la calidad de productos y servicios; Statista, también de Alemania, una nueva empresa que ayuda a visualizar información públicamente disponible.

En todo el mundo actualmente hay una oleada de nuevas empresas en este sector, que

cubren naturalmente una amplia gama de áreas; por ejemplo, Timetric, que apunta a “reinventar los estudios de negocios”, OpenCorporates, Kasabi, Infochimps y Data Market. Muchas de estas son experimentos, pero de conjunto pueden considerarse una señal importante de cambio.

Y están los medios públicos, que en términos de periodismo de datos, son un gigante dormido. En Alemania, € 7200 millones van a este sector anualmente. El periodismo es un producto especial: si se hace bien, no solo se trata de ganar dinero, sino que sirve un rol importante en la sociedad. Una vez que queda en claro que el periodismo de datos puede ofrecer visiones más confiables y de modo más fácil, parte de este dinero podría usarse para nuevos empleos en las redacciones.

En el caso del periodismo de datos no se trata solo de ser el primero si no de ser una fuente de información confiable. En este mundo multicanal, se puede generar atención en abundancia, pero la *confianza* es un recurso cada vez más escaso. Los periodistas de datos pueden ayudar a filtrar, sintetizar y presentar fuentes de información diversas y a menudo difíciles de un modo que le da al público una visión real de asuntos complejos. En vez de solo reciclar comunicados de prensa y repetir las historias que han escuchado en otras partes, los periodistas de datos pueden dar a los lectores una perspectiva clara, comprensible y preferentemente adecuada a esos lectores, con gráficos interactivos y acceso directo a fuentes primarias. No trivial y sin duda valioso.

¿Entonces cuál es el mejor enfoque para que quienes aspiran a periodistas de datos exploren este campo y convenzan a la gerencia de que apoyen proyectos innovadores?

El primer paso debiera ser buscar oportunidades inmediatas cerca de donde están: fruta que cuelga del árbol. Por ejemplo usted puede tener ya colecciones de textos y datos estructurados que puede usar. Un gran ejemplo de esto es la “base de datos de homicidios” de Los Ángeles Times. Aquí los datos y las visualizaciones son el centro, no algo secundario. Los editores recogen información sobre todos los crímenes que encuentran y recién entonces escriben artículos basados en ello. Con el tiempo tales colecciones se están volviendo mejores, más profundas y más valiosas.

Esto podría no funcionar la primera vez. Pero con el tiempo si lo hará. Un indicador que da muchas esperanzas es que el Texas Tribune y ProPublica, que podría decirse que son ambas compañías de medios de la era posterior a los diarios impresos, informaron que la financiación de sus organizaciones de periodismo sin fines de lucro superó sus metas mucho antes de lo planificado.

Volverse eficiente en todo lo relacionado con datos –como generalista o como especialista concentrado en un aspecto de la cadena alimenticia de datos- genera una perspectiva valiosa para la gente que cree en el periodismo. Como dijo un muy conocido editor en Alemania recientemente en una entrevista: “Existe este nuevo grupo que se llaman periodista de datos. Y ya no están dispuestos a trabajar por moneditas”.

— Mirko Lorenz, Deutsche Welle

Estudio de casos

En esta sección analizamos con más profundidad el detrás de escena de numerosos proyectos de periodismo de datos, desde aplicaciones desarrolladas en un día, hasta investigaciones de 9 meses de duración. Nos informamos sobre cómo han sido usadas fuentes de datos para aumentar y mejorar la cobertura de diferentes temas, desde elecciones hasta gasto, de disturbios hasta corrupción, desde el nivel educativo de las escuelas hasta el precio del agua. Junto a organizaciones de grandes medios, tales como la BBC, el Chicago Tribune, The Guardian, el Financial Times, Helsingin Sanomar, La Nación, el Wall Street Journal, y el Zeit Online, también presentamos iniciativas más pequeñas tales como las de California Watch, Hack/HackersBeunos Aires, ProPublica y un grupo de ciudadanos-periodistas brasileños llamados amigos de Januária.

Qué contiene este capítulo?

La brecha de oportunidades

The Opportunity Gap (La Brecha de Oportunidades, usó datos de derechos civiles nunca antes difundidos del departamento de Educación de Estados Unidos y mostró que algunos estados, como Florida, han creado una situación equitativa ofreciendo a estudiantes ricos y pobres un acceso equitativo en términos generales a cursos de alto nivel, mientras que otros estados, como Kansas, Maryland y Oklahoma, ofrecen menos oportunidades en distritos con familias más pobres.

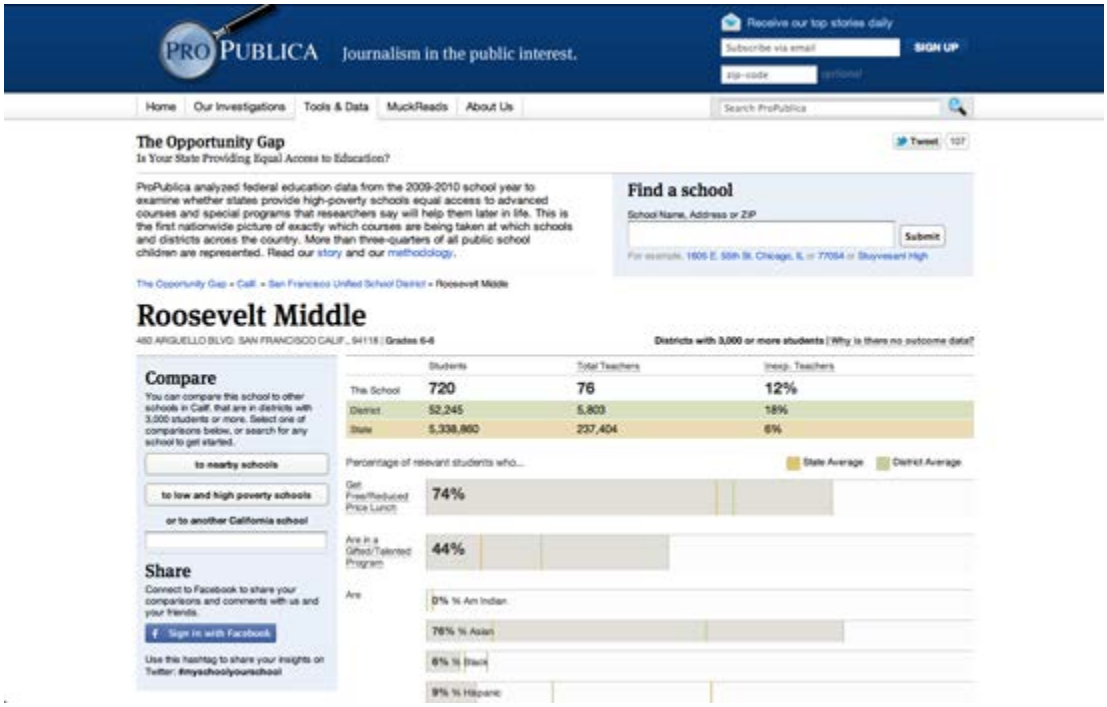


Figure 1. The Opportunity Gap project (ProPublica)

Los datos incluyen a las escuelas públicas de todo distrito con 3000 estudiantes o más. Estan representados más de 3 cuartos de todos los alumnos de escuelas públicas. Un reportero de nuestra redacción obtuvo los datos y nuestro Director de Informes Asistidos por Computadora los depuró en profundidad.

Fue un proyecto que llevó aproximadamente 3 meses. En total 6 personas trabajaron en la historia y la aplicación de noticias: 2 editores, un redactor, una persona de Informes asistidos por computadora y 2 programadores. La mayoría no trabajó exclusivamente en este proyecto durante ese período.

El proyecto requirió realmente nuestras capacidades combinadas: profundo conocimiento del tema, una comprensión de las mejores prácticas con datos, capacidades de diseño y programación. Lo que es más importante, requirió la capacidad de encontrar la historia en los datos. También exigió edición, no solo para la historia que la acompaña, sino también para la aplicación de noticias.

Para la depuración y análisis de los datos usamos principalmente Excel y rutinas de depuración, así como MS Access. La aplicación de noticias fue desarrollada con el programa Ruby on Rails y usa abundantemente JavaScript.

Además de un artículo que da el marco general, nuestra cobertura incluyó una aplicación de noticias interactiva que permite a los lectores comprender y encontrar ejemplos que se relacionen con su propia situación dentro de esta gran base de datos nacional. Utilizando nuestra aplicación de noticias, el lector podía encontrar su escuela local –digamos, por ejemplo, Central High School en Newark, N.J._ y ver inmediatamente el desempeño relativo de la escuela en una gran variedad de áreas. Entonces podía cliquear un botón que dice "comparar con Escuelas de Alta y Baja Pobreza", e inmediatamente ver otras escuelas medias y su pobreza relativa, y la medida en la que ofrecen matemática avanzada, Advanced Placement (conocido con la sigla AP, un programa de la dirección de Colleges, que ofrece currícula y exámenes de nivel de College para estudiantes de secundaria en Estados Unidos, n. del t.) y otros cursos importantes. En nuestro ejemplo, Central High tiene como referencia a Millburn Sr. High. La Brecha de Oportunidades muestra que sólo el 1% de los estudiantes de Milburn recibe almuerzo gratis o a precio reducido y el 72% de ellos hace al menos un curso de AP. En el otro extremo, en el International High el 85% de sus estudiantes recibe almuerzo gratis o a precios reducidos y solo 1% toma cursos AP.

A través de este ejemplo el lector puede usar información que conoce –de una escuela media local- para averiguar algo que no sabe: la distribución de la accesibilidad educativa y en qué medida la pobreza predice esa accesibilidad.

También integramos la aplicación con Facebook, de modo que los lectores pudieran loguearse esta a esta red social y nuestra aplicación automáticamente les haría saber de escuelas que podrían interesarles.

El tráfico hacia todas nuestras aplicaciones de noticias es excelente y estamos particularmente orgullosos del modo en que ella cuenta una historia compleja; y, lo que va más al grano, ayuda a los lectores a definir su propia historia.

Tal como sucede con muchos proyectos que parten de información oficial, los datos requirieron mucha depuración. Por ejemplo, si bien sólo hay alrededor de 30 posibles cursos de Advanced Placement, algunas escuelas informaban que tenían cientos de ellos. Esto exigió muchos chequeos manuales y llamadas telefónicas a escuelas para confirmaciones y correcciones.

También trabajamos fuerte para asegurarnos que la aplicación ofreciera una versión "lejana" y una versión "cercana" de la historia. Es decir, la aplicación tenía que presentar al lector un cuadro nacional amplio y abstracto; una manera de comparar a los estados en materia de acceso educativo. Pero dado que la abstracción a veces genera confusión en los

lectores respecto de lo que los datos significan para ellos, también queríamos que los lectores pudieran encontrar sus escuelas locales y compararlas con escuelas de baja pobreza en su área.

Si quisiera aconsejar a quienes quieren ser periodistas de datos y abordar proyectos de este tipo, diría que tienen que conocer el material y ser inquisitivos. Todas las reglas que valen para otros tipos de periodismo, valen también aquí. Hay que tener datos ciertos, asegurarse de contar bien la historia y -cuestión crucial- asegurarse de que su aplicación de noticias no contradiga la historia que está escribiendo. Si lo hace, una de las 2 podría estar equivocada.

Además, si usted quiere aprender a programar, lo más importante es empezar. Usted puede preferir aprender a través de clases, libros o videos, pero asegúrese de tener una idea realmente buena para un proyecto y un plazo para completarlo. Si tiene una historia en la cabeza que solo puede expresarse a través de una aplicación de noticias, entonces no saber programar no lo va a detener.

— Scott Klein, ProPublica

Una investigación de 9 meses sobre Fondos Estructurales Europeos

En 2010, el Financial Times y el Bureau of Investigative Journalism (BIJ) se unieron para investigar los Fondos Estructurales Europeos. La intención era revisar quiénes son los beneficiarios de esos fondos y verificar si el dinero se usó para bien. Con € 347.000 millones a lo largo de 7 años, los Fondos Estructurales son el segundo programa de subsidios de la Unión Europea. El programa existe desde hace décadas, pero fuera de informes generales, había poca transparencia respecto de quiénes eran los beneficiarios. Como parte de un cambio de reglas en la actual ronda de otorgamiento de fondos, las autoridades están obligadas a hacer pública una lista de beneficiarios, incluyendo la descripción de los proyectos y el monto de fondos de la UE y nacionales recibidos.

Figure 2. Investigación de Fondos Estructurales de la UE (Financial Times y el Bureau of Investigative Journalism)

El equipo del proyecto estaba compuesto por 12 periodistas y un programador tiempo completo colaborando por 9 meses. La recolección de los datos por sí sola llevó varios meses.

El proyecto se publicó en 5 días de cobertura en el Financial Times y el BIJ, un documental radial de la BBC y varios documentales de TV.

Antes de abordar un proyecto con este nivel de esfuerzo hay que estar seguro de que lo descubierto es original y que se terminará teniendo buenas historias que nadie más tiene.

El proceso se dividió en una serie de pasos diferentes.

1. Identificar quién registra los datos y cómo

El Directorio General de las Regiones de la Comisión Europea tiene un portal de los sitios de autoridades regionales que publican los datos. Creíamos que la Comisión tendría una base de datos general de proyectos a la que podríamos acceder directamente o que podríamos obtener a través de un pedido de acceso a la información. No existe tal base de datos con el nivel de definición requerido. Rápidamente advertimos que muchos de los vínculos provistos por la comisión eran erróneas y que la mayoría de las autoridades publicaban los datos en formato PDF, en vez de formatos que faciliten el análisis tales como CSV o XML.

Un equipo de 12 personas trabajó para identificar los datos más actualizadas y ordenar los vínculos reuniéndolos en una planilla de cálculo que usamos para colaboración. Dado que los campos de datos no eran uniformes (por ejemplo, los encabezados estaban en distintos idiomas, algunos conjuntos de datos usaban diferentes divisas, y algunos incluían descomposición en fondos de UE y nacionales) tuvimos que ser lo más precisos posible en la traducción y [line-through]*la*descripción de los campos de datos disponibles en cada conjunto.

2. Descargar y preparar los datos

El siguiente paso consistió en descargar todas las planillas de cálculo, PDF y, en algunos casos, recopilar datos originales en la red.

Cada conjunto de datos tuvo que ser estandarizado. Nuestra mayor tarea fue extraer datos de cientos de páginas en formato .PDF. Gran parte de esto se hizo utilizando UnPDF y ABBYY FineReader, que permiten extraer datos a formatos tales como CSV o Excel.

También significó verificar y volver a verificar que las herramientas de extracción de PDF hubiesen captado los datos correctamente. Esto se hizo filtrando, ordenando y sumando

totales (para asegurarnos que se correspondieran con lo publicado en los PDF).

3. Crear una base de datos

El programador del equipo creó una base de datos SQL. Cada uno de los archivos preparados fue utilizado entonces como unidad para la construcción de la base de datos SQL general. Con un proceso diario se cargaba todos los archivos individuales de datos en una gran base de datos SQL, en la que se podían realizar búsquedas en cualquier momento a través de su interfaz con palabras claves.

4. Doble verificación y análisis

El equipo analizó los datos de 2 maneras principales:

Vía la interfaz de la base de datos

> Esto significó tipear palabras claves de interés (por ejemplo, “tabaco”, “hotel”, “compañía A” en el motor de búsquedas. Con la ayuda de Google Translate, que fue incorporado a la funcionalidad de búsquedas de nuestra base de datos, esas palabras claves se traducían a 21 idiomas, obteniendo los resultados apropiados. Estos se podían descargar y los periodistas podían continuar su investigación en proyectos individuales de su interés.

Por macro-análisis usando toda la base de datos

> Ocasionalmente descargábamos un conjunto de datos completo, que entonces podía ser analizado (por ejemplo, usando palabras clave o agregando datos por país, región, tipo de gasto, número de proyectos por beneficiarios, etc.)

Nuestras historias se conformaron con ambos métodos, pero también a través de investigación de campo y de escritorio.

Hacer la doble verificación de la integridad de los datos (agregando y verificando en comparación con lo que las autoridades dijeron que fue asignado) llevó una gran cantidad de tiempo. Uno de los principales problemas fue que las autoridades mayormente solo divulgaban la cantidad de “fondos de la UE y nacionales”. Bajo las reglas de la UE, cada programa puede cubrir un porcentaje del costo total usando fondos de la UE. El nivel de financiación por la UE es determinado, al nivel del programa, por la llamada tasa de co-financiación. Cada programa (por ejemplo, competitividad regional) está compuesto de numerosos proyectos. Al nivel de los proyectos, técnicamente, uno podría recibir ciento por ciento de financiación de la UE y otro nada, mientras el monto total de la financiación por la UE al nivel de los programas no superara la tasa de co-financiación aprobada.

Esto significó que tuvimos que verificar con cada compañía beneficiaria el monto de

financiación de la UE que citamos en nuestras historias.

— *Cynthia O'Murchu, Financial Times*

El colapso de la Eurozona

Estamos cubriendo el colapso de la Eurozona. Todos los aspectos. El dramatismo de los enfrentamientos entre gobiernos y la pérdida de los ahorros de toda la vida; la reacción de los líderes mundiales, las medidas de austeridad, y las protestas en contra de estas medidas. Todos los días en el Wall Street Journal hay cuadros sobre pérdidas de empleos, caída de PBI y hundimiento de los mercados mundiales. Es incremental. Y aturde.

Los editores de tapa convocan una reunión para debatir ideas sobre la cobertura de fin de año y en momentos en que me voy de la reunión, me pregunto: ¿Cómo será vivir esto?

¿Es esto como 2008 cuando me echaron y las malas noticias eran incesantes? Hablábamos de empleo y dinero todas las noche en la cena, casi sin pensar en cómo podía intranquilizar a mi hija. Y los fines de semana eran lo peor. Yo trataba de negar el temor que parecía dominarme permanentemente y la ansiedad que no me dejaba respirar. ¿Así vive una familia ahora mismo en Grecia? ¿En España?

Me volví y seguí a Mike Allen, el editor de tapa, a su oficina le propuse la idea de contar la crisis a través de familias en la Eurozona mirando primero los datos, encontrando perfiles demográficos para entender la composición familiar y luego sacando eso a luz junto con las imágenes y entrevistas, audio de las distintas generaciones. Usaríamos hermosos elementos de retrato, las voces ... y los datos.

Cuando volví a mi escritorio escribí un resumen y dibujé un logo.

Figure 3. El colapso de la Eurozona: resumen (Wall Street Journal)

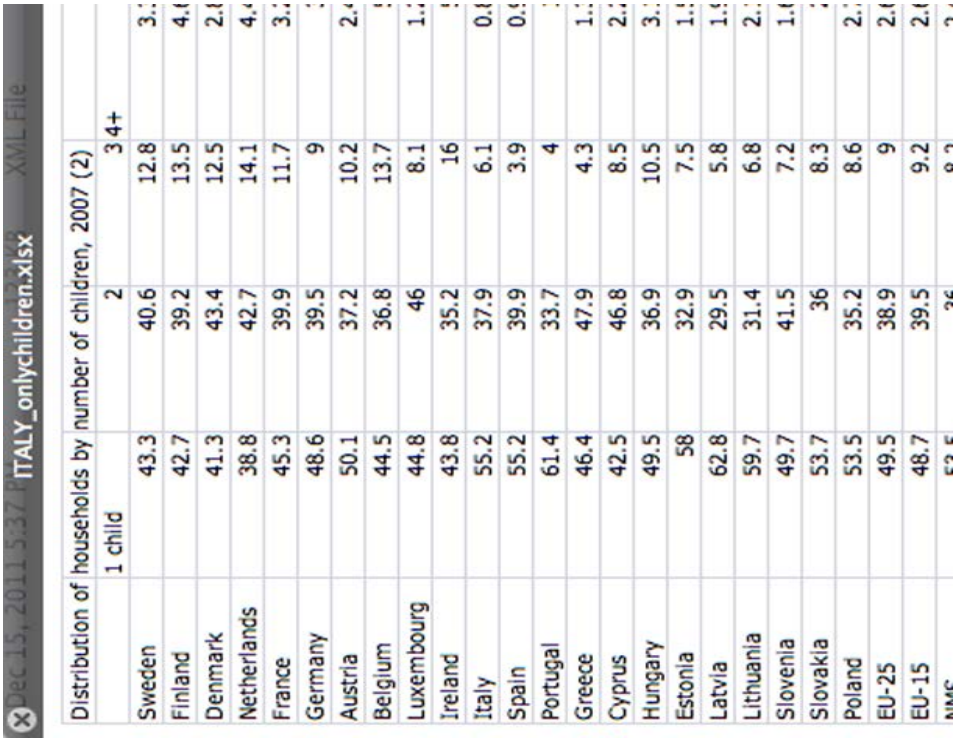
Durante las siguientes 3 semanas perseguí cifras: métricas sobre matrimonio, mortalidad, el tamaño de las familias y gasto en salud. Leí sobre condiciones de vida y tasas de divorcio, miré encuestas sobre bienestar y tasas de ahorro. Estudié estadísticas nacionales, llamé al bureau de población de la ONU, el FMI, Eurostat, y la OCDE hasta que encontré un economista que había pasado su carrera siguiendo familias. Me conectó con una estudiosa sobre composición familiar. Me indicó trabajos sobre mi tema.

Con mi editor, Sam Enriquez, redujimos el número de países. Reunimos un equipo para debatir el enfoque visual y qué periodistas producirían palabras, audio y la historia. Matt Craig, el editor fotográfico de tapa, se puso a trabajar para encontrar fotógrafos. Matt Murray, el subeditor ejecutivo para cobertura mundial, envió un memo a los jefes de sección pidiendo ayuda de los periodistas. (Esto fue crucial: la orden de la máxima jerarquía).

Pero primero los datos. Por la mañana yo exportaba datos a planillas de cálculo y hacía cuadros para ver tendencias: caída del ahorro, desaparición de pensiones, la vuelta de madres al trabajo, gasto en salud, junto con deuda pública y desempleo. Por la tarde analizaba esos datos agrupados, comparando los países para encontrar historias.

Lo hice durante una semana antes de enredarme en los yuyos y comenzar a dudar de mi misma. Quizás fuera un enfoque equivocado. Quizás no debía tratarse de países, sino de padres y madres, y niños y abuelos. Los datos aumentaron.

Y se redujeron. A veces pasaba horas reuniendo información que en definitiva no me decía nada. Había buscado un conjunto de cifras equivocado. En algunos casos los datos eran simplemente demasiado viejos.


Dec 15, 2011 5:37 PM ITALY_onlychildren.xlsx XML File

Distribution of households by number of children, 2007 (2)				
	1 child	2	3	4+
Sweden	43.3	40.6	12.8	3.[illegible]
Finland	42.7	39.2	13.5	4.[illegible]
Denmark	41.3	43.4	12.5	2.[illegible]
Netherlands	38.8	42.7	14.1	4.[illegible]
France	45.3	39.9	11.7	3.[illegible]
Germany	48.6	39.5	9	[illegible]
Austria	50.1	37.2	10.2	2.[illegible]
Belgium	44.5	36.8	13.7	[illegible]
Luxembourg	44.8	46	8.1	1.[illegible]
Ireland	43.8	35.2	16	[illegible]
Italy	55.2	37.9	6.1	0.[illegible]
Spain	55.2	39.9	3.9	0.[illegible]
Portugal	61.4	33.7	4	[illegible]
Greece	46.4	47.9	4.3	1.[illegible]
Cyprus	42.5	46.8	8.5	2.[illegible]
Hungary	49.5	36.9	10.5	3.[illegible]
Estonia	58	32.9	7.5	1.[illegible]
Latvia	62.8	29.5	5.8	1.[illegible]
Lithuania	59.7	31.4	6.8	2.[illegible]
Slovenia	49.7	41.5	7.2	1.[illegible]
Slovakia	53.7	36	8.3	[illegible]
Poland	53.5	35.2	8.6	2.[illegible]
EU-25	49.5	38.9	9	2.[illegible]
EU-15	48.7	39.5	9.2	2.[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Figure 4. Juzgar la utilidad de un conjunto de datos puede ser una tarea que lleve mucho tiempo Sarah Slobin)

Luego los datos volvieron a aumentar al advertir que aún tenía interrogantes y no entendía las familias.

Necesitaba verlo, moldearlo. Por lo que hice una serie rápida de gráficos en Illustrator y comencé a ordenarlos y editarlos.

Al emerger los cuadros, también apareció una imagen cohesionada de las familias.

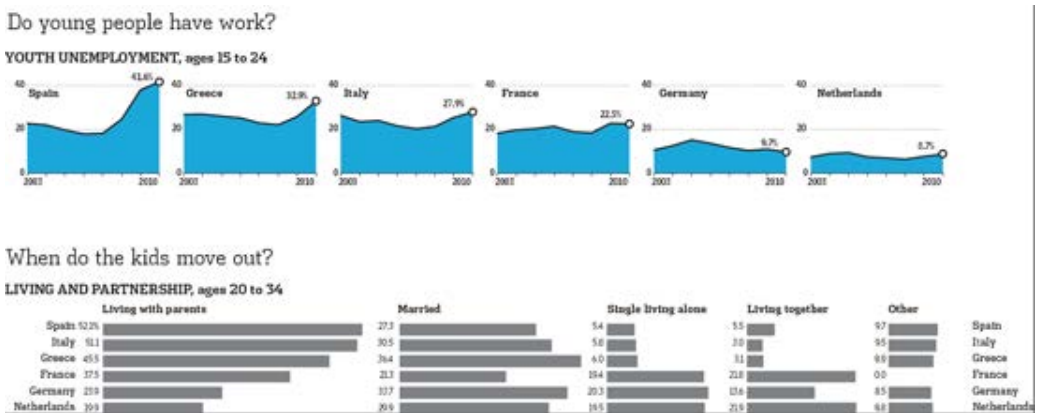


Figure 5. Visualizaciones gráficas: encontrar sentido a tendencias y patrones escondidos en los conjuntos de datos (Sarah Slobin)

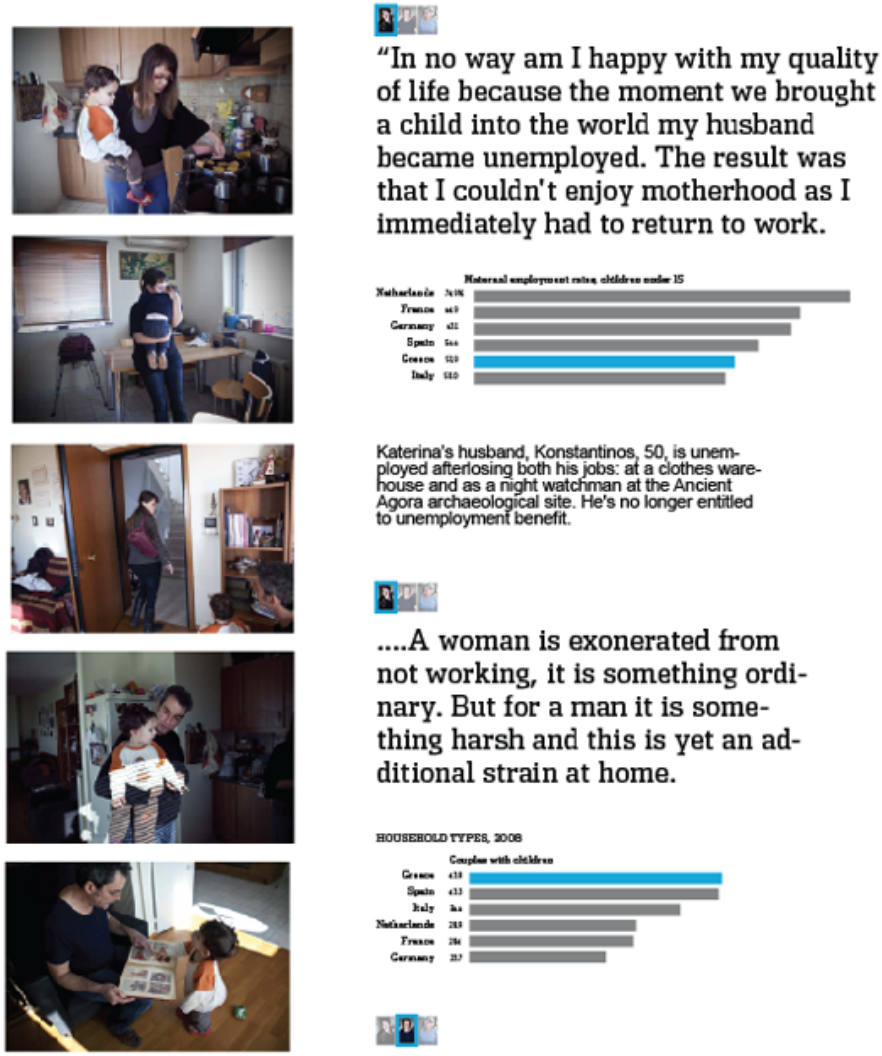


Figure 6. Las cifras son gente: el valor de los datos está en las historias individuales que representan (Wall Street Journal)

Lanzamos el proyecto. Llamé a cada periodista. Les mandé los cuadros, la idea general y una invitación abierta a encontrar historias que ellos consideraran significativas, que acercaran la crisis a nuestros lectores. Necesitábamos una familia pequeña en Ámsterdam y familias más grandes en España e Italia. Queríamos saber de múltiples generaciones para ver cómo la historia personal moldea las respuestas.

A partir de aquí, me levantaba temprano para ver mi correo electrónico teniendo en cuenta la brecha de horarios. Los periodistas respondieron con temas hermosos, síntesis y sorpresas que no había previsto.

En cuanto a fotografías, sabíamos que queríamos retratos de generaciones. La visión de Matt era lograr que sus fotógrafos siguieran a cada miembro de la familia a lo largo de un día en sus vidas. Escogió periodistas visuales que hubiesen cubierto el mundo, cubierto noticias e incluso guerras. Matt quería que cada sesión terminara en la cena. Sam sugirió que incluyéramos los menúes de las comidas.

A partir de allí era cuestión de esperar a ver qué historia contaban las fotos. Esperar a ver qué decían las familias. Diseñamos el aspecto del material interactivo. Robé una paleta de colores de una novela de Tintin y trabajamos la interacción. Y cuando reunimos todo en paneles, agregamos nuevamente algunos (no todos, algunos) de los cuadros originales. Lo suficiente para puntuar cada historia, lo suficiente para endurecer los temas. Los datos se convirtieron en una pausa en la historia, una manera de bajar un cambio.

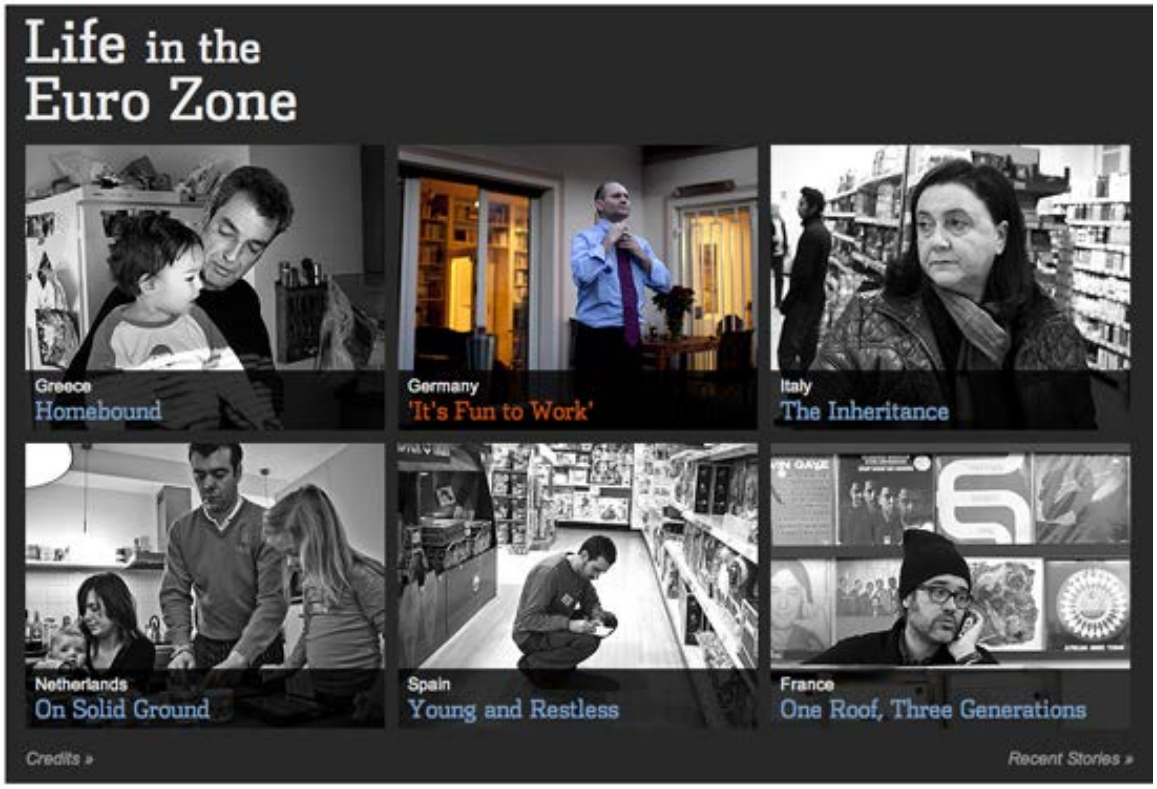


Figure 7. La vida en la Eurozona (Wall Street Journal)

Al final, los datos eran la gente; eran las fotografías y las historias. Era lo que enmarcaba cada narración y provocaba la tensión entre países.

Para cuando publicamos el proyecto, justo antes de fin de año, mientras todos contemplábamos lo que había en el horizonte, ya conocía a todos los miembros de las familias por su nombre. Me sigo preguntando cómo estarán ahora. Y si esto no parece un proyecto de datos, no hay problema. Porque los momentos que quedaron documentados en la *Vida en la zona del Euro*, esas historias de sentarse a comer y hablar sobre el trabajo y la vida con su familia es algo que pudimos compartir con nuestros lectores. Entender los datos es lo que lo hizo posible.

— *Sarah Slobin, Wall Street Journal*

Cubrir el gasto público con OpenSpending.org

En 2007, Jonathan vino a la Open Knowledge Foundation con una propuesta de una carilla para un proyecto llamado Where Does My Money Go (A dónde va mi dinero, que apuntaba a facilitarle a los ciudadanos británicos la comprensión de cómo se gastan los fondos públicos. La intención era que fuera una demostración de un concepto para un proyecto mayor que representara visualmente la información pública, basándonos en trabajos pioneros del Istoype Institute de Otto y Marie Neurath de la década del '40.

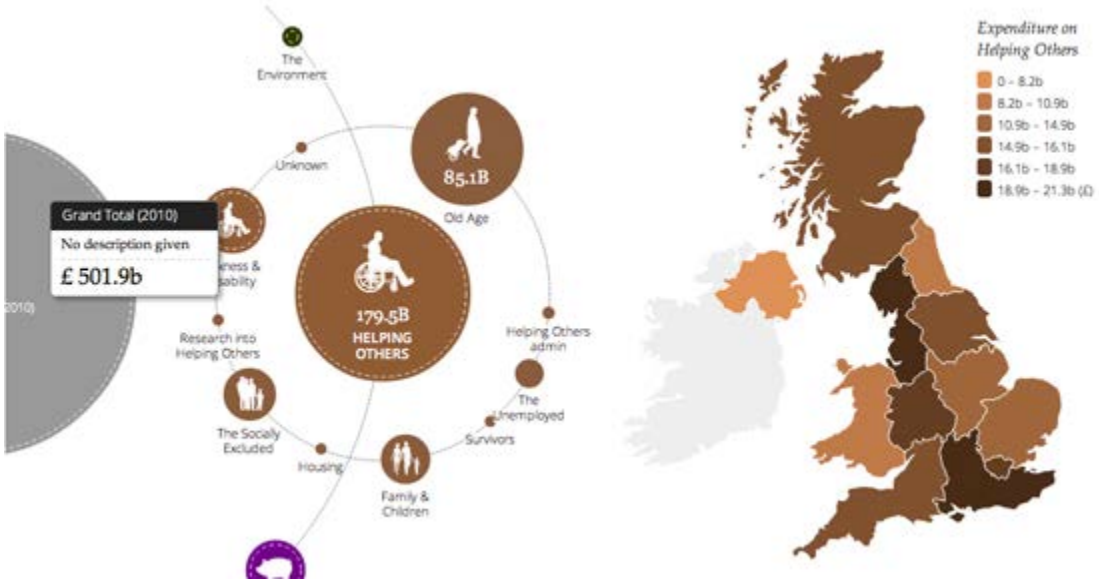


Figure 8. ¿A dónde va mi dinero? (Open Knowledge Foundation)

El proyecto Where Does My Money Go? permitió a los usuarios explorar datos públicos de una amplia variedad de fuentes usando herramientas de código abierto intuitivas. Obtuvimos apoyo para desarrollar un prototipo del proyecto, y luego recibimos fondos del 4IP de Channel 4, para convertir esto en una aplicación de la red plenamente funcional.

El gurú del diseño informático, David McCandless (de Information is Beautiful; creó varias vistas distintas de los datos que ayudan a la gente a ubicarse respecto de las grandes cifras, incluyendo el "Analisis del País y Regional", que muestra cómo se gastan los fondos en distintas partes del país, y "Daily Bread" (Pan diario, que muestra a los ciudadanos un desglose de sus contribuciones fiscales por día en libras y centavos.

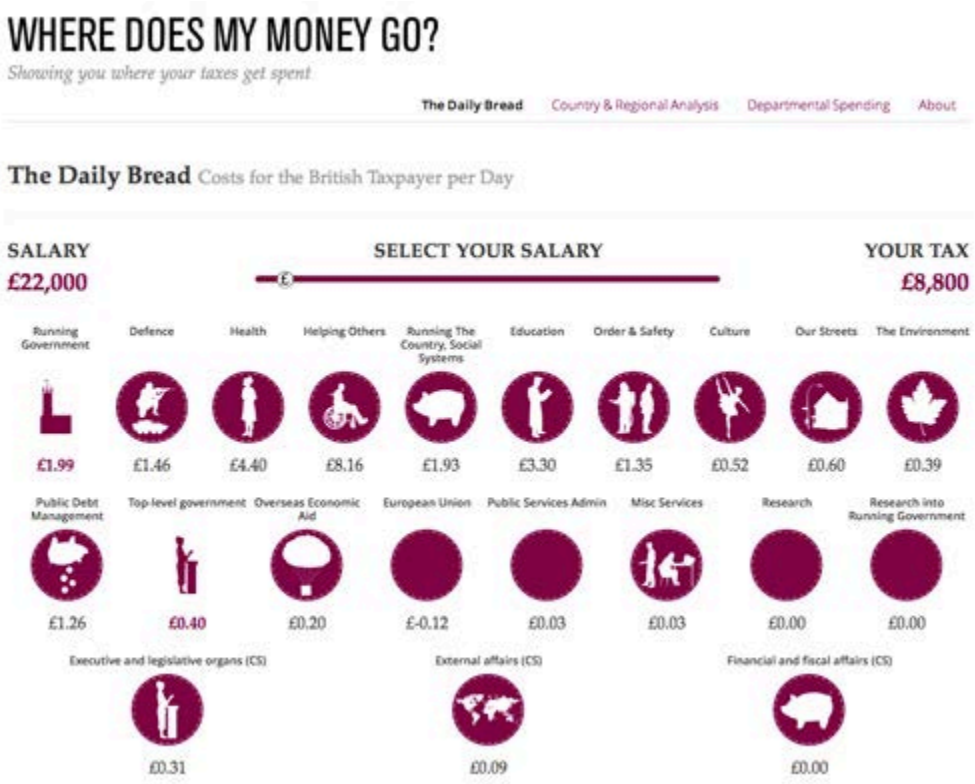


Figure 9. Calculador impositivo Daily Bread de ¿A dónde va mi dinero? (Open Knowledge Foundation)

En aquel tiempo, el santo grial para el proyecto eran los datos de lo que se llamaba Combined Online Information System (COINS, Sistema de Información Combinada Online, que era la base de datos más abarcativa y detallada de finanzas públicas británicas. Trabajando con Lisa Evans (antes de que se sumara al equipo del Datablog en The Guardian), Julian Todd y Francis Irving (conocidos por Scraperwiki), Martin Rosenbaum (BBC) y otros, presentamos numerosos pedidos de datos, muchos de ellos con éxito (la saga está parcialmente documentada por Lisa en el cuadro de texto "Using FOI to Understand Spending") (Usar LDI para entender el gasto, en la página 120 de este manual.)

Cuando los datos fueron finalmente difundidos a mediados de 2010, fue considerado un golpe en favor de la transparencia. Se nos dio acceso por adelantado a los datos para poder cargarlos en nuestra aplicación en la red y recibimos significativa atención de la prensa cuando se hizo público este hecho. El día en que se puso a disposición del público, tuvimos docenas de periodistas que aparecieron en nuestro canal de chat para debatir y preguntar

sobre el hecho, así como averiguar cómo abrir la aplicación y explorarla (los archivos tenían decenas de gigabytes). Si bien algunos críticos sostuvieron que la publicación masiva de datos era tan complicada que en los hechos era oscurecer las cosas de tanta transparencia, muchos periodistas valientes se metieron a investigar en los datos para dar a sus lectores un cuadro sin precedentes del gasto público. The Guardian transmitió el evento en vivo en su blog y otros medios lo cubrieron y ofrecieron conclusiones basadas en los datos.

No tardaron mucho en llegar pedidos y preguntas respecto de proyectos similares en otros países del mundo. Poco después de lanzar OffenerHaushalt -una versión del proyecto para el presupuesto estatal alemán creado por Friedrich Lendenberg- lanzamos OpenSpending, una versión internacional del proyecto, que apunta a ayudar a los usuarios a seguir el gasto público de todo el mundo, un poco como el OpenStreetMap ayudó a hacer el mapa de accidentes geográficos. Implementamos nuevos diseños con ayuda del talentoso Gregor Aisch, basados parcialmente en los diseños originales de David McCandless.

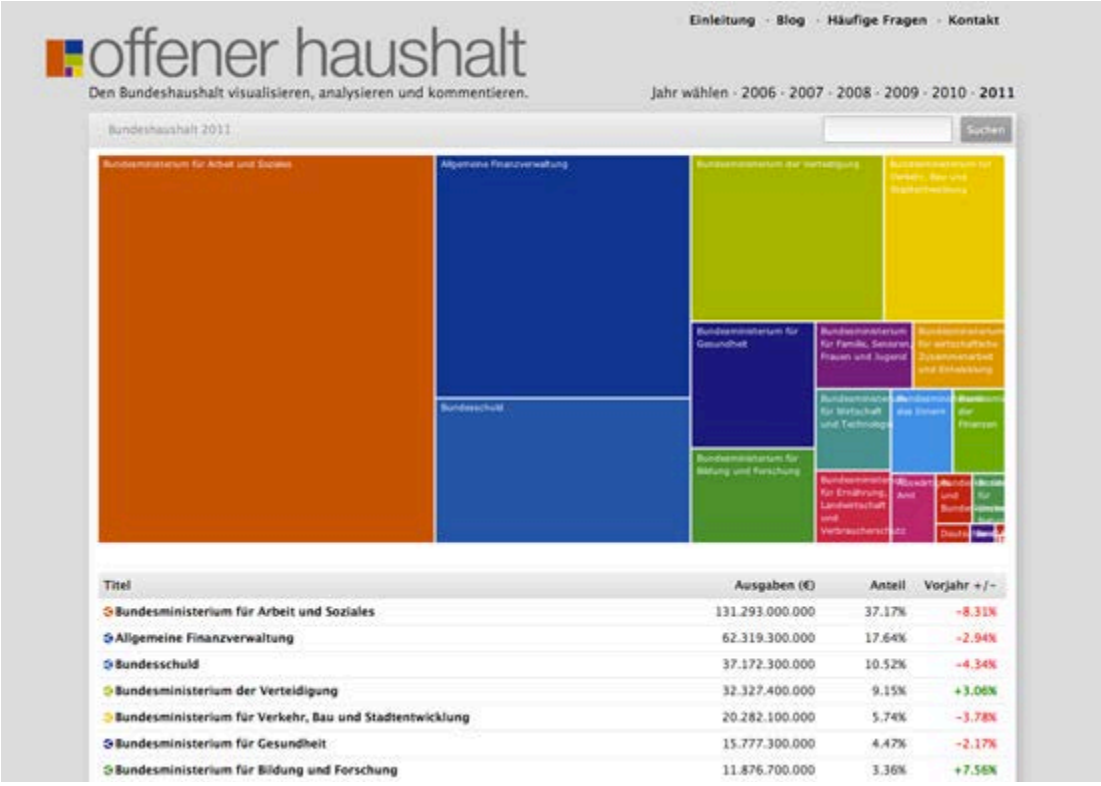


Figure 10. OffenerHaushalt, la versión alemana de ¿A dónde va mi dinero? (Open Knowledge Foundation)

Con OpenSpending, hemos trabajado extensamente con periodistas para adquirir, representar, interpretar y presentar datos de gasto público. El proyecto es en primer lugar una base de datos enorme del gasto público –tanto información presupuestaria de alto nivel como gasto efectivo al nivel de las transacciones- en la que se puede hacer búsquedas. Sobre esto se ha construido una serie de visualizaciones tales como "treemaps"(gráficos de

rectángulos anidados) y "bubbletrees" (gráficos de burbujas anidadas). Cualquiera puede cargar los datos de su municipalidad y producir visualizaciones.

Inicialmente creímos que habría mayor demanda de nuestras visualizaciones más sofisticadas, pero luego de hablar con organizaciones de noticias advertimos que había necesidades más básicas que debían ser satisfechas primero, tales como la capacidad de insertar tablas dinámicas de datos en sus blogs. Deseosos de alentaras a dar acceso público a los datos junto con sus historias, también creamos una aplicación para esto.

Nuestro primer gran lanzamiento fue en la época del primer Festival Internacional de Periodismo en Perugia. Un grupo de programadores, periodistas y empleados públicos colaboraron para cargar datos italianos en la plataforma de OpenSpending, que daba una rica visión de cómo se dividía el gasto entre las administraciones regionales y locales y central. Apareció en Il Fatto Quotidiano, Il Post, La Stampa, Repubblica, y Wired Italia, así como en The Guardian.

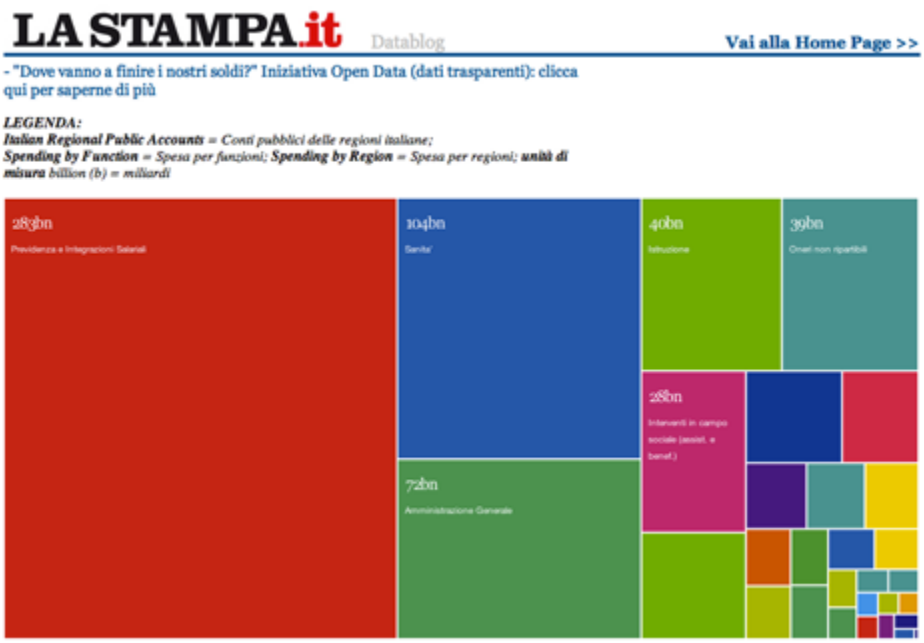


Figure 11. Versión italiana de ¿A dónde va mi dinero? (La Stampa)

En 2011 trabajamos con Publish What You Fund (Publique lo que financia), y el Overseas Development Institute (Instituto de Desarrollo en el Extranjero, para rastrear la ayuda financiera a Uganda de 2003-2006. Esto era nuevo porque por primera vez se podía ver los flujos de ayuda junto con el presupuesto nacional, lo que permite ver en qué medida las prioridades de los donantes están alineadas con las prioridades de los gobiernos. Hubo algunas conclusiones interesantes, por ejemplo tanto los programas contra el HIV como la planificación familiar resultaron estar financiadas casi completamente por donantes externos. Esto fue cubierto en The Guardian.

También hemos estado trabajando con ONGs y grupos interesados para cruzar los datos del gasto con otras fuentes de información. Por ejemplo, Privacy International se conectó con nosotros trayendo una larga lista de compañías de tecnología de vigilancia y una lista de entes que participaron de una feria internacional de la vigilancia muy famosa, que se conoce como la "fiesta de los que colocan micrófonos ocultos". Cruzando nombres de empresas con conjuntos de datos de gasto, fue posible identificar qué compañías tenían contratos oficiales, los que a partir de allí podían seguirse a través de pedidos de acceso a la información al Estado. Esto fue cubierto por The Guardian.

Actualmente, estamos trabajando para aumentar el entendimiento de los datos fiscales por periodistas y el público en general como parte de un proyecto llamado Spending Stories (Historias de Gastos, que permite a los usuarios vincular datos de gasto público con historias relacionadas, para ver las cifras detrás de las noticias y las noticias a partir de los números.

A través de nuestro trabajo en esta área aprendimos que:

- Los periodistas a menudo no están acostumbrados a trabajar con datos en crudo y muchos no consideran tenerlos como base para sus informes. Basar historias en información cruda sigue siendo una idea relativamente nueva.
- Analizar y comprender datos es un proceso que exige mucho tiempo, incluso si se tiene las capacidades requeridas. Es difícil encajar esto en un ciclo de noticias de corto plazo, por lo que el periodismo de datos a menudo es utilizado en proyectos de investigación de más largo plazo.
- Los datos difundidos por los gobiernos a menudo están incompletos o son viejos. Muy a menudo, las bases de datos públicas no pueden ser usadas para propósitos de investigación sin el agregado de piezas de información más específicas requeridas a través de las normas de acceso a la información pública.
- Grupos de interesados, estudiosos e investigadores a menudo tienen más tiempo y recursos para realizar investigaciones basadas en datos más extensas que los periodistas. Puede ser muy fructífero hacer equipo con ellos.

— Lucy Chambers and Jonathan Gray, Open Knowledge Foundation

Elecciones parlamentarias finlandesas y financiación de campañas

En los últimos meses ha habido juicios relacionados con financiación de campañas en las elecciones generales finlandesas de 2007.

Después de esos comicios la prensa descubrió que las leyes sobre publicidad de la financiación de las campañas no tenía efecto sobre los políticos. Básicamente, se ha utilizado la financiación de campañas para comprar los favores de políticos que no declararon su financiación tal como lo ordena la ley finlandesa.

A partir de estos incidentes, las leyes se volvieron más estrictas. Luego de la elección general de marzo de 2011, Helsingin Sanomat decidió explorar cuidadosamente todos los datos disponibles sobre financiación de campañas. La nueva ley estipula que se debe declarar la financiación electoral, y solo las donaciones de menos de 1500 euros pueden ser anónimas.

1. Encontrar datos y programadores

Helsingin Sanomat ha organizado hackatones HS Open desde marzo 2011. Invitamos programadores, periodistas y diseñadores gráficos finlandeses al sótano de nuestro edificio. Los participantes son divididos en grupos de 3 personas y se los alienta a desarrollar aplicaciones y visualizaciones. Hemos tenido alrededor de 60 participantes en cada uno de nuestros 3 eventos hasta la fecha. Decidimos que los datos de finanzas de campaña debían ser el centro de HS Open #2, en mayo de 2011.

La Oficina Nacional de Auditoría de Finlandia es la autoridad que lleva registro de las finanzas de campaña. Esa fue la parte fácil. El jefe de información, Jaakko Hamunen, construyó un sitio en la red que da acceso en tiempo real a su base de datos de finanzas de campaña. La Oficina de Auditoría lo hizo solo en 2 meses después de nuestro pedido.

El sitio Vaalirahoitus.fi proveerá al público y la prensa información de las finanzas de campaña para cada elección a partir de ahora.

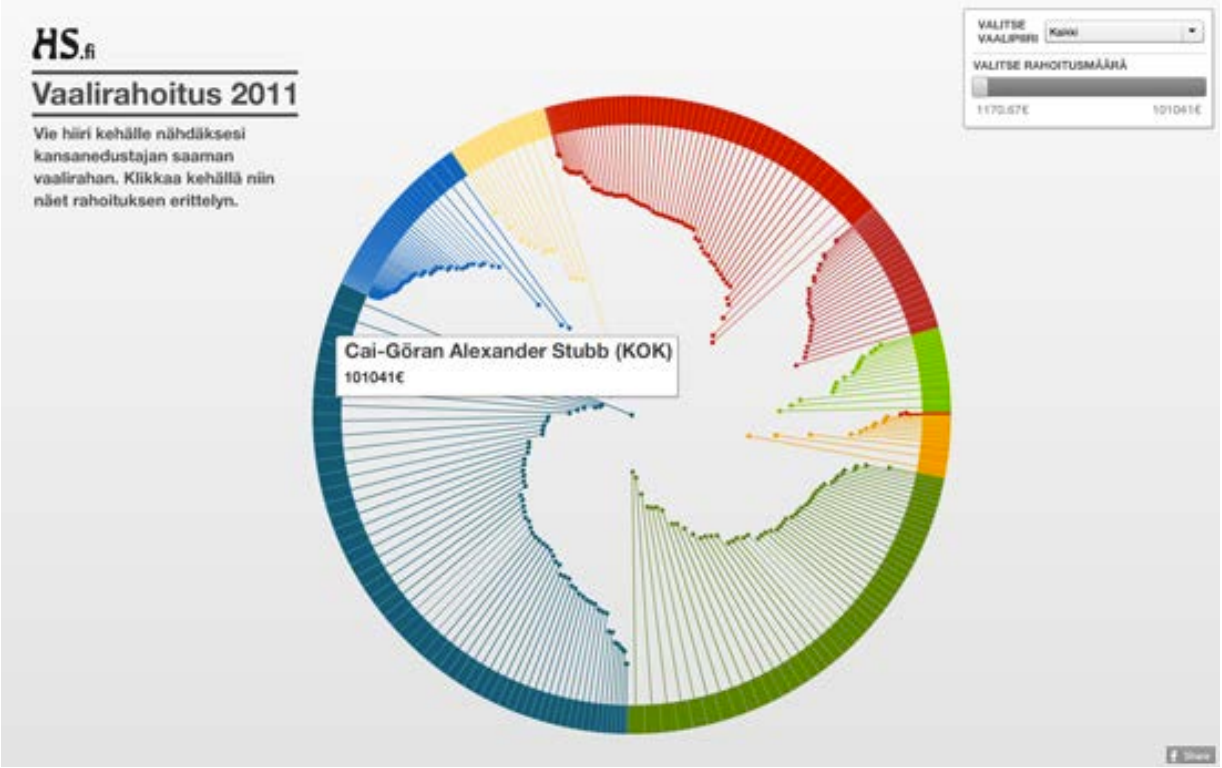


Figure 12. Finanzas electorales (Helsingin Sanomat)

2. Tormenta de ideas

Los participantes de HS Open 2 generaron veinte prototipos distintos respecto de qué hacer con los datos. Puede encontrar todos los prototipos en nuestro sitio, (texto en finlandés).

El investigador de bio-informática Janne Peltola señaló que los datos de las finanzas de campaña se veían parecidos a los datos de genes que ellos investigan, en términos de contener muchas interdependencias.

En la bio-informática hay una herramienta de código abierto llamada Cytoscape que se usa para mapear estas interdependencias. Por lo que procesamos los datos con Cytoscape, y obtuvimos un prototipo muy interesante.

3. Implementar la idea en papel y en la red

La ley de financiación de campañas dice que los miembros electos del parlamento deben declarar su financiación 2 meses después de las elecciones. En la práctica esto significó que obtuvimos los datos reales a mediados de junio. En HS Open solo tuvimos datos de parlamentarios que habían presentado su información antes del vencimiento del plazo.

También hubo un problema con el formato de los datos. La Oficina Nacional de Auditoría los proveyó en 2 archivos CSV. Uno contenía el presupuesto total de las campañas, el otro listaba todos los donantes. Tuvimos que combinar estos 2 creando un archivo que contenía 3 columnas: donantes, receptor y monto. Si los políticos habían usado su propio dinero, en nuestro formato de datos se veía como que el Político A donó X euros al Político A. Quizás resulte contra-intuitivo, pero funcionó para Cytoscape.

Cuando los datos fueron depurados y reformateados, lo corrimos con Cytoscape. Entonces nuestro departamento interactivo hizo un gráfico a toda página.

Finalmente creamos una hermosa visualización en nuestro sitio. Este no fue un gráfico de análisis de redes. Queríamos ofrecer a la gente una manera fácil de explorar los fondos de campaña y quién los dona. La primera vista muestra la distribución de fondos entre parlamentarios. Cuando se cliquea en un parlamentario se tiene el desglose de su financiación. También se puede votar si este donante particular es bueno o no. La visualización fue hecha por Juha Rouvinen y Jukka Kokko, de una agencia publicitaria llamada Satumaa.

La versión de la red de la visualización de finanzas de campaña usa los mismos datos que el análisis de redes.

4. Publicar los datos

Por supuesto que la Oficina Nacional de Auditoría ya publica los datos, por lo que no hay

necesidad de volver a publicarlos. Pero, como habíamos depurado los datos y les habíamos dado una mejor estructura, decidimos publicarlos. Damos nuestros datos con una licencia de Creative Commons Attribution. Después varios programadores independientes hicieron visualizaciones de los datos, algunas de las cuales hemos publicado.

Las herramientas que usamos para el proyecto fueron Excel y Google Refine para la depuración y análisis de los datos; Cytoscape para el análisis de redes; e Illustrator y Flash para las visualizaciones. El Flash debió haber sido HTML5, pero se nos acabó el tiempo.

¿Qué aprendimos? Quizás la lección más importante fue que las estructuras de datos pueden ser muy difíciles. Si los datos originales no están en un formato adecuado, recalcular y convertirlos lleva mucho tiempo.

Hack electoral en tiempo real (Hacks/Hackers Buenos Aires)

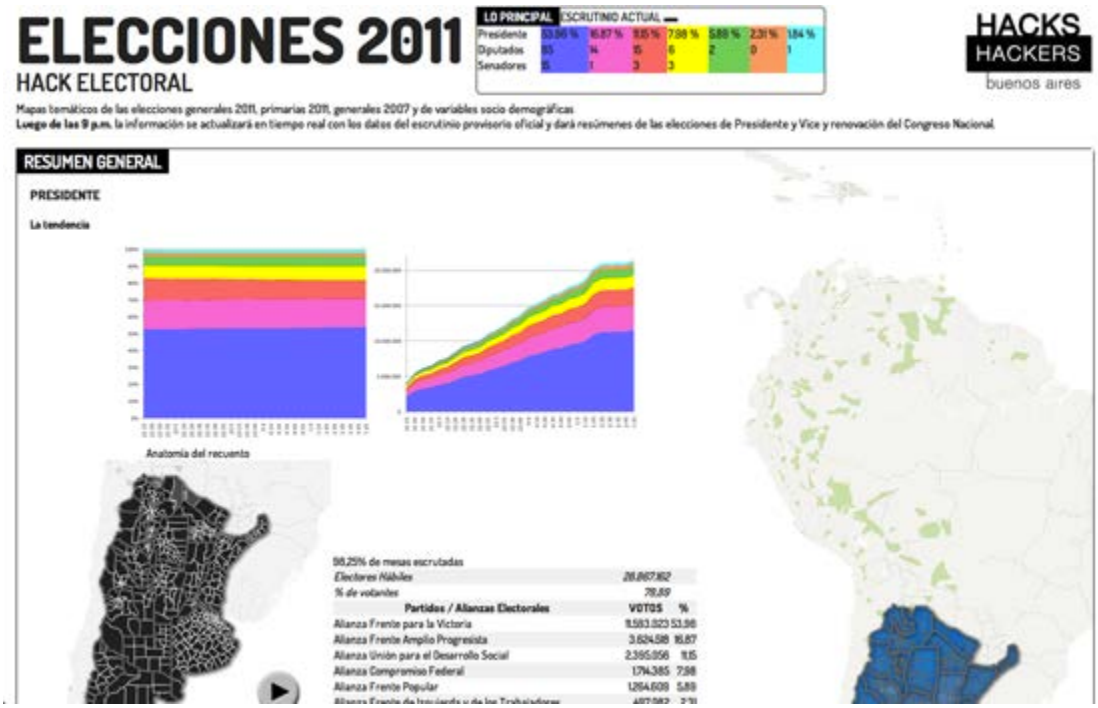


Figure 13. Elecciones 2011 (Hacks/Hackers Buenos Aires)

Electoral Hack es un proyecto de análisis político que visualiza datos de los resultados provisionales de las elecciones de octubre de 2011 en la Argentina. El sistema también incluye información de anteriores elecciones y estadísticas demográficas de todo el país. El proyecto fue actualizado en tiempo real con información del recuento provisional de las elecciones nacionales de 2011 en ese país y dio resúmenes de los resultados. Fue una iniciativa de Hacks/Hackers Buenos Aires con el analista político Andy Tow, y un esfuerzo colaborativo de periodistas, programadores, diseñadores, analistas, cientistas políticos e otros integrantes del capítulo local de Hacks/Hackers.

¿Qué datos usamos?

Todos los datos provienen de fuentes oficiales: la Dirección Nacional Electoral dio acceso a los datos del recuento provisional por Indra; el Ministerio del Interior dio información sobre cargos electorales y candidatos de los distintos partidos políticos; un proyecto universitario dio información biográfica y las plataformas políticas de cada lista electoral; mientras que la información socio-demográfica provino del Censo Nacional de 2001 de Población y Vivienda (INDEC), el censo de 2010 (INDEC) y el ministerio de Salud.

¿Cómo se desarrolló?

La aplicación fue generada durante el Hackatón Electoral 2011 de Hacks/Hackers Buenos Aires, el día antes de las elecciones del 23 de octubre de 2011. El hackatón tuvo la participación de 30 voluntarios de una variedad de especialidades. El Hack Electoral fue desarrollado como una plataforma abierta que podría mejorarse con el tiempo. Para la tecnología usamos Google Fusion Tables, Google Maps y bibliotecas de gráficos vectoriales.

Trabajamos en la construcción de polígonos para presentar mapeado geográfico y demografía electoral. Combinando polígonos en software GIS y geometrías de tablas públicas en las Tablas de Fusión Google, generamos tablas con claves correspondientes a la base de datos electorales del ministerio del Interior, Indra y datos socio-demográficos de INDEC. A partir de esto creamos visualizaciones en Google Maps.

Usando el API Google Maps, publicamos varios mapas temáticos representando la distribución espacial de la votación con distintos tonos de color, donde la intensidad del color representaba el porcentaje de votos para varias candidaturas presidenciales en distintos departamentos administrativos y centros de votación, con particular énfasis en centros urbanos importantes: de la ciudad de Buenos Aires, los 24 distritos del Gran Buenos Aires y las ciudades de Córdoba y Rosario.

Usamos la misma técnica para generar mapas temáticos de elecciones anteriores, a saber las primarias presidenciales de 2011 y la elección de 2007, así como la distribución de los datos socio-demográficos, tales como los de pobreza, mortalidad infantil y condiciones de vida, permitiendo análisis y comparaciones. El proyecto también mostró la distribución espacial de las diferencias porcentuales de votos obtenidos por cada candidatura en la elección general de octubre, comparado con la elección primaria de agosto.

Luego, usando datos de recuentos parciales, creamos un mapa animado presentando la anatomía del recuento, en el que se muestra el avance del mismo desde el cierre de la votación hasta la mañana siguiente.

Pros

- Nos propusimos representar datos y lo logramos. Teníamos la base de datos socio-demográfica infantil de UNICEF, a mano así como la base de datos creada por el yoquierosaber.org de la Universidad Torcuato Di Tella. Durante el hackatón reunimos un gran volumen de datos adicionales que terminamos no incluyendo.
- Claramente el trabajo periodístico y de programación se vio enriquecido por los estudios académicos. Sin la contribución de Andy Tow e Hilario Moreno Campos, el proyecto no se hubiera podido realizar.

Contras

- Los datos socio-demográficos que pudimos utilizar no estaban actualizados (la mayor parte era del censo de 2001) y no era muy granular. Por ejemplo, no incluía detalles de PBI promedio local, principal actividad económica, nivel educativo, número de escuelas, médicos per cápita y muchas otras cosas que hubiera sido bueno tener.
- Originalmente el sistema debía ser una herramienta que pudiera usarse para combinar y mostrar datos arbitrariamente, de modo que el periodista pudiera mostrar fácilmente datos que le interesaran en la red. Pero tuvimos que dejar esto para otro momento.
- Dado que el proyecto fue creado por voluntarios en un plazo breve, fue imposible hacer todo lo que queríamos. De todos modos avanzamos mucho en el sentido adecuado.
- Por el mismo motivo, todo el trabajo colaborativo de 30 personas terminó condensado en un solo programador cuando los datos del gobierno comenzaron a aparecer, y tuvimos problemas para importar datos en tiempo real. Estos problemas se resolvieron en cuestión de horas.

Implicancias

La plataforma de Hack Electoral tuvo gran impacto en los medios, con cobertura en televisión, radio, medios impresos y online. Mapas del proyecto fueron utilizados por varias plataformas de medios durante las elecciones y en días subsecuentes. Con el paso del tiempo, los mapas y visualizaciones fueron actualizados, incrementando aún más el tráfico. El día de la elección, el sitio creado ese mismo día recibió alrededor de 20.000 visitantes diferentes y sus mapas fueron reproducidos en la tapa del diario Página/12 2 días consecutivos, así como en artículos en La Nación. Algunos mapas aparecieron en las ediciones impresas del diario Clarín. Fue la primera vez en la historia del periodismo argentino que se utilizó un despliegue interactivo de mapas en tiempo real. En los mapas centrales se podía ver claramente la victoria abrumadora de Cristina Fernández de Kirchner por el 54 por ciento de los votos, desglosada por la saturación de color. También sirvió para ayudar a los usuarios a entender casos específicos donde candidatos locales tuvieron victorias por amplio margen en las provincias.

— Mariano Blejman, Mariana Berruezo, Sergio Sorín, Andy Tow, and Martín Sarsale from Hacks/Hackers Buenos Aires

Datos en las noticias: WikiLeaks

Comenzó con uno de los integrantes del equipo de periodismo investigativo preguntando: "¿Ustedes son buenos con las planillas de cálculo verdad?" Y esta era una enorme planilla de cálculo: 92.201 filas de datos, cada una conteniendo un análisis de un evento militar en Afganistán. Estos fueron los registros de la guerra de WikiLeaks. En realidad, la primera parte. Siguieron 2 episodios más: Irak y los cables. El término oficial fue SIGACTS: la base de datos de Acciones Significativas de las Fuerzas Armadas de Estados Unidos.

Los registros de guerra de Afganistán –compartidos con The New York Times y Der Spiegel- fueron periodismo de datos en acción. Lo que queríamos hacer era permitir a nuestro equipo de periodistas especializados obtener grandes historias humanas a partir de la información y queríamos analizarlos para tener el cuadro general, mostrar cómo iba la guerra realmente.

Desde el comienzo fue central para lo que íbamos a hacer saber que no publicaríamos toda la base de datos. WikiLeaks ya iba a hacer eso y queríamos estar seguros de no revelar los nombres de los informantes, o poner en peligro innecesariamente tropas de la OTAN. Al mismo tiempo, teníamos que hacer más fácil el uso de los datos para nuestro equipo de periodistas investigadores encabezados por David Leigh y Nick Davies (que habían negociado la difusión de los datos con Julian Assange). También queríamos simplificar el acceso a información clave en el mundo real, haciéndola tan clara y abierta como pudiéramos.

Los datos llegaron a nosotros como un inmenso archivo Excel, más de 92.201 filas de datos, algunas conteniendo nada o mal formateadas. No le servía a los periodistas que trataban de buscar historias y era demasiado grande como para hacer informes significativos.

Nuestro equipo creó una base de datos interna simple usando SQL. Los periodistas podían a partir de allí buscar por medio de palabras clave o eventos. De pronto el conjunto de datos se volvió accesible y generar historias se hizo más fácil.

Los datos estaban bien estructurados: cada evento tenía los siguientes datos claves: hora, día, descripción, cifras de bajas y, crucialmente, latitud y longitud detalladas.

También comenzamos a filtrar los datos para ayudarnos a contar una de las historias claves de la guerra: el aumento de los ataques con DEI (dispositivos explosivos improvisados), bombas caseras al costado del camino que son impredecibles y difíciles de combatir. Este conjunto de datos seguía siendo enorme pero más fácil de manejar. Hubo alrededor de 7500 explosiones o emboscadas con DEI (una emboscada es donde el ataque se combina, por ejemplo, con fuego de armas pequeñas o de misiles con granadas) entre 2004 y 2009. Hubo otros 8000 DEI descubiertos y desactivados. Queríamos ver cómo cambiaban con el tiempo y hacer comparaciones. Estos datos nos permitieron ver que el sur, donde estaban las tropas

británicas y canadienses, era la zona más golpeada, lo que confirmaba lo que sabían nuestros corresponsales que habían cubierto la guerra.

La difusión de los registros de la guerra de Irak en octubre de 2010 descargó otros 391.000 registros de la guerra de Irak en la escena pública.

Esto estaba en una categoría diferente de la filtración sobre Afganistán; se puede decir que [line-through]*esto*convirtió a esta en la guerra más documentada de la historia. Ahora contábamos con cada detalle menor para analizarlo y desglosarlo. Pero se destaca un factor: el volumen de las muertes, la mayoría de las cuales eran de civiles.

Tal como en el caso de Afganistán, The Guardian decidió no volver a publicar la base de datos completa, en gran medida porque no podíamos estar seguros de que el conjunto no contuviera detalles confidenciales de informantes y demás.

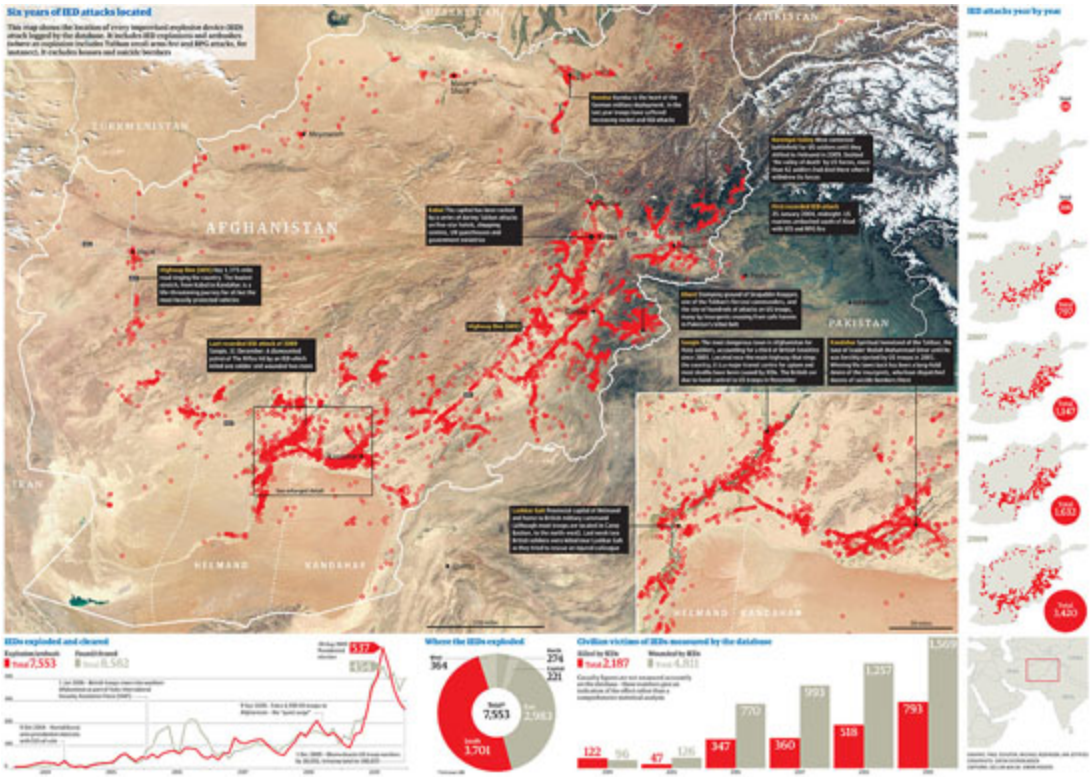


Figure 14. Los registros de guerra de The WikiLeaks (The Guardian)

Pero sí permitimos a nuestros usuarios descargar una planilla de cálculo que contenía los registros de cada incidente en el que alguien murió, casi 60.000 en total. Eliminamos el sumario por lo que solo estaban los datos básicos: el encabezado militar, la cantidad de muertes y la ubicación geográfica.

También tomamos todos estos incidentes en los que murió alguien y los pusimos en un mapa usando Google Fusion Tables. No era perfecto, pero sí un comienzo para tratar de mapear los patrones de destrucción que habían devastado Irak.

Los cables se difundieron en diciembre de 2010. Esto entraba en otra liga, un inmenso conjunto de datos de documentos oficiales: 251.287 despachos, de más de 250 embajadas y consulados estadounidenses. Es un cuadro único de lenguaje diplomático de EE.UU., incluyendo más de 50.000 documentos que cubren la actual administración Obama. ¿Qué incluían los datos?

Los cables mismos vinieron vía el inmenso Secret Internet Protocol Router Network (Red de Ruteo del Protocolo Secreto de Internet), conocido por la sigla SIPRNet. Es el sistema de Internet mundial militar de Estados Unidos, que se mantiene separado de la Internet civil común y es dirigida por el departamento de Defensa en Washington. Desde los ataques de septiembre de 2001 había habido una iniciativa en Estados Unidos de vincular archivos de información gubernamental, con la esperanza de que datos claves de inteligencia ya no quedaran atrapados en compartimentos estancos o presentados fuera de contexto. Un número creciente de embajadas de EE.UU. han sido conectados a SIPRNet en la última década, de modo que pueda compartirse la información militar y diplomática. Para 2001, había 125 embajadas en SIPRNet; para 2005 la cifra había crecido a 180 y a esta altura la gran mayoría de las misiones de EE.UU. en el mundo entero están vinculadas con el sistema, que es el motivo por el que grueso de estos cables son de 2008 y 2009. Como escribió David Leigh:

Un despacho de embajada marcado como SIPDIS es descargado automáticamente al sitio clasificado de la embajada. Allí no solo puede verlo cualquiera en el departamento de Estado, sino cualquiera de las fuerzas armadas de EE.UU. que tenga acceso de seguridad de nivel “Secreto”, una clave, y una computadora conectada a SIPRNet.

1. lo que asombrosamente abarca a 3.000.000 de personas. Hay varias capas de datos aquí; llegando hasta la clasificación de *SECRET NOFORN*, que significa que no podrán mostrarse jamás a quienes no sean ciudadanos estadounidenses. En cambio se supone que son para que los lean funcionarios en Washington hasta el nivel de la secretaria de Estado, Hillary Clinton. Los cables normalmente son redactados por el embajador local o subordinados. No se puede acceder a los documentos de “Secreto Máximo” y por encima de documento de inteligencia extranjera desde SIPRNet.

A diferencia de las anteriores entregas, esta era predominantemente de texto, no cuantificada ni con datos idénticos. Esto es lo que incluía:

Una fuente

La embajada o el ente que envió los datos

Una lista de receptores

Normalmente los cables eran enviados a una cantidad de embajadas y entes.

Un campo de tema

Una síntesis del cable.

Etiquetas

Cada cable estaba etiquetado con una cantidad de abreviaturas claves.

Cuerpo del texto

El cable mismo. Optamos por no publicar estos completos por razones obvias.

Un detalle interesante de esta historia es cómo los cables casi crearon filtraciones a demanda. Durante semanas ocuparon el centro de las noticias al ser publicada; ahora, cada vez que aparece una historia acerca de algún régimen corrupto o un escándalo internacional, el acceso a los cables nos da nuevas historias.

El análisis de los cables es una tarea enorme que quizás nunca se termine por completo.

— Esta es una versión editdada de un capítulo publicado en Facts are Sacred: The Power of Data de Simon Rogers, the Guardian (published on Kindle)

Hackatón Mapa76

Inauguramos el capítulo de Hacks/Hackers de Buenos Aires en abril de 2011. Fuimos anfitriones de 2 encuentros iniciales para difundir la idea de mayor colaboración entre periodistas y programadores que incluyó entre 120 y 150 personas en cada evento. Para una tercera reunión tuvimos un hackatón de 30 horas con 8 personas en una conferencia de periodismo digital en la ciudad de Rosario, a 300 kilómetros de Buenos Aires.

Un tema recurrente en estas reuniones fue el deseo de recoger grandes volúmenes de datos de la red y luego representarlos visualmente. Para ayudar con esto, nació un proyecto llamado Mapa76.info, que ayuda a los usuarios a extraer datos y luego desplegarlos usando mapas y líneas de tiempo. Una tarea nada fácil.

Figure 15. Mapa76 (Hacks/Hackers Buenos Aires)

¿Por qué Mapa76? El 24 de marzo de 1976 hubo un golpe de Estado en la Argentina, que duró hasta 1983. En ese período hubo según se estima 30.000 desaparecidos, miles de muertes y 500 niños nacidos en cautiverio apropiados por la dictadura militar. Pasados más de 30 años, la cantidad de gente condenada en la Argentina por crímenes de lesa humanidad cometidos durante la dictadura llega a 262 personas (septiembre de 2011). En este momento hay 14 juicios en curso y 7 con fecha de comienzo establecida. Hay 802 personas en varios casos en las cortes.

Estos juicios generan grandes volúmenes de datos que son difíciles de procesar para los investigadores, periodistas, organizaciones de derechos humanos, jueces, fiscales y otros. Los datos se producen de modo distribuido y los investigadores a menudo no recurren a herramientas de software para ayudarse a interpretarlos. Esto significa que a menudo no son tenidos en cuenta y las hipótesis son limitadas. Mapa76 es una herramienta de investigación que da acceso abierto a esta información con propósitos periodísticos, legales, jurídicos e históricos.

Para preparar el hackatón creamos una plataforma que programadores y periodistas pudieran usar para colaborar en el día del evento. Martín Sarsale desarrolló algunos algoritmos básicos para extraer datos estructurados de documentos de texto simples. También se usaron algunas bibliotecas del proyecto DocumentCloud.org, pero no demasiadas. La plataforma analiza y extrae de manera automática nombres, fechas y lugares de textos y permite a los usuarios explorar datos claves sobre distintos casos (por ejemplo, fecha de nacimiento, lugar de arresto, supuesto lugar de desaparición y así siguiendo).

Nuestra meta era proveer una plataforma para la extracción automática de datos sobre los juicios contra la dictadura militar en la Argentina. Queríamos una manera de desplegar automáticamente (o al menos semi-automáticamente) datos claves relacionados con casos entre 1976 y 1983 basado en evidencias escritas, argumentos y juicios. Los datos extraídos (nombres, lugares y fechas) son recogidos, almacenados y pueden ser analizados y refinados por el investigador, así como explorados usando mapas, líneas de tiempo y herramientas de análisis de redes.

El proyecto permitirá a periodistas e investigadores, fiscales y testigos seguir la historia de vida de una persona, incluyendo por supuesto su cautiverio y posterior desaparición o liberación. Donde falte información, los usuarios pueden buscar en un vasto número de documentos que podrían ser de posible relevancia para el caso.

Para el hackatón hicimos un anuncio público a través de Hacks/Hackers Buenos Aires, que entonces tenía alrededor de 200 miembros (en el momento de escribir este informe hay alrededor de 540). También contactamos muchas asociaciones de derechos humanos. De la reunión participaron unas cuarenta personas, incluyendo periodistas, organizaciones de defensa de los derechos humanos, programadores y diseñadores.

Durante el hackatón identificamos tareas que distintos tipos de participantes podían desarrollar de forma independiente para ayudar a que las cosas funcionaran bien. Por ejemplo, pedimos a diseñadores que trabajaran en una interfaz que combinara mapas y líneas de tiempos, pedimos a programadores que analizaran maneras de extraer datos estructurados y logaritmos para eliminar ambigüedades relacionadas con nombres, y pedimos a periodistas que investigaran qué había pasado con gente específica, para comparar distintas versiones de historias y analizar documentos para narrar historias sobre casos particulares.

Probablemente el principal problema que tuvimos después del hackatón fue que nuestro proyecto era muy ambicioso, nuestros objetivos de corto plazo exigentes, y es difícil coordinar una red de voluntarios dispersos. Casi todos los involucrados con el proyecto tenían empleos que les ocupaban mucho tiempo y muchos participaban además de otros eventos y proyectos. Hacks/Hackers Buenos Aires tuvo 9 reuniones en 2011.

El proyecto está actualmente en desarrollo activo. Hay un equipo central de 4 personas trabajando con más de una docena de colaboradores. Tenemos una lista de correo pública y un centro de almacenado de código a través del cual cualquiera puede involucrarse en el proyecto.

— Mariano Blejman, Hacks/Hackers Buenos Aires

Cobertura de los disturbios en el Reino Unido por el Datablog de The Guardian

Durante el verano de 2011, hubo una oleada de disturbios en el Reino Unido. En aquel momento, algunos políticos sugirieron que estas acciones categóricamente no estaban vinculadas con la pobreza y los que saquearon fueron simplemente criminales. Lo que es más, el primer ministro, junto con los principales políticos conservadores, culparon a los medios sociales por causar los disturbios, sugiriendo que había habido incitación desde estas plataformas y que los disturbios fueron organizados a través de Facebook, Twitter y Blackberry Messenger (BBM). Hubo reclamos para cerrar temporariamente los medios sociales. Debido a que el gobierno no hizo una investigación de por qué se dieron los disturbios, The Guardian, en colaboración con la London School of Economics, creó un proyecto innovador para abordar estas cuestiones, llamado Reading the Riots (Leer los Disturbios),

Figure 16. Los disturbuios en Reino Unido: todos los incidentes verificados (The Guardian)

El diario usó periodismo de datos a gran escala para permitir al público comprender mejor quién saqueaba y por qué. También trabajaron con otro equipo de académicos, encabezados por el profesor Rob Procter de la universidad de Manchester para entender mejor el rol de los medios sociales, que The Guardian mismo había usado abundantemente para sus informes durante los disturbios. El equipo de *Reading the Riots* fue encabezado por Paul Lewis, el Editor de Proyectos Especiales de The Guardian. Durante los disturbios Paul reportó desde el lugar de los eventos en ciudades de toda Inglaterra (fundamentalmente a través de su cuenta de Twitter @paullewis). Este segundo equipo trabajó a partir de

26.000.000 de tweets sobre los disturbios puestos a disposición por Twitter. El objetivo principal de este trabajo con los medios sociales fue ver cómo circulan los rumores en esa red social, la función que tienen distintos usuarios/actores en la propagación y difusión de flujos de información, para ver si se usó la plataforma para incitar, y para examinar otras formas de organización.

En términos del uso del periodismo de datos y visualizaciones, es útil distinguir 2 períodos claves: el período de los disturbios mismos y las maneras en que los datos ayudaron a narrar historias mientras se desarrollaban los disturbios; y luego un segundo período de investigación mucho más intensa con 2 conjuntos de equipos académicos trabajando con The Guardian, para recolectar datos, analizarlos, y escribir informes con análisis de fondo sobre lo descubierto. Los resultados de la primera fase del proyecto *Reading the Riots* fueron publicados durante una semana de cobertura extensiva a comienzos de diciembre de 2011. A continuación aparecen algunos ejemplos claves de cómo se usó el periodismo de datos durante ambos períodos.

Primera fase: los disturbios mientras sucedían

Usando mapas simples, el equipo de datos de The Guardian mostró localizaciones de lugares de disturbios confirmados y combinando datos de pobreza con los lugares donde se dieron los disturbios, se comenzó a dejar sin sustento el discurso político central de que no había ningún vínculo con la pobreza. En ambos ejemplos se utilizaron herramientas de mapeo preexistentes y, en el segundo caso, se combinaron datos de ubicación con otro conjunto de datos para comenzar a establecer otras conexiones y vínculos.

En relación al uso de medios sociales durante los disturbios (en este caso, Twitter), el diario creó una visualización de hashtags relacionadas con los disturbios usadas durante este período, lo que destacó que Twitter fue usado principalmente para responder a disturbios en vez de para organizar a gente para saquear, con #riotcleanup, la campaña espontánea para limpiar las calles después de los disturbios, mostrando el salto más significativo durante el período de los disturbios.

Fase 2: análisis de los disturbios

Con relación al momento en que el diario informó sus conclusiones luego de meses de investigaciones intensivas trabajando en estrecha colaboración con 2 equipos académicos, se destacan 2 visualizaciones que han sido ampliamente debatidas. La primera, un video corto, muestra los resultados de combinar los lugares conocidos donde la gente protagonizó disturbios con sus lugares de vivienda y mostrando lo que se llamó “viaje a los disturbios”. Aquí el diario trabajó con un especialista en mapeo de transporte, ITO World, para hacer un

modelo de la ruta más probable utilizada por quienes protagonizaron los disturbios al dirigirse a los distintos lugares donde saquearon, lo que destaca patrones diferentes para distintas ciudades, con viajes largos en algunas de ellas.

La segunda se refiere a las maneras en que se extienden los rumores en Twitter. En debate con el equipo académico, se escogieron 7 rumores para su análisis. El equipo académico entonces recolectó todos los datos relacionados con cada rumor y diseñó un código que identifica cada tweet de acuerdo a los 4 códigos principales: gente que simplemente repite el rumor (afirma algo), lo rechaza (afirma algo contrario), lo cuestiona (interrogación) o simplemente lo comenta (comentario). Todos los tweets fueron codificados por triplicado y los resultados fueron visualizados por el equipo interactivo de The Guardian. El equipo de The Guardian escribió acerca de cómo construyó las visualizaciones.

Lo llamativo de esta visualización es que muestra de manera potente lo que es muy difícil de describir y que es la naturaleza viral de los rumores y las maneras en que se desarrolla su ciclo vital a lo largo del tiempo. El rol de los principales medios es evidente en algunos de estos rumores (por ejemplo, rechazándolos abiertamente, o confirmándolos rápidamente como noticias), al igual que la naturaleza correctiva de Twitter mismo en términos de responder a tales rumores. Esta visualización no solo ayudó mucho a narrar la historia, sino que también dio una visión real de cómo funcionan los rumores en Twitter, lo que aporta información útil para responder a eventos futuros.

Lo que resulta claro a partir del último ejemplo es la poderosa sinergia entre el diario y un equipo académico capaz de un análisis profundo de 2.600.000 de tweets producidos en los disturbios. Si bien el equipo académico creó un conjunto de herramientas para hacer su análisis, ahora están trabajando para hacer que estas estén disponibles para cualquiera que desee utilizarlas ofreciendo un centro de trabajo para su análisis. Combinado con la explicación de cómo hacer las cosas aportada por el equipo de The Guardian, constituye un estudio de caso que es útil porque muestra cómo el análisis de medios sociales y las visualizaciones pueden ser usadas para narrar historias importantes.

— Farida Vis, University of Leicester

Evaluaciones de escuelas de Illinois

Cada año la Dirección Estadual de Educación de Illinois difunde “evaluaciones” de escuelas, datos sobre la demografía y el desempeño de todas las escuelas públicas de Illinois. Es un conjunto de datos masivo. El informe de este año tenía 9500 columnas de ancho. El problema con esa cantidad de datos es decidir qué presentar. (Como sucede con cualquier proyecto de software, lo difícil no es crear el software, sino crear el software correcto).

Trabajamos con los periodistas y el editor de Educación para escoger los datos más

relevantes. (hay muchos datos que parecen interesantes, pero que un periodista le dirá que en realidad son falsos o engañosos).

También encuestamos y entrevistamos gente con hijos en edad escolar en nuestra redacción. Hicimos esto por la existencia de una brecha de empatía: ninguno de los miembros del equipo de aplicaciones de noticias tiene chicos en edad escolar. Por esta vía descubrimos muchas cosas acerca de nuestros usuarios y de la practicidad (o falta de ella) de la versión anterior de nuestro sitio sobre escuelas.

Figure 17. 2011 Los boletines de las escuelas de Illinois (Chicago Tribune)

Nos orientamos a diseñar para un par de usuarios y casos de uso específicos:

- Padres con un niño en la escuela que quieren saber cómo es el desempeño de su escuela
- Padres que trataban de determinar dónde les convenía vivir, dado que la calidad de las escuelas a menudo tiene un gran impacto sobre esa decisión

La primera vez el sitio sobre escuelas fue un proyecto de 2 diseñadores de alrededor de 6 semanas. La actualización de 2011 fue un proyecto de 2 diseñadores de 4 semanas. (en realidad hubo 3 personas trabajando activamente en el proyecto más reciente, pero ninguna de ellas era full-time, por lo que equivalen a 2).

Una pieza clave de este proyecto fue el diseño de la información. Aunque presentamos mucho menos datos de los que hay disponibles, siguen siendo *muchos* datos, y hacerlos digeribles fue un desafío. Por suerte, pudimos tomar alguien prestado de nuestra mesa de gráficos, un diseñador especializado en presentar información complicada. Nos enseñó mucho acerca del diseño de cuadros y, en general, nos guió para producir una presentación

que es legible, pero no subestima la capacidad o el deseo del lector de entender las cifras. El sitio fue creado con Python y Django. Los datos están almacenados en MongoDB: los datos sobre escuelas son heterogéneos y jerárquicos, lo que hace que no funcionen bien en una base de datos relacional (de otro modo probablemente hubiésemos usado PostgreSQL).

Por primera vez experimentamos con el marco de interfaz de usuario Bootstrap de Twitter en este proyecto y los resultados nos dejaron contentos. Los gráficos fueron dibujados con Flot.

La aplicación también alberga las muchas historias sobre evaluación escolar que hemos escrito. En ese sentido es una especie de portal; cuando hay una nueva historia de evaluación de escuelas la ubicamos a la cabeza de la aplicación, junto con listas de escuelas que son relevantes para la historia (y cuando aparece una nueva historia, a los lectores de chicagotribune.com se los reorienta hacia la aplicación, no el artículo).

Los primeros indicios muestran que a los lectores les encanta la aplicación sobre las escuelas. La retroalimentación que hemos recibido en gran medida ha sido positiva (o al menos constructiva) y la cantidad de visitas es enorme. Como premio, estos datos mantendrán su interés todo un año, por lo que aunque prevemos que se reducirán las visitas al ir desapareciendo las historias sobre escuelas en la página de inicio, nuestra experiencia nos indica que los lectores recurren a esta aplicación todo el año.

Algunas ideas claves que surgieron del proyecto son:

- Los diseñadores gráficos son nuestros amigos. Son buenos para hacer digerible información compleja.
- Hay que pedir ayuda a la redacción. Este es el segundo proyecto para el que realizamos una encuesta y entrevistas en la redacción, y es una gran manera de tener opiniones de gente reflexiva que, como nuestro público, es diversa en cuanto a sus inclinaciones y en general se siente incómoda con las computadoras.
- ¡Muestre su trabajo! Gran parte de la retroalimentación tomó la forma de pedidos de los datos que usó la aplicación. Pusimos muchos datos a disposición del público vía una API, y pronto difundiremos todo lo que no incluimos inicialmente.

— Brian Boyer, Chicago Tribune

Facturación de hospitales

Periodistas de investigación de California Watch recibieron informes de que una gran cadena de hospitales de ese estado norteamericano podía estar haciendo trampas sistemáticamente contra el programa federal Medicare que paga los tratamientos médicos de estadounidenses de 65 años o más. La trampa denunciada se llama *upcoding* (subir el código), que significa reportar pacientes con problemas más complicados de salud –con reembolsos más elevados- que los reales. Pero una fuente clave era un sindicato que estaba

enfrentado con la administración de la cadena de hospitales, y el equipo de California Watch sabía que era necesaria una verificación independiente para que la historia tuviera credibilidad.

Por suerte, el departamento de Salud de California tiene registros públicos que dan información muy detallada sobre cada caso tratado en todos los hospitales del estado. Las 128 variables incluyen hasta 25 códigos de diagnóstico del manual de "Clasificación Estadística Internacional de Enfermedades y Problemas de Salud Relacionados" (conocido comúnmente como ICD-9) publicado por la Organización Mundial de la Salud (OMS). Aunque no se identifica a los pacientes por su nombre, si aparece la edad del paciente, cómo se pagó por el tratamiento y qué hospital lo trató. Los periodistas advirtieron que con estos registros, podían ver si los hospitales propiedad de la cadena estaban informando ciertas enfermedades inusuales en proporciones significativamente mayores que en otros hospitales.

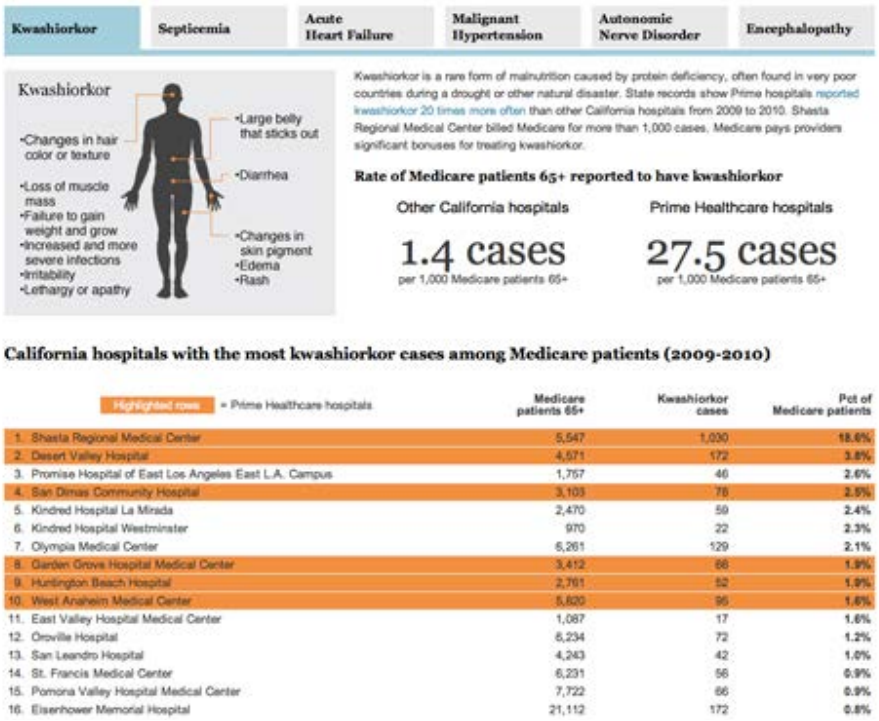


Figure 18. Kwashiorkor (California Watch)

Los conjuntos de datos eran grandes: casi 4.000.000 de registros por año. Los periodistas querían estudiar los registros de 6 años para ver cómo cambiaban los patrones a lo largo del tiempo. Pidieron los datos al ente estatal; llegaron en varios CD-ROM que se copiaron fácilmente a una computadora de escritorio. El periodista que hizo el análisis de los datos usó un sistema llamado SAS para trabajar con los datos. SAS es muy poderoso (permitiendo el análisis de muchos millones de registros) y es usado por numerosos entes estatales, incluyendo el departamento de Salud de California, pero es costoso. Se pudo haber hecho el

mismo tipo de análisis usando una variedad de herramientas de bases de datos, tales como el Access de Microsoft o MySQL de código abierto.
Con los datos y los programas para estudiarlos, encontrar patrones sospechosos fue relativamente simple. Por ejemplo, una acusación era que la cadena estaba informando de gente con diversos grados de desnutrición con porcentajes mucho más altos que lo que se veía en otros hospitales. Usando SAS, el analista de datos extrajo tablas de frecuencia que muestran la cantidad de casos de desnutrición informados cada año por cada uno de los más de 300 hospitales de agudos de California. Las tablas de frecuencia luego eran importadas a Microsoft Excel para un análisis más fino de los patrones de cada hospital; la capacidad de Excel de ordenar, filtrar y calcular tasas a partir de las cifras en bruto facilitó la tarea de encontrar patrones.

Eran particularmente llamativos los informes de una enfermedad llamada Kwashiorkor, un síndrome de deficiencia de proteínas que se ve casi exclusivamente en infantes que mueren por desnutrición en países en desarrollo afectados por hambrunas. Pero la cadena estaba informando que sus hospitales diagnosticaban Kwashiorkor entre personas mayores de California en cantidades 770 veces mayores que el promedio de los hospitales del estado.

Para otras historias, los análisis usaron técnicas similares para examinar las cantidades reportadas de enfermedades como septicemia, encefalopatía, hipertensión maligna y desórdenes nerviosos autonómicos. Otro estudio analizó las denuncias de que la cadena estaba admitiendo en internación, provenientes de sus salas de emergencias, porcentajes inusualmente elevados de pacientes de Medicare, cuya fuente de pagos de cuidados hospitalarios es más segura que lo que sucede con muchos otros pacientes atendidos en salas de emergencias.

En síntesis, historias como estas son posibles cuando se usan datos para producir evidencias que evalúan de forma independiente acusaciones de fuentes que pueden tener sus propios objetivos. Estas historias también son un buen ejemplo de la necesidad de leyes de registro público robustas; el motivo por el que el estado requiere que los hospitales informen estos datos es para que se pueda hacer este tipo de análisis, ya sea por el propio estado o por académicos, investigadores o incluso ciudadanos periodistas. El tema de estas historias es importante porque examina si se está gastando como corresponde millones de dólares de fondos públicos.

— Steve Doig, Walter Cronkite School of Journalism, Arizona State University

Crisis de los geriátricos

Una investigación del Financial Times sobre geriátricos sacó a luz como algunos inversores de capitales privados convierten el cuidado de las personas mayores en una máquina de

obtener ganancias, y destacó los costos mortales de un modelo de negocios que promueve las ganancias por encima de los cuidados.

El análisis se hizo en un buen momento, porque los problemas financieros de Southern Cross, entonces el mayor operador de geriátricos del país, estaban llegando a un punto álgido. El gobierno había impulsado durante décadas la privatización en el sector de los geriátricos y seguía aplaudiendo al sector privado por sus prácticas de negocios astutas.

Nuestra investigación comenzó con el análisis de datos que obtuvimos del ente regulador británico a cargo de inspeccionar los geriátricos. La información era pública, pero se requirió mucha persistencia para conseguir los datos en una forma que fuera utilizable.

Los datos incluían calificaciones (ahora eliminadas) del desempeño de geriátricos individuales y un desglose de si eran privados, estatales o sin fines de lucro. La Comisión de Calidad de Cuidados (CQC) hasta junio de 2010 calificaba a los geriátricos de acuerdo a su calidad (0 estrellas = mala, 3 estrellas = excelente).

El primer paso requirió mucha depuración de datos, ya que la información provista por la Comisión de Calidad de Cuidados contenían categorizaciones que no eran uniformes. Esto se hizo primordialmente usando Excel. También determinamos –a través de investigaciones de escritorio y telefónicas- si había geriátricos particulares que fueran propiedad de grupos de capitales privados. Antes de la crisis financiera, el sector de los geriátricos era un imán para el capital privado e inversores inmobiliarios, pero varios de ellos -tales como Southern Cross- habían comenzado a tener serias dificultades financieras. Queríamos establecer qué efecto, si es que había alguno, tenía el hecho de la presencia de capitales privados en la calidad de los cuidados.

Un conjunto de cálculos relativamente simples con Excel nos permitieron establecer que los geriátricos sin fines de lucro y estatales en promedio tenían un desempeño significativamente mejor que los del sector privado. Algunos grupos de geriátricos de capitales privados funcionaban por encima del promedio y otros por debajo.

Junto con informes in situ, estudios de casos de abandono, un análisis profundo de las fallas de las políticas regulatorias, así como otros datos sobre niveles de paga, tasas de rotación, etc., nuestro análisis nos permitió armar un cuadro del estado real de los geriátricos.

Algunos consejos:

- Asegúrese de tomar notas de cómo manipula los datos originales.
- Tenga una copia de los datos originales y nunca los modifique.
- Verifique y vuelva a verificar los datos. Haga el análisis varias veces (si es necesario, a partir de cero).
- Si menciona compañías o individuos particulares, deles derecho a réplica.

— Cynthia O'Murchu, Financial Times

El teléfono que lo dice todo

La comprensión de la mayoría de las personas de lo que puede hacerse con los datos que nos proveen nuestros celulares es teórica; había pocos ejemplos de la vida real. Es por eso que Malte Spitz del partido Verde Alemán decidió publicar sus propios datos. Para acceder a la información tuvo que presentar una demanda contra el gigante de las telecomunicaciones Deutsche Telekom. Los datos, contenidos en un inmenso documento de Excel, fueron la base para el mapa interactivo del Zeit Online. Cada una de las 35.831 filas de la planilla de cálculo representa una instancia en la que el teléfono de Spitz transfirió información en un período de medio año.

Vistas por separado, cada pieza de datos es casi inofensiva. Pero tomadas de conjunto aportan lo que los investigadores llaman un perfil de llamadas: un claro cuadro de los hábitos y preferencias de una persona y por cierto de su vida. Este perfil revela cuándo Spitz caminaba por la calle, cuánto tomó un tren, cuándo estaba en un avión. Muestra que trabaja principalmente en Berlín y qué ciudades visitó. Muestra cuándo estaba despierto y cuándo dormía.

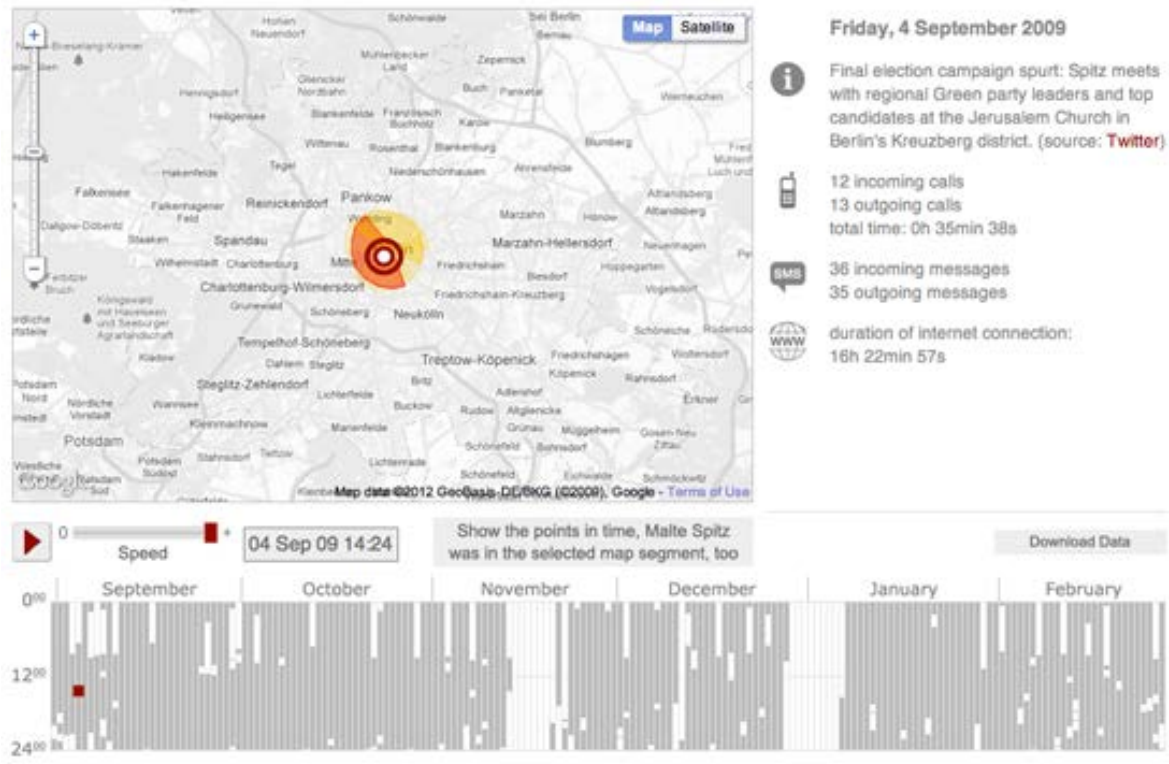


Figure 19. El teléfono que lo dice todo (Zeit Online)

El conjunto de datos de Deutsche Telekom mantenía en privado una parte del registro de los datos de Spitz, a saber, a quién llamó y quién lo llamó a él. Ese tipo de información no solo podría infringir la privacidad de mucha otra gente relacionada con él, también –aunque los números estuviesen encriptados- revelaría demasiado acerca de Spitz (pero los agentes del gobierno en el mundo real tendrían acceso a esta información).

Pedimos a Lorenz Matzat y Michael Kreil de OpenDataCity que exploraran los datos y encontraran una solución para la presentación visual. “Al principio usamos herramientas como Excel y Fusion Tables para comprender los datos. Luego comenzamos a desarrollar una interfaz del mapa que permitiera al público interactuar con los datos de un modo no lineal”, dijo Matzat. Para ilustrar hasta qué punto pueden obtenerse detalles de la vida de alguien a partir de estos datos almacenados, se le sumó información del dominio público acerca de su actividad (Twitter, entradas en blogs, información partidaria como entradas en el calendario público de su sitio en la red). Es el tipo de proceso que cualquier buen investigador usaría probablemente para hacer el perfil de una persona en observación. Junto con los gráficos del propio Zeit Online y los del equipo de investigación y desarrollo, se creó una gran interfaz para navegar: apretando el botón de play se inicia un viaje a través de la vida de Malte Spitz.

Luego de un lanzamiento muy exitoso del proyecto en Alemania, advertimos que recibíamos muchísimo tráfico de fuera de Alemania y decidimos crear una versión en inglés de la aplicación. Luego de recibir el premio Grimme Online Alemán, el proyecto recibió un premio ONA en septiembre de 2011, lo que fue la primera vez que lo recibía un sitio de noticias alemán. Todos los datos están disponibles en una planilla de cálculo de Google Docs. Lea la historia en Zeit Online.

— *Sascha Venohr, Zeit Online*

Tasas de reprobación de distintos modelos de auto en la prueba MOT

En enero de 2010 la BBC obtuvo datos sobre aprobaciones y rechazos en la prueba del Ministerio de Transporte (MOT, Ministry of Transport Test) para distintas marcas y modelos de autos. Esta es la prueba que evalúa si un auto es seguro y está en condiciones para andar por la calle; todo auto de más de 3 años tiene que pasar una prueba MOT anual.

Obtuvimos los datos bajo la ley de acceso a la Información luego de una larga batalla con VOSA, el ente del departamento de Transporte que supervisa el sistema MOT. VOSA rechazó nuestro pedido de estas cifras con el argumento de que violaría la confidencialidad comercial. Sostuvo que podría *causar daño comercial* a fabricantes de vehículos con altas tasas de rechazo. Entonces apelamos al Comisionado de información, que dictaminó que dar a conocer la información iría en favor del interés del público. Entonces VOSA entregó los datos, 18 meses después de que los pidiéramos.

Analizamos las cifras, concentrándonos en los modelos más populares y comparando autos de la misma antigüedad. Esto mostró grandes discrepancias. Por ejemplo, entre los autos de 3 años de antigüedad, 28% de los Renault Mégane no aprobaron su MOT, en contraste con solo el 11% de los Toyota Corolla. Las cifras se difundieron por televisión, radio y online.

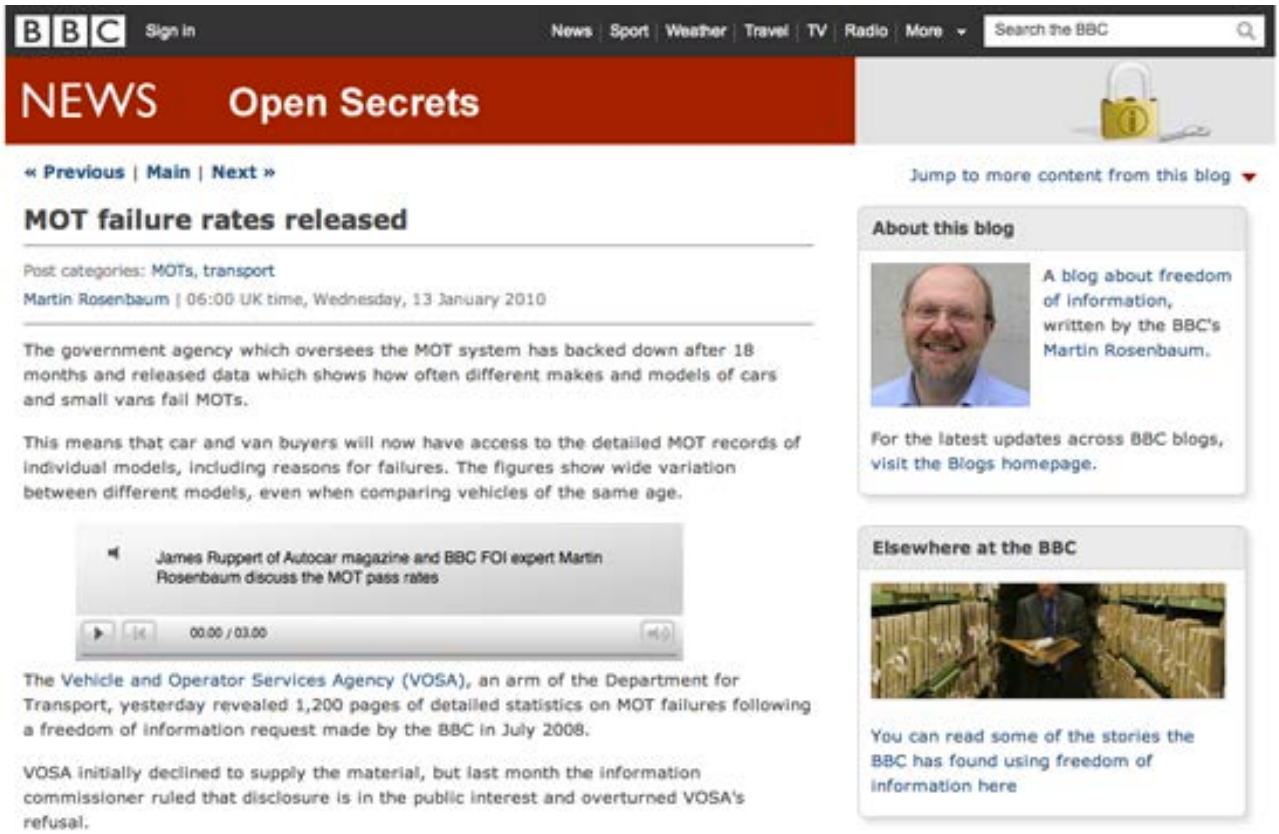


Figure 20. Difusión de las tasas de rechazo en la prueba MOT (BBC)

Nos entregaron los datos en la forma de un documento PDF de 1200 páginas, que tuvimos que convertir en planilla de cálculo para hacer el análisis. Además de informar nuestras conclusiones, publicamos la planilla de cálculo Excel (con más de 14.000 líneas de datos) en el sitio de BBC News junto con nuestra historia. Esto permitió el acceso a los datos en formato usable a todos.

El resultado fue que entonces otros usaron estos datos para sus propios análisis, que nosotros no tuvimos tiempo de hacer por el apuro de difundir la historia rápidamente (y que en algunos casos hubiera superado nuestra capacidad técnica de aquel momento). Esto incluyó el examen de las tasas de rechazo para autos de otras antigüedades, comparar los registros de fabricantes en vez de modelos individuales y crear bases de datos para buscar los resultados de modelos individuales. Agregamos vínculos a estos sitios en nuestra historia online, de modo que los lectores pudieran conocer estos trabajos.

Esto ilustra algunas de las ventajas de publicar los datos en crudo junto con una historia basada en datos. Puede haber excepciones (por ejemplo si piensa usar los datos para otras historias posteriores y quiere quedárselos mientras tanto), pero en general publicar los datos tiene varios beneficios importantes:

- Su trabajo es descubrir cosas y contarle a los ciudadanos. Si se tomó el trabajo de obtener los datos es parte de su trabajo difundirlos.

- Otras personas pueden descubrir cuestiones de interés significativo que usted no vio o simplemente detalles que les importan a ellos, aunque no le importaran lo suficiente a usted como para incluirlos en su historia.

- Otros pueden basarse en su trabajo para desarrollar un análisis más detallado, o usar distintas técnicas para presentar o visualizar las cifras, usando sus propias ideas o capacidades técnicas que pueden sondear los datos de modo productivo y de maneras alternativas.

- Es parte de incorporar la rendición de cuentas y la transparencia al proceso periodístico. Otros pueden entender sus métodos y verificar su trabajo si quieren.

— Martin Rosenbaum, BBC

Subsidios a colectivos en Argentina

Desde 2002 los subsidios para el sistema de transporte público de pasajeros en la Argentina han estado creciendo de modo exponencial, rompiendo un record cada año. Pero en 2011, luego de ganar las elecciones, el nuevo gobierno argentino anunció reducciones de los subsidios para los servicios públicos a partir del mes de diciembre de ese año. Al mismo tiempo, decidió transferir la administración de líneas locales de ómnibus y del subte al Gobierno de la Ciudad de Buenos Aires. Dado que no se ha clarificado la transferencia de subsidios a este gobierno municipal y hay falta de fondos locales para garantizar la seguridad el sistema de transporte, el Gobierno porteño rechazó esta decisión.

Mientras esto sucedía, junto con mis colegas en La Nación nos reunimos por primera vez para discutir cómo iniciar nuestra propia operación de periodismo de datos. Nuestro editor de la sección financiera sugirió que los datos sobre subsidios publicados por la secretaría de Transporte sería un buen desafío para comenzar, considerando que era muy difícil encontrarles sentido debido al formato y la terminología.

Las malas condiciones del sistema de transporte público afectan la vida de más de 5800000 pasajeros diarios. Demoras, huelgas, desperfectos de vehículos, o incluso accidentes suceden a menudo. Por tanto, decidimos analizar a dónde van los subsidios para el sistema

de transporte público en la Argentina y poner estos datos a disposición de todos los ciudadanos argentinos por medio de un “Explorador de Subsidios del Transporte”, que actualmente está en construcción.

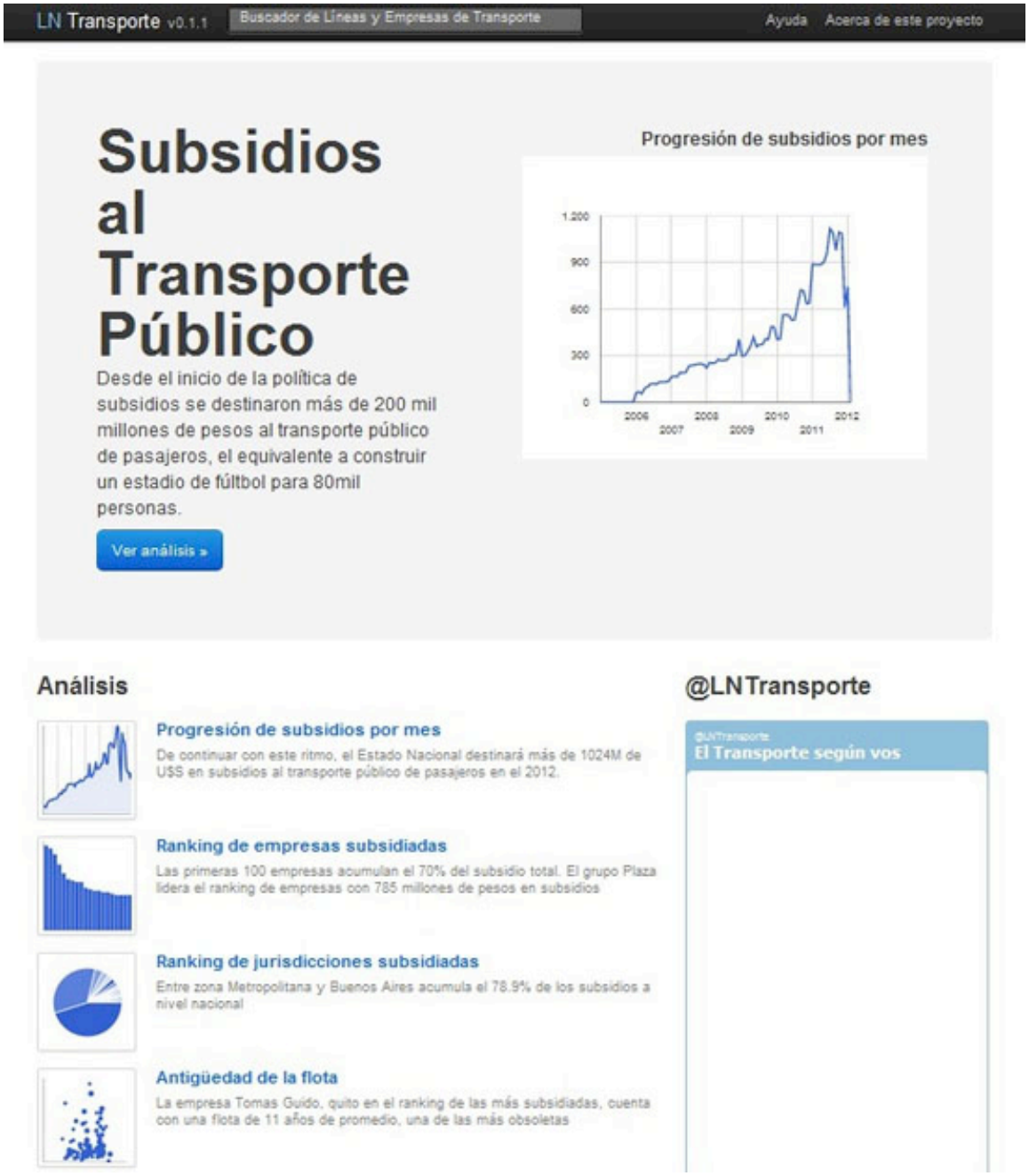


Figure 21. El explorador de subsidios al transporte (La Nación)

Comenzamos por calcular cuánto reciben cada mes las compañías de ómnibus de parte del Estado. Para hacerlo, tomamos los datos publicados en el sitio del departamento de Transporte, donde se publican desde 2006 más de 400 PDF conteniendo pagos mensuales en efectivo a más de 1300 compañías.

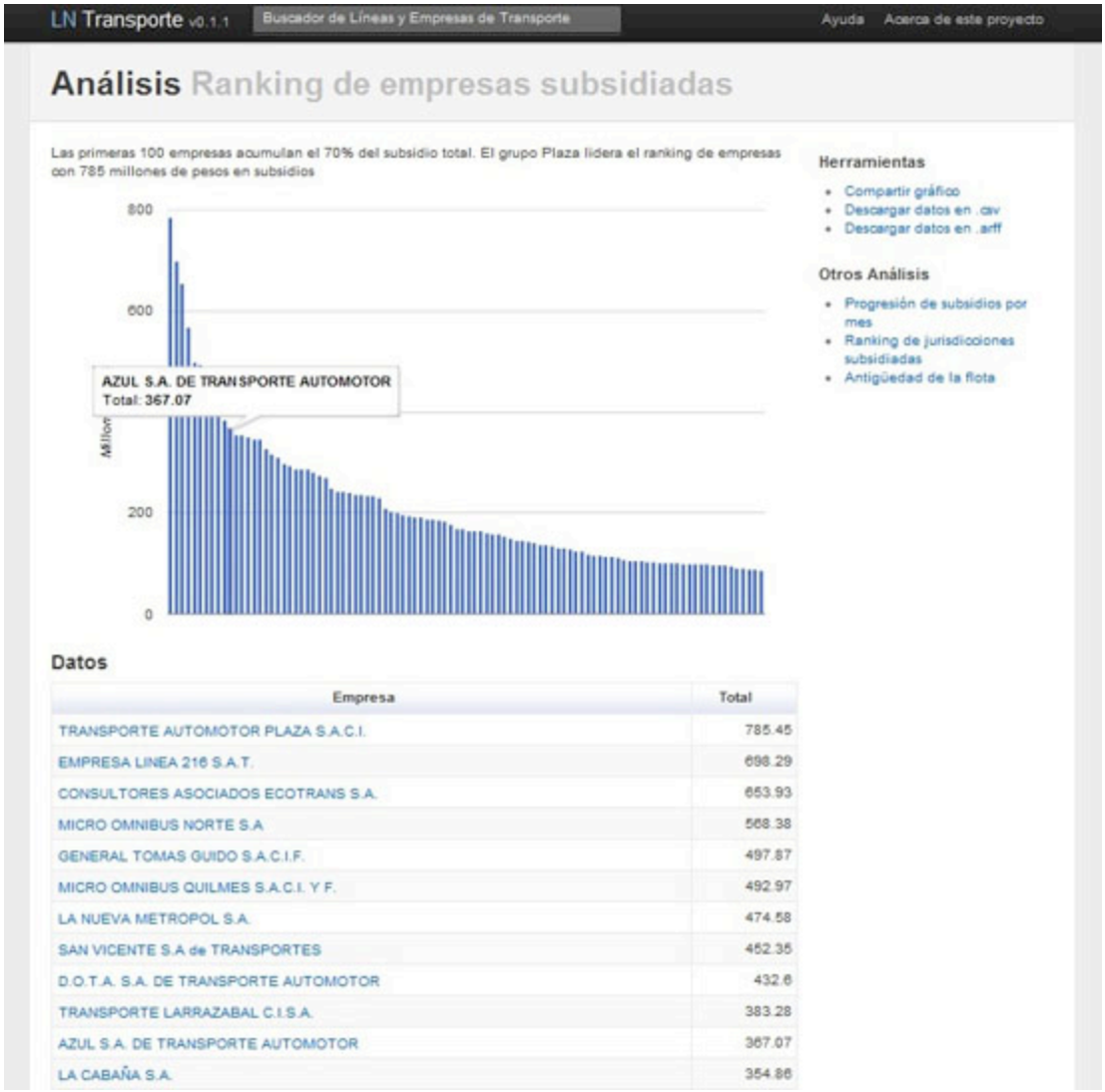


Figure 22. Ranking de empresas de transporte subsidiadas (La Nación)

Formamos equipo con un programador experimentado para desarrollar un recopilador de información de modo de automatizar la descarga regular y la conversión de estos PDF en archivos de Excel y Base de datos. Estamos usando el conjunto de datos resultante con más de 285.000 registros para nuestras investigaciones y visualizaciones, tanto en versión impresa como online. Además, estamos difundiendo estos datos en formato legible por computadora para que todo argentino pueda utilizarlos y compartirlos.

El siguiente paso fue identificar cuánto le cuesta en promedio al gobierno el mantenimiento mensual de un vehículo de transporte público. Para descubrirlo consultamos otro sitio oficial, el de la Comisión Nacional de Regulación del Transporte, responsable de la regulación del transporte en la Argentina. En este sitio encontramos una lista de compañías de ómnibus que poseen en total 9000 vehículos. Desarrollamos un normalizador para permitirnos conciliar los nombres de las compañías de transporte y hacer referencias cruzadas entre los 2 conjuntos de datos.

Para continuar, necesitábamos el número de registro de cada vehículo. En el sitio de la CNRT encontramos una lista de vehículos discriminados por línea de colectivo y compañía, con sus números de licencia. En Argentina, estos registros están compuestos de letras y cifras que se corresponden con la edad del vehículo. Por ejemplo, mi auto tiene el número de registro IDF234, y la "I" corresponde a marzo-abril 2011. Hicimos el cálculo inverso a partir de las licencias de los ómnibus propiedad de las compañías registradas, para descubrir la edad promedio de los ómnibus y mostrar cuánto dinero recibe cada compañía y finalmente comparar los montos en base a la edad promedio de sus vehículos.

En medio de este proceso, cambió misteriosamente el contenido de los PDF oficiales con los datos, aunque las URL y los nombres de los archivos no se modificaron. En algunos PDF ahora faltaban los "totales" verticales, lo que hace imposible cruzar los mismos en todo el período investigado, 2002-2011.

Llevamos este caso a un hackatón organizado por Hacks/Hackers en Boston, donde el programador Matt Perry generosamente creó lo que llamamos el "Espía de PDF". Esta aplicación ganó la categoría "más intrigante" en ese evento. El Espía de PDFs apunta a una página web llena de PDF y verifica si el contenido dentro de los PDF ha cambiado. "Nunca serán engañados nuevamente por la supuesta "transparencia del gobierno"", escribe Matt Perry.

Figure 23. Comparación de antigüedad de flotas con el monto de dinero que reciben del Estado (La Nación)

¿Quién trabajó en el proyecto?

Un equipo de 7 periodistas, programadores y un diseñador interactivo durante 13 meses.

Las capacidades que necesitamos para este proyecto fueron:

- Periodistas con conocimiento sobre cómo funcionan los subsidios para el sistema de transporte público y cuáles eran los riesgos; conocimiento del mercado de compañías de ómnibus.
- Un programador capacitado en recopilar datos de la red, su análisis, normalización y extracción de datos de PDF a planillas de cálculo Excel.
- Un especialista en estadística para el análisis de los datos y los distintos cálculos.
- Un diseñador para producir las visualizaciones interactivas de datos.

¿Qué herramientas utilizamos?

Usamos VBasic para aplicaciones, Excel Macros, Tableau Public y la Plataforma Abierta de datos Junar, así como Ruby on Rails, la API de cuadros Google, y Mysql para el Explorador de Subsidios.

El proyecto tuvo gran impacto. Hemos tenido decenas de miles de visitas y la investigación apareció en la primera plana de la edición impresa de La Nación.

El éxito de este primer proyecto de periodismo de datos nos ayudó internamente para argumentar en favor de la creación de una operación de datos que cubra periodismo de investigación y provea servicio al público. Esto resultó en Data.lanacion.com.ar, una plataforma donde publicamos datos abiertos sobre distintos tópicos de interés público en formatos procesables por computadora.

— Angélica Peralta Ramos, La Nación (Argentina)

Ciudadanos periodistas de datos

No solo las grandes redacciones pueden trabajar en historias basadas en datos. Las mismas capacidades que son útiles para los periodistas de datos también pueden ayudar a ciudadanos periodistas a acceder a datos sobre sus localidades y convertirlos en historias.

Ese fue la principal motivación para el proyecto de medios ciudadanos de Amigos de Januária, en Brasil, que recibió un subsidio (de Rising Voices, la rama de extensión de Global Voices Online y apoyo adicional de la organización Article 19. Entre septiembre y octubre de 2011, un grupo de jóvenes residentes de un pequeño pueblo localizado al norte del estado de Minas Gerais, una de las regiones más pobres de Brasil, fue capacitado en técnicas básicas de periodismo y control de presupuesto. También aprendió cómo hacer pedidos de acceso a la información y cómo obtener información pública de bases de datos oficiales en internet.

Figure 24. El proyecto de medios ciudadanos Amigos de Januária da capacidades claves a los ciudadanos para convertirlos en periodistas de datos

Januária, un pueblo de aproximadamente 65.000 residentes, también es conocido por las fallas de sus políticos locales. En 3 períodos de 4 años tuvo 7 alcaldes diferentes. Casi todos fueron removidos de sus funciones por mal desempeño en sus administraciones, incluyendo acusaciones de corrupción.

Los pequeños pueblos como Januária a menudo no atraen la atención de los medios brasileños, que tienden a concentrarse en ciudades mayores y capitales de estado. Sin embargo hay una oportunidad para que los residentes de pequeños pueblos se conviertan en aliados potenciales en el monitoreo de la administración pública, porque conocen mejor que nadie los desafíos cotidianos que enfrentan las comunidades locales. Teniendo a Internet como otro aliado importante, los residentes ahora pueden acceder mejor a datos del presupuesto y otra información local.

Luego de participar de 12 talleres, algunos de los nuevos ciudadanos periodistas de Januária comenzaron a demostrar cómo este concepto de acceder a datos públicos en pequeños pueblos puede ponerse en práctica. Por ejemplo, Soraia Amorim, una periodista ciudadana de 22 años, escribió una historia sobre una cantidad de doctores que está en la nómina municipal según datos del gobierno federal. Sin embargo, descubrió que la cifra oficial no se correspondía con la situación en el pueblo. Para escribir esta pieza, Soraia tuvo acceso a datos de salud, que están disponibles online en el sitio del SUS (Sistema Único de Saúde,un

programa federal que provee ayuda médica gratuita a la población brasileña. Según los datos de US, Januária debiera tener 71 doctores en varias especialidades de salud.

El número de doctores indicado por los datos de SUS no se correspondía con lo que Soraia sabía acerca de los doctores de la zona: los residentes siempre se quejaban de la falta de doctores y algunos pacientes tenían que viajar a pueblos vecinos para ver un profesional. Más tarde entrevistó a una mujer que había estado recientemente en un accidente de motocicleta, y no pudo conseguir ayuda médica en el hospital de Januária porque no había ningún doctor disponible. También habló con el secretario de Salud del pueblo, que reconoció que había menos doctores en el pueblo de lo que indicaba la cifra publicada por el SUS.

Estas conclusiones iniciales plantean muchos interrogantes respecto de los motivos de estas diferencias entre la información oficial publicada online, y la realidad del pueblo. Uno de ellos es que los datos federales pueden estar equivocados, lo que significaría que hay una importante falta de información de salud en Brasil. Otra posibilidad puede ser que Januária está reportando incorrectamente la información al SUS. Ambas posibilidades debieran llevar a una investigación más profunda para encontrar la respuesta definitiva. Sin embargo, la historia de Soria es una parte importante de esta cadena porque destaca una inconsistencia y puede también alentar a otros a analizar esta cuestión con más detenimiento.

“Yo antes vivía en el campo y terminé la secundaria con mucha dificultad”, dice Soraia. “Cuando la gente me preguntaba qué quería hacer de mi vida, siempre dije que quería ser periodista. Pero imaginaba que era casi imposible debido al mundo en el que vivía”. Luego de participar en la capacitación de Amigos de Januária, Soraia cree que el acceso a datos es una herramienta importante para cambiar la realidad de su pueblo. “Me siento capaz de ayudar a cambiar mi pueblo, mi país, el mundo”, agrega.

Otro periodista ciudadano del proyecto es Alyson Montiériton, de 20 años, que también usó datos para un artículo. Fue durante la primera clase del proyecto, cuando los periodistas ciudadanos caminaron por la ciudad en busca de temas que pudieran convertirse en historias, que Alysson decidió escribir sobre un semáforo roto ubicado en una intersección muy importante, que había permanecido en ese estado desde el comienzo del año. Luego de aprender a conseguir datos en Internet, buscó la cantidad de vehículos que existe en el pueblo y la cantidad de impuestos que pagan los dueños de autos. Escribió:

La situación en Januária empeora debido al alto número de vehículos en el pueblo. Según el IBGE (el instituto de investigaciones estadísticas más importante de Brasil), Januária tenía 13771 vehículos (entre ellos 7979 motos) en 2010... Los residentes del pueblo creen que la demora en arreglar el semáforo no es resultado de la falta de recursos. Según el Secretario

del Tesoro del estado de Minas Gerais, el pueblo recibió 470.000 reales en impuestos sobre vehículos en 2010.

Teniendo acceso a los datos, Alysson pudo mostrar que Januária tiene muchos vehículos (casi 1 por cada 5 residentes) y que un semáforo roto podía poner en peligro a mucha gente. Lo que es más, pudo decirle a su público la cantidad de fondos recibidos por el pueblo de impuestos pagados por dueños de vehículos y basado en ello cuestionar si este dinero no sería suficiente para reparar el semáforo garantizando condiciones de seguridad a conductores y peatones.

Si bien las 2 historias escritas por Soraia y Alysson son muy simples, muestran que los datos pueden ser usados por cronistas ciudadanos. No se necesita estar en una gran redacción con muchos especialistas para usar datos en sus artículos. Luego de 12 talleres, Soraia y Alysson, ninguno de los cuales ha estudiado periodismo, pudieron trabajar en historias basadas en datos y escribir piezas interesantes sobre su situación local. Además sus artículos muestran que los datos mismos pueden ser útiles incluso a escala pequeña. Dicho de otro modo también hay información valiosa en conjuntos de datos y tablas pequeñas, no solo en bases de datos inmensas.

— Amanda Rossi, Friends of Januária

El gran cuadro de resultados electorales

Los resultados electorales ofrecen grandes oportunidades para contar historias de forma visual para cualquier organización de noticias, pero durante años esta fue para nosotros una oportunidad perdida. En 2008 con los diseñadores gráficos nos propusimos cambiar eso.

Queríamos encontrar una manera de desplegar resultados que contara una historia y que no se viera como simplemente una mezcla de cifras en una tabla o mapa. En anteriores elecciones eso es exactamente lo que hicimos.

No es que una gran bolsa de números –lo que llamo el “modelo CNN” de tablas, tablas y más tablas- tenga algo de malo necesariamente. Funciona porque da al lector lo que quiere saber: quién ganó.

Y es peligroso meterse con algo que no está roto. Al hacer algo radicalmente diferente y alejarnos de lo que la gente espera podríamos haber hecho más confusas las cosas.

Por fin, fue Shan Carter de la mesa de diseño el que dio la respuesta adecuada, lo que terminamos llamando el “gran cuadro”. Cuando vi los bosquejos por primera vez, fue literalmente una cachetada a la cara.

Era exactamente lo que había que hacer.

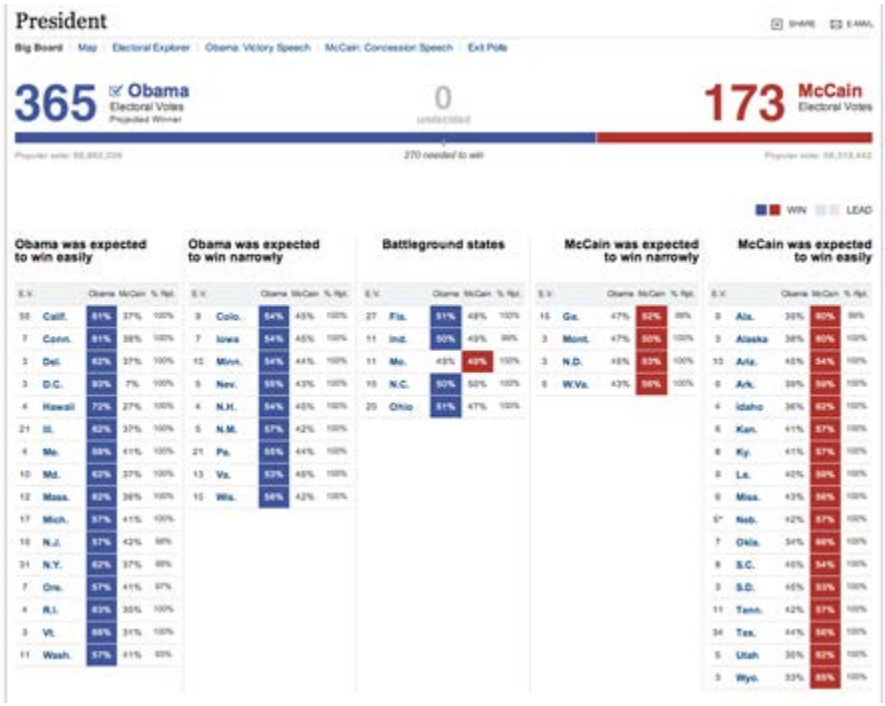


Figure 25. El gran cuadro de resultados electorales (New York Times)

¿Qué es lo que hace de esto una gran pieza de periodismo visual? Por empezar, la mirada del lector es atraída inmediatamente a la gran barra que muestra los votos del colegio electoral arriba, lo que en el contexto periodístico podríamos llamar el _copete._Le dice al lector exactamente lo que quiere saber y lo hace de modo rápido, simple y sin ruido visual.

A continuación el lector es atraído al agrupamiento de estados en 5 columnas más abajo, organizado de acuerdo a la probabilidad que el Times asignaba a que un estado dado se inclinara por uno u otro candidato. En la columna del medio está lo que en el contexto periodístico podríamos llamar nuestro *gráfico central*, donde explicamos por qué Obama ganó. El interactivo lo deja totalmente claro: Obama se quedó con los estados que se preveía y 4 de los 5 más disputados.

Para mi esta construcción en 5 columnas es un ejemplo de cómo el periodismo visual difiere de otras formas de diseño. Idealmente una gran pieza de periodismo visual será tanto hermosa como informativa. Pero cuando tiene que decidir entre la historia y la estética, el periodista debe volcarse para el lado de la historia. Aunque este diseño puede no ser la manera en que un diseñador puro podría preferir presentar los datos, presenta la historia muy, pero muy bien.

Y finalmente, como cualquier buen recurso interactivo de la red, este invita al lector a profundizar más. Hay detalles como porcentajes de votos, estado por estado, informes de la cantidad de votos electorales y porcentajes deliberadamente colocados en un segundo plano para no competir con lo principal de la historia.

Todo esto hace que el “gran cuadro” sea una gran pieza de periodismo visual que hace un mapa casi perfecto siguiendo el esquema probado de la pirámide invertida.

— Aron Pilhofer, New York Times

Consulta sobre el precio del agua

Desde marzo de 2011, la información sobre el agua de la canilla en toda Francia se obtiene a través de un experimento de consulta a la población. En solo 4 meses, mas de 5000 personas hartas del control corporativo del mercado de agua se tomaron el tiempo de buscar su factura, escanearla y cargarla en el proyecto Prix de l’Eau (“precio del agua”); El resultado es una investigación sin precedentes que reunió técnicos, ONG y medios tradicionales para mejorar la transparencia en torno de proyectos de agua.

Figure 26. El precio del agua (Fundación France Liberté)

El mercado de servicios de agua consiste en más de 10.000 clientes (ciudades que compran agua para distribuir a sus contribuyentes) y sólo un puñado de compañías. La relación de fuerzas en este oligopolio está distorsionado en favor de las corporaciones, que en algunos casos cobran precios distintos a pueblos vecinos.

La ONG francesa France Libertés ha estado tratando con cuestiones de agua en todo el mundo en los últimos 25 años. Ahora se concentra en mejorar la transparencia del mercado francés y en dar poder a ciudadanos y alcaldes que negocian acuerdos de servicios de agua. El gobierno francés decidió enfrentar el problema hace 2 años con un censo nacional del

precio y la calidad el agua. Hasta ahora sólo se ha recogido el 3% de los datos. Para ir más rápido, France Libertés quería involucrar ciudadanos directamente.

Junto con el equipo OWNI diseñé una interfaz para la consulta en la que los usuarios estudiaban su factura de agua e ingresaban el precio que pagaban por el agua de la canilla en prixdeleau.fr/. En los últimos 4 meses, 8500 se inscribieron y sean cargado y validado más de 5000 facturas.

Si bien esto no permite una evaluación perfecta de la situación del mercado, le mostró a los interesados, tales como los entes de supervisión del agua, que había una preocupación genuina, a nivel popular, por el precio del agua corriente. Al principio eran escépticos respecto de la transparencia, pero cambiaron de idea en el curso de la operación, sumándose progresivamente a France Libertés en su lucha contra la opacidad y la mala praxis corporativa. ¿Qué pueden aprender de esto las organizaciones de medios?

Asociarse con ONG

> Las ONG necesitan gran cantidad de datos para diseñar trabajos de política. Estarán más dispuestas a pagar por una operación e recolección de datos que un ejecutivo de diario.

Los usuarios pueden aportar datos en crudo

> Las consultas funcionan del mejor modo cuando los usuarios cumplen una tarea de recolección de datos o refinado de datos.

Pedir la fuente de la información

> Evaluamos si pedir a los usuarios una copia de la factura original, pensando que disuadiría a algunos de ellos (especialmente dado que nuestro público era mayor en promedio). Si bien pudo haber sido una traba para algunos, aumentó la credibilidad de los datos.

Crear un mecanismo de validación

> Diseñamos un sistema de puntaje y un mecanismo de revisión por los pares para controlar los aportes de los usuarios. Esto demostró ser demasiado engorroso para los usuarios, que tenían pocos incentivos para hacer visitas repetidas al sitio. Pero fue utilizado por el equipo de France Libertés, cuyos empleados, alrededor de 10, se sintieron motivados por el sistema de puntaje.

Mantenerlo simple

> Creamos un mecanismo de correo automatizado de modo que los usuarios pudieran presentar un pedido de acceso a la información respecto de precios del agua con solo unos pocos clics. Aunque innovador y bien diseñado, este recurso no generó un número sustancial de pedidos (solo 100 fueron enviados).

Defina su público

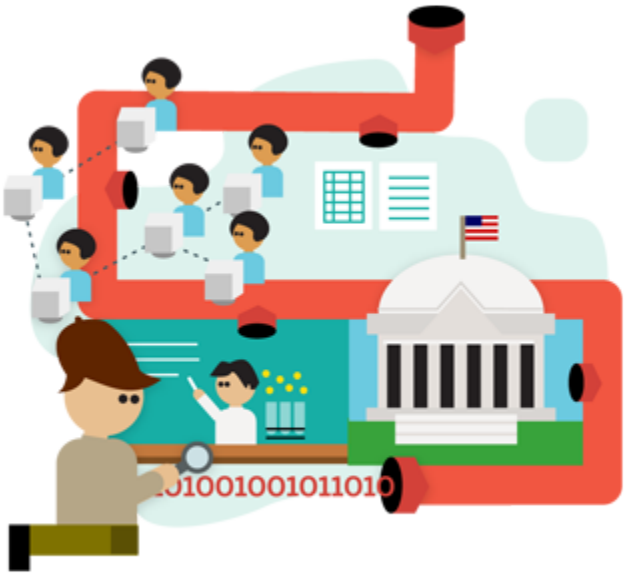
France Libertés se asoció con la revista dedicada a los derechos de los consumidores *60 Millions de Consommateurs*, que lograron una gran participación de su comunidad. Fue la unión prefecta para esta operación.

Elija cuidadosamente sus indicadores claves de desempeño

El proyecto tuvo solo 45.000 visitantes en 4 meses, equivalente a 15 minutos de tráfico en nytimes.com. Lo importante es que 1 de cada 5 se inscribió y 1 de cada 10 se tomó el tiempo de escanear y subir su factura.

— *Nicolas Kayser-Bril, Journalism++*

Obtener datos



Así que está listo para comenzar con su primer proyecto de periodismo de datos. ¿Y ahora qué? Primero necesita algunos datos. Esta sección analiza de dónde puede obtenerlos. Aquí aprenderemos cómo encontrar datos en la red, cómo pedirlos usando las leyes de acceso a la información, cómo usar el "screen scraping" (peinado de pantalla) para recoger datos de fuentes no estructuradas, y cómo usar la “colaboración del público” (crowdsourcing) para obtener sus propios conjuntos de datos de sus lectores. Finalmente analizamos lo que dicen las leyes respecto de la re-edición de conjuntos de datos, y cómo usar herramientas legales simples para permitir a otros reutilizar sus datos.

Qué contiene este capítulo?

Una guía para trabajos de campo de 5 minutos

¿Busca datos sobre un tópico o cuestión particular? ¿No está seguro de qué es lo que hay o dónde encontrarlo? ¿No sabe por dónde empezar? En esta sección analizamos cómo comenzar la búsqueda de fuentes de datos públicos en la red.

Ajustar la búsqueda

Aunque pueden no ser siempre fáciles de encontrar, muchas bases de datos en la red están indexadas por motores de búsqueda, fuera ello o no la intención del editor. Unos cuantos consejos:

- Cuando busque datos asegúrese de incluir tanto términos de búsqueda relacionados con el contenido de los datos que trata de encontrar, como algo de información sobre el formato o la fuente en la que prevé encontrarlos. Google y otros motores de búsqueda le permiten buscar por tipo de archivo. Por ejemplo, puede buscar solo planillas de cálculo (agregando a su búsqueda "filetype:XLS filetype:CSV"), datos geográficos ("filetype:shp"), o extractos de bases de datos ("filetype:MDB, filetype:SQL, filetype:DB). Si así lo desea incluso puede buscar PDF ("filetype:pdf").
- También puede buscar con una parte de una URL. Hacer una búsqueda en Google de "inurl:donwloads filetype:xls" o "inurl:descargas filetype:xls" significa buscar todos los archivos Excel que tienen "downloads" o "descargas" en su dirección de la red (si encuentra una sola descarga, a menudo vale la pena simplemente verificar qué otros resultados existen para la misma carpeta en el servidor de la red). También puede limitar su búsqueda solo a aquellos resultados dentro de un solo nombre de dominio, buscando "site:agency.gov", por ejemplo.
- Otro truco popular es no buscar determinado contenido directamente, sino lugares donde puede haber datos disponibles en gran cantidad. Por ejemplo "sitio:ente.gov Directory Listing" puede darle algunos listados generados por el servidor de la red con fácil acceso a archivos en bruto, mientas que "sitio:ente.gov Database Download" buscará listados creados intencionalmente.

Ir Directo a la fuente

El primer truco que uso para obtener datos que están en manos de un ente público es tratar de ir directo a quien tiene los datos, no la persona de relaciones públicas, ni a través de un pedido de acceso a la información (PAI). Podría por supuesto hacer un PAI o un pedido de registros públicos, pero eso hace que los engranajes comiencen a girar con lentitud. Es probable que reciba la respuesta de que los datos no están en el formato que solicité o (tal como ha sucedido en algunos casos) que el ente oficial usa un software propio y no puede extraer los datos en el formato que requerí. Pero si empiezo por llegar a la persona que maneja los datos para esa organización, puedo hacer preguntas respecto de qué datos tienen sobre el tema y cómo los guardan. Puedo conocer el formato, hablar en el lenguaje de los datos y descubrir lo que necesito saber para pedir los datos y tener éxito. ¿Las barreras que se enfrentan en este caso? A menudo es difícil llegar a estas personas. El encargado/a de Información Pública (EIP) va a querer que trate directamente con él/ella. En esos casos he

descubierto que lo mejor es tratar de organizar una llamada colectiva o, aún mejor, una reunión en persona con el/la EIP, el gurú de datos, y yo. Y lo puedo organizar de un modo que les resulte difícil decir que no. “No quiero darles trabajo”, digo. “No quiero crear una carga innecesaria ni hacer un pedido demasiado amplio, de modo que una reunión me ayudará a entender exactamente lo que tienen y cómo pedir exactamente lo que necesito”.

Si este método no funciona, la alternativa es hacer un pedido de conocer cómo está organizado su archivo y su diccionario de datos. Entonces pido los datos efectivamente. A veces pregunto también cómo guardan y qué sistema usan. De ese modo puedo investigar de qué modo exportar los datos antes de escribir mi pedido.

Por último, mi mejor historia de éxito es de cuando estaba trabajando en un pequeño diario en Montana. Necesitaba algunos datos de países, me dijeron que no podían exportarse de la computadora central. Investigué un poco y ofrecí ir a ayudarlos. Trabajé con la persona de datos, hicimos un pequeño guión y copiamos los datos a un disquete (esto fue hace mucho tiempo). Tenía mis datos y el condado ahora estaba en condiciones de proveer los datos a cualquiera que los pidiera. No querían que eso sucediera, pero a veces ellos también necesitaban extraer datos y no entendían su sistema por completo, de modo que nos ayudamos entre todos.

— Cheryl Philips, The Seattle Times

Explore sitios y servicios de datos

En los últimos años han aparecido una cantidad de portales y centros de datos dedicados y otros sitios de datos en la red. Son buenos lugares para llegar a conocer los tipos de datos que hay. Para empezar podría ver:

Figure 1. datacatalogs.org (Open Knowledge Foundation)

Portales oficiales de datos

La disposición del gobierno a entregar ciertos conjuntos de datos varía de país en país. Un número creciente de países está lanzando portales de datos (inspirados por el data.gov de EE.UU. y el data.gov.uk del R.U.) para promover la reutilización civil y comercial de información oficial. Se puede encontrar un índice global actualizado de tales sitios en datacatalogs.org/. Otro sitio práctico es el Guardian World Government Data, un meta-motor de búsquedas que incluye muchos catálogos de datos gubernamentales internacionales.

El Data Hub

Un recurso comunitario manejado por la Open Knowledge Foundation que facilita buscar, compartir y re-utilizar fuentes de datos abiertamente disponibles, especialmente de maneras automatizadas.

Scraperwiki

Una herramienta online para hacer que el proceso de extraer “datos útiles sea más fácil de modo que puedan ser utilizados en otras aplicaciones o que periodistas e investigadores puedan *scrapear* en ellos”. La mayoría de los "scrapers" y sus bases de datos son públicos y pueden ser reutilizados.

Portales de datos del Banco Mundial y las Naciones Unidas::

Estos servicios ofrecen indicadores de alto nivel para todos los países y en muchos casos cubren muchos años.

Buzzdata, Infochimps y DataMarket:

Sitios nuevos que apuntan a crear comunidades dedicadas a compartir datos y su reventa.

DataCouch

Un lugar donde subir, refinar, compartir y visualizar sus datos.

Freebase

Una interesante subsidiaria de Google que ofrece “un gráfico de entidades de gente, lugares y cosas, creado por una comunidad amante de la información abierta”.

Datos de investigación

Hay compiladores nacionales y disciplinarios de datos de investigación como el UK Data Archive. Si bien hay mucha información gratuita en el punto de acceso, también hay muchos datos que requieren una suscripción, o que no pueden ser reutilizados o redistribuidos sin obtener autorización.

Obtener datos de archivos de papel

Justo después de la difusión por WikiLeaks de documentos militares de EE.UU. sobre Afganistán e Irak, decidimos adaptar el concepto para conmemorar el 50 aniversario de la Guerra de Argelia publicando los Diarios de la Guerra de Argelia. Nos propusimos obtener y

digitalizar los archivos del Ejército Francés en Argelia. Estos están disponibles en el archivo del ministerio de Guerra en París, aunque en formato impreso. Enviamos a nuestros periodistas y estudiantes a tomar fotografías de los documentos. Tratamos de escanearlos usando un scanner Canon P-150 portátil, pero no funcionó principalmente porque gran parte de los archivos están abrochados.

Por fin se recogieron alrededor de 10000 páginas en pocas semanas. Las pasamos por un software de reconocimiento de texto (ABBYY FineReader) que produjo resultados pobres. Lo que es más, el ministerio arbitrariamente negó acceso a las cajas más interesantes de archivos. Por encima de todo, el ministerio prohíbe reeditar documentos que pueden ser fotografiados libremente en el lugar, por lo que decidimos que no se justificaba el riesgo y el proyecto quedó en suspenso.

— *Nicolas Kayser-Bril, Journalism++*

Pregunte en un foro

Busque respuestas existentes o haga una pregunta en Get The Data o Quora. GetTheData es un sitio de preguntas y respuestas donde puede hacer sus preguntas relacionadas con datos, incluyendo donde encontrar datos relacionados con un asunto particular, cómo interrogar o encontrar una determinada fuente de datos, qué herramientas usar para explorar de modo visual, como expurgar datos, o ponerlos en un formato con el que pueda trabajar.

Pregunte en una lista de correo

Las listas de correo aprovechan la sabiduría de una comunidad entera sobre un tópico particular. Para los periodistas de datos, la Data-Driven Journalism List y la NICAR-L son excelentes puntos de partida. Ambas listas están pobladas de periodistas de datos y expertos en Periodismo Asistido por Computadora (Computer-Assisted Reporting – CAR) que trabajan en todo tipo de proyectos. Es posible que alguien haya hecho una historia como la suya y puede tener una idea de por dónde empezar, si es que no un vínculo directo con los datos que busca. También podría probar con Project Wombat; (“una lista de discusión para preguntas de referencia difíciles”), las muchas listas de correo de la Open Knowledge Foundation, listas de correo en the Info, o buscar listas de correo sobre el tópico o en la región que está interesado.

Súmese a Hacks/Hackers

Hacks/Hackers es una organización periodística internacional de base en rápida expansión con docenas de secciones y miles de miembros en 4 continentes. Su misión es crear una red

de periodistas (“Hacks”) y tecnólogos (Hackers”) que reflexionan sobre el futuro de las noticias y la información. Con una red tan amplia, tiene grandes probabilidades de encontrar a alguien que sepa dónde encontrar lo que busca.

Pregunte a un experto

Profesores, empleados públicos y gente de los distintos sectores a menudo saben dónde buscar. Llámelos. Mándeles un correo electrónico. Abórdelos en eventos. Aparézcase en su oficina. Pregunte amablemente. “Estoy escribiendo una historia sobre X. ¿Dónde encuentro esto? ¿Sabe quién tiene esto?”

Conozca la TI (Tecnología Informática) de los entes oficiales

A menudo ayuda entender el contexto técnico y administrativo en el que los entes oficiales tienen su información cuando se quiere acceder a datos. Se trate de CORDIS, COINS o THOMAS, las grandes bases de datos a menudo resultan más útiles cuando uno conoce algo del objetivo con el que se crearon.

Encuentre los cuadros organizativos de los entes oficiales y busque departamentos/unidades con una función que los atraviese (por ejemplo, informaciones, servicios TI), luego explore sus sitios en la red. Muchos datos se archivan en distintos departamentos y mientras que para uno de ellos la base de datos que le interesa puede ser su tesoro, otro puede dársela sin problemas.

Busque infografías dinámicas de sitios oficiales. Estas a menudo se basan en fuentes de datos estructurados/API que pueden ser usadas de modo independiente (por ejemplo, aplicaciones que rastrean vuelos, aplicaciones Java que pronostican el clima).

Investigar registros de llamadas telefónicas

Hace pocos meses quise analizar los registros de llamadas telefónicas del gobernador de Texas, Rick Perry (por entonces candidato presidencial). Fue el resultado de un pedido, largamente esperado, de registros públicos estaduales. Los datos vinieron esencialmente en el formato de más de 120 páginas de documentos en calidad de fax. Era un esfuerzo que requería ingresar datos y expurgarlos, seguido del uso de una aplicación que permitiera buscar en la guía los titulares de los teléfonos con los que se había comunicado el gobernador.

Combinando nombres con datos electorales estaduales y federales, descubrimos que Perry tomó contacto con donantes a su campaña y con súper comités de acción política (los

llamados super PAC, que supuestamente no deben organizar la recolección de fondos) desde teléfonos de oficinas públicas estaduales, práctica mal vista y que planteó interrogantes sobre los vínculos entre él y un “super PAC” que trabaja para él.

— *Jack Gillum, Associated Press*

Busque nuevamente

Cuando sepa más sobre lo que está buscando, vuelva a buscar usando frases y conjuntos de palabras improbables que descubrió desde la última vez. ¡Quizá tenga más suerte con los motores de búsqueda!

Escriba un pedido de acceso a la información

Si usted cree que un ente oficial tiene los datos que necesita, un Pedido de Acceso a Información puede ser su mejor herramienta. Vea la siguiente sección para más información respecto de cómo presentarlo.

— *Brian Boyer (Chicago Tribune), John Keefe (WNYC), Friedrich Lindenberg (Open Knowledge Foundation), Jane Park (Creative Commons), Chrys Wu (Hacks/Hackers)*

Cuando falla la ley

Luego de leer un artículo académico que explica que publicar el resultado de inspecciones de higiene en restaurantes redujo la cantidad de enfermedades relacionadas con alimentos en Los Ángeles, pedí a los servicios de higiene parisinos la lista de inspecciones. Siguiendo el procedimiento establecido por la ley de Acceso a la Información francesa, esperé 30 días su negativa a contestar, entonces fui a la Comisión de Acceso a los Datos públicos (CADA en francés), que determina la legitimidad de los pedidos de acceso a información. CADA apoyó mi pedido y ordenó a la administración entregar los datos. La administración a continuación pidió dos meses más y CADA lo aceptó. Dos meses más tarde la administración aún no había hecho nada.

Traté de conseguir el apoyo de defensores del libre acceso a la información famosos (y con muchos recursos) para presentar una demanda legal (lo que hubiera costado € 5000 y se hubiera ganado sin duda con el apoyo de CADA), pero temían complicar sus relaciones con los programas de datos abiertos oficiales. Este ejemplo es uno entre muchos en los que la administración francesa simplemente ignora la ley y las iniciativas oficiales no hacen nada para apoyar pedidos de datos de periodistas comunes.

— *Nicolas Kayser-Bril, Journalism++*

Su Derecho a la Información

Antes de hacer un pedido de acceso a información, debiera verificar si los datos que está buscando ya están disponibles o si otros ya los han pedido. El capítulo anterior tiene algunas sugerencias respecto de dónde puede averiguar. Si ha estado mirando y aún no pudo conseguir los datos que necesita, entonces puede querer presentar un pedido formal. Algunos consejos que pueden ayudar a hacer más efectivo su pedido.

Planifique anticipadamente para ahorrar tiempo

Piense en presentar un pedido formal de acceso cuando se proponga buscar información. Es mejor no esperar hasta haber agotado todas las demás posibilidades. Ahorrará tiempo presentado un pedido al comienzo de su investigación y desarrollando otras investigaciones paralelamente. Esté preparado para las demoras: a veces los entes públicos tardan en procesar pedidos, por lo que es mejor prever esto.

Verifique las normas respecto de aranceles

Antes de comenzar a presentar un pedido, verifique las normas respecto de aranceles para presentar pedidos o recibir información. De ese modo, si un funcionario público de pronto le pide dinero, sabrá cuáles son sus derechos. Puede pedir documentos electrónicos para evitar costos de copiado y correo, mencione en su pedido que prefiere tener la información en formato electrónico. De ese modo evitará pagar un arancel, a menos por supuesto que la información no esté disponible electrónicamente, aunque en estos tiempos por lo general es posible escanear documentos que no están digitalizados aún y luego enviarlos como agregado por correo electrónico.

Conozca sus derechos

Sepa cuáles son sus derechos antes de comenzar, de modo de saber donde está parado y qué cosas están obligadas a hacer las autoridades y qué cosas no. Por ejemplo, la mayoría de las leyes de libre acceso a información establecen un plazo para que las autoridades respondan. Globalmente, en la mayoría de las leyes los plazos varían de unos pocos días a un mes. Asegúrese de conocer el plazo antes de comenzar y anote la fecha en la que presenta su pedido.

Los entes oficiales no están obligados a procesar los datos para usted, pero debieran darle todos los datos que tienen, y si son datos que debieran tener para cumplir con sus obligaciones legales, por cierto que debieran entregárselos.

Diga que conoce sus derechos

Habitualmente no se requiere que usted mencione las leyes de acceso a información o de libertad de información, pero esto se recomienda porque muestra que conoce sus derechos y esto probablemente promueva una respuesta acorde con el derecho

vigente. Señalamos que en el caso de pedidos a la UE, es importante mencionar que es un pedido de acceso a documentos y es mejor mencionar específicamente la Norma 1049/2001.

Hágalo simple

En todos los países es mejor comenzar con un simple pedido de información y luego agregar más preguntas cuando obtiene la información inicial. De ese modo no corre el riesgo de que el ente público pida extensión del plazo por tratarse de un "pedido complejo".

Concentre su pedido

Un pedido de información que solo está en manos de una parte de un ente público probablemente tenga respuesta más rápida que un pedido que requiere una búsqueda en todo un ente. Un pedido que involucra que el ente consulte a terceros (p.ej., una empresa privada que aportó la información, otro gobierno que se ve afectado por la misma) puede llevar un tiempo particularmente prolongado. Sea persistente.

Piense que hay dentro del archivo

Intente averiguar qué datos se recogen. Por ejemplo, si recibe una copia en blanco del formulario que llena la policía después de accidentes de tráfico, puede ver qué información toman en cuenta y cual no respecto de choques de autos.

Sea específico

Antes de presentar su pedido piense: ¿es ambiguo en algún sentido? Esto es especialmente importante si piensa comparar datos de distintos entes públicos. Por ejemplo, si pide cifras de los *últimos 3 años*, algunos entes le enviarán información de los últimos 3 años calendario y otros de los 3 últimos años financieros, los que no podrá comparar directamente. Si decide ocultar su verdadero pedido en otro más general, entonces debe hacer su pedido lo suficientemente amplio como para que abarque la información que quiere pero no tanto como para resultar poco claro o como para desalentar a las autoridades a responder. Los pedidos específicos y claros tienden a tener respuestas más celeras y mejores.

Presente múltiples pedidos

Si no está seguro donde presentar su pedido, nada le impide presentar su pedido a 2, 3 o más entes al mismo tiempo. En algunos casos, los varios entes le darán distintas respuestas, pero esto en realidad le puede ser de ayuda en cuanto a darle un cuadro más completo de la información disponible en la materia que investiga.

Presente pedidos internacionales

Cada vez hay más posibilidades de presentar pedidos por vía electrónica, por lo que no importa donde vive. Alternativamente, si no vive en el país en el que quiere presentar su pedido, puede en algunos casos enviar el pedido a la embajada y desde

allí deben transferir el pedido al ente público competente. Tendrá que verificar en la embajada correspondiente si están en condiciones de hacer esto: a veces el personal de la embajada no está capacitado en la cuestión del derecho a la información y si este parece ser el caso, es más seguro presentar le pedido directamente al ente público correspondiente.

Haga una prueba

Si piensa mandar el mismo pedido a muchos entes públicos, empiece por enviar un primer texto del pedido a unos pocos entes como ejercicio piloto. Esto le mostrará si está usando la terminología adecuada para obtener el material que quiere y si es factible que contesten sus preguntas, de modo de poder revisar el pedido si fuera necesario antes de enviarlo a todos los destinatarios.

Anticipe las excepciones

Si cree que pueden aplicarse excepciones a su pedido entonces, cuando prepare sus preguntas, separe las preguntas relativas a información potencialmente delicada del resto de la información que el sentido común diría que no tiene porque ser motivo de una excepción. Luego divida sus preguntas en 2 y presente los 2 pedidos por separado.

Pida acceso a los archivos

Si vive cerca del lugar donde se guarda la información (por ej., en la capital en la que se guardan los documentos), también puede pedir inspeccionar los documentos originales. Esto puede ser de ayuda en la investigación de información que puede estar contenida en una gran cantidad de documentos que le gustaría ver. Tal inspección debiera ser gratuita y debe poder realizarse en un momento que sea razonable y conveniente para usted.

¡Guarde registro!

Haga su pedido por escrito y guarde una copia o un archivo de modo que en el futuro pueda demostrar que envío su pedido, en caso de tener que apelar por falta de respuesta. Esto también le da evidencias de haber presentado el pedido si piensa hacer un artículo sobre el tema.

Hágalo público

Acelere las respuestas haciendo público que presentó un pedido: si escribe o transmite la información de que se ha presentado el pedido puede crear presión sobre la institución pública para que procese y responda al pedido. Puede actualizar la información cuando reciba respuesta a su pedido si pasa el plazo y no hay respuesta, puede transformar esto en una noticia también. Hacer esto tiene el beneficio adicional de educar al público respecto del derecho de acceso a la información y cómo funciona en la práctica.

Note También hay varios servicios excelentes que puede usar para hacer público su pedido y

toda respuesta subsecuente, poniéndolas a disposición del público en la red, tales como ¿Qué saben? para entes públicos en el RU, Frag den Staat para entes públicos alemanes, y Ask the EU) para instituciones de la UE. El proyecto Alaveteli está ayudando a crear servicios similares en docenas de países en todo el mundo.

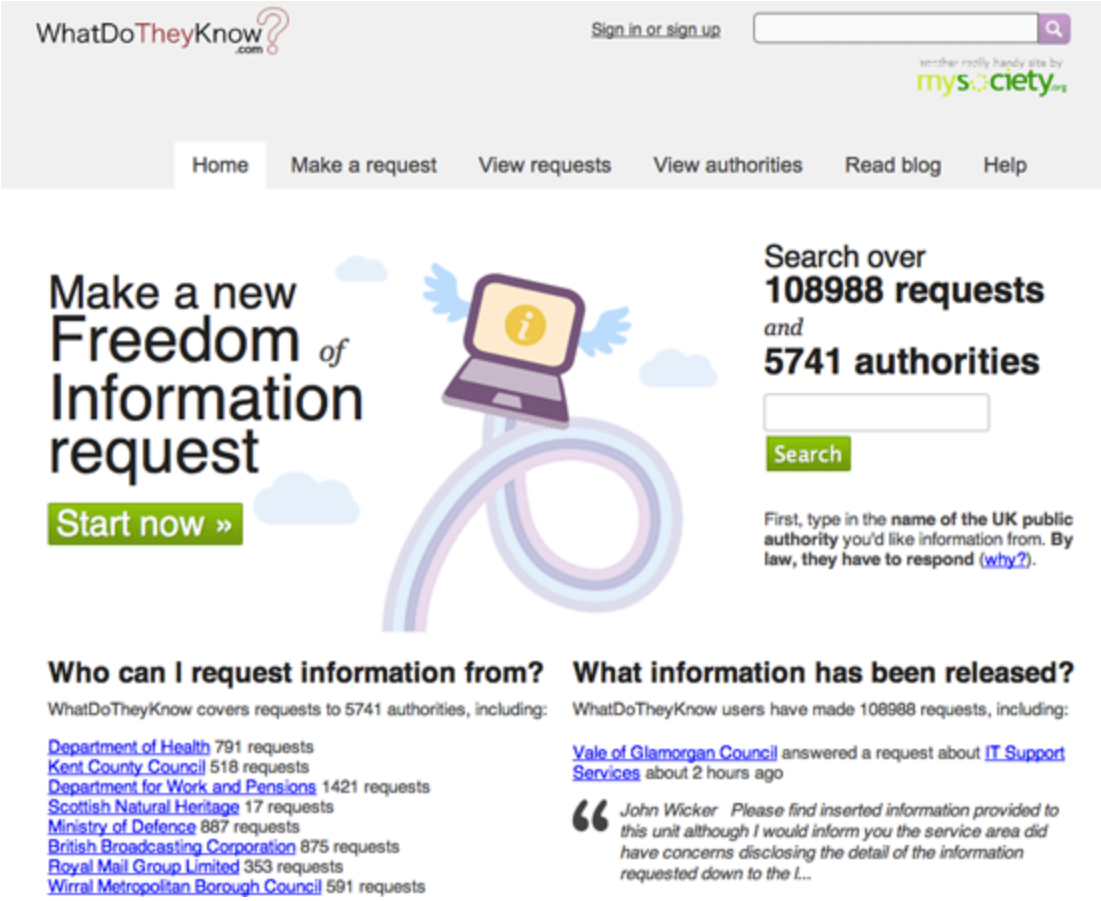


Figure 2. ¿Qué saben? (My Society)

Involucre a colegas

Si sus colegas son escépticos respecto del valor de los pedidos de acceso a la información, una de las mejores maneras de convencerlos es escribir un artículo basado en información que obtuvo usando una ley de acceso a la información. También se recomienda mencionar en el artículo final o en su alocución por radio o televisión que usó la ley, como un modo de subrayar su valor y aumentando la conciencia del público de la existencia de ese derecho.

Pida datos en crudo

Si quiere analizar, explorar, o manejar datos usando una computadora, entonces debe pedir explícitamente datos en formato electrónico que la máquina pueda leer. Puede clarificar esto especificando, por ejemplo, que requiere una información presupuestaria en un formato “adecuado para su análisis con software contable”.

También puede querer pedir explícitamente la información en forma desagregada o granular. Puede leer más acerca de esto en este informe (http://bit.ly/access-report)

Preguntar sobre organizaciones eximidas de las leyes de acceso a la información::

Usted puede querer investigar acerca de ONG, compañías privadas, organizaciones religiosas y/u otras organizaciones que no están obligadas a entregar documentación bajo las leyes de acceso a la información. Sin embargo es posible encontrar información acerca de ellas a través de entes públicos que sí están cubiertos por las leyes de acceso a la información. Por ejemplo, puede preguntar a un departamento o ministerio si han dado fondos o tratado con una compañía privada u ONG específica y pedir documentos que respalden la información. Si necesita más ayuda para hacer su pedido de acceso a la información puede consultar también el Legal Leaks

— Helen Darbishire (Access Info Europe), Djordje Padejski (Knight Journalism Fellow, Stanford University), Martin Rosenbaum (BBC), y Fabrizio Scrollini (London School of Economics and Political Science)

Usar pedidos de acceso a la información para entender el gasto

He usado pedidos de acceso a información de un par de maneras diferentes para ayudar a cubrir COINS, la mayor base de datos de gasto, presupuesto e información financiera del estado británico. Al comienzo de 2010 George Osborne sostuvo que si era nombrado al frente del Tesoro, daría acceso a COINS para facilitar una mayor transparencia. En ese momento pareció una buena idea investigar los datos y la estructura de COINS por lo que envíe unos cuantos pedidos de acceso a la información, uno para el esquema de la base de datos, otro para la orientación que reciben los trabajadores del Tesoro cuando trabajan con COINS y un tercero para el contrato del Tesoro con el proveedor de la base de datos. Todo lo cual resultó en la publicación de datos útiles. También pedí todos los códigos de gasto en la base de datos, información que también fue publicada. Todo esto ayudó a entender COINS cuando George Osborne llegó al Tesoro en mayo de 2010 y publicó COINS en junio de 2010. Los datos de COINS fueron usados en una cantidad de sitios de la red alentando al público a investigar los mismos, incluyendo OpenSpending.org y el Coins Data Explorer de The Guardian.

Luego de investigar un poco más pareció que faltaba una gran parte de la base de datos: la Whole of Government Accounts (WGA) que son 1500 conjuntos de cuentas para entes con financiación estatal. Usé un pedido de acceso a la información para solicitar los datos WGA de 2008/09 pero no obtuve resultados. También pedí el informe de la oficina de auditoría para WGA, que esperaba que explicara los motivos por los que la WGA no estaba en condiciones de publicarse. Eso también se me negó.

En diciembre de 2011 la WGA fue publicada en los datos COINS. Sin embargo quería asegurarme de que hubiera suficiente orientación para crear un conjunto completo de

cuentas para cada uno de los 1500 entes incluidos en el ejercicio de la WGA. Esto me lleva a la segunda manera en que usé un pedido de acceso a información: para asegurarme de que los datos difundidos bajo el plan de transparencia británico estuvieran bien explicados y contuvieran lo que debían. Presenté un pedido de acceso a la información del conjunto de cuentas para cada ente público incluido en la WGA.

— *Lisa Evans, the Guardian*

El Wobbing* funciona. ¡Úselo!

- N. del t. Wobbing es un neologismo surgido de la jerga periodística holandesa. La legislación de libre acceso a la información en Holanda se conoce por la sigla WOB. De allí se deriva el término.

Usar la legislación de acceso a la información –o wobbing, como se lo llama a veces- es una herramienta excelente pero requiere método y, a menudo, persistencia. A continuación, 3 ejemplos de mi propio trabajo como periodista de investigación que ilustran los puntos fuertes y los desafíos que plantea el wobbing.

Estudio de caso 1: subsidios agropecuarios

Todos los años la UE paga casi € 60.000 millones a productores y el sector agropecuario. Todos los años. Esto sucede desde fines de la década de 1950 y el discurso político era que los subsidios ayudan a los productores más pobres. Sin embargo, un primer logro en base a un pedido de acceso a la información en Dinamarca en 2004 mostró que esto eran solo palabras. Los pequeños productores estaban en graves dificultades, de lo que a menudo se quejaban en privado y en público, y en realidad la mayor parte de los fondos iban a unos pocos grandes terratenientes y a la gran industria agropecuaria. De modo que obviamente quise saber: ¿Esto es un patrón que abarca a toda Europa?

En el verano de 2004 le pedí los datos a la Comisión Europea. Todos los años en febrero la Comisión recibe datos de los estados miembros. Los datos muestran quien solicita fondos de la UE, cuánto reciben los beneficiarios y si lo reciben por explotar su tierra, desarrollar su región o para exportar leche en polvo. En aquel momento la Comisión recibía las cifras como archivos CSV en un CD. Muchos datos, pero con los que en principio era fácil trabajar. Es decir, si uno podía obtenerlos.

En 2004 la Comisión se negó a entregar los datos; el argumento clave fue que los datos estaban cargados en una base de datos y recuperarlos exigía mucho trabajo. Argumento que el Ombudsman Europeo llamó *mala administración*. Puede encontrar todos los documentos de este caso en el sitio sobre wobbing.eu. Allá por 2004 no teníamos tiempo de enredarnos en cuestiones legales. Queríamos los datos.

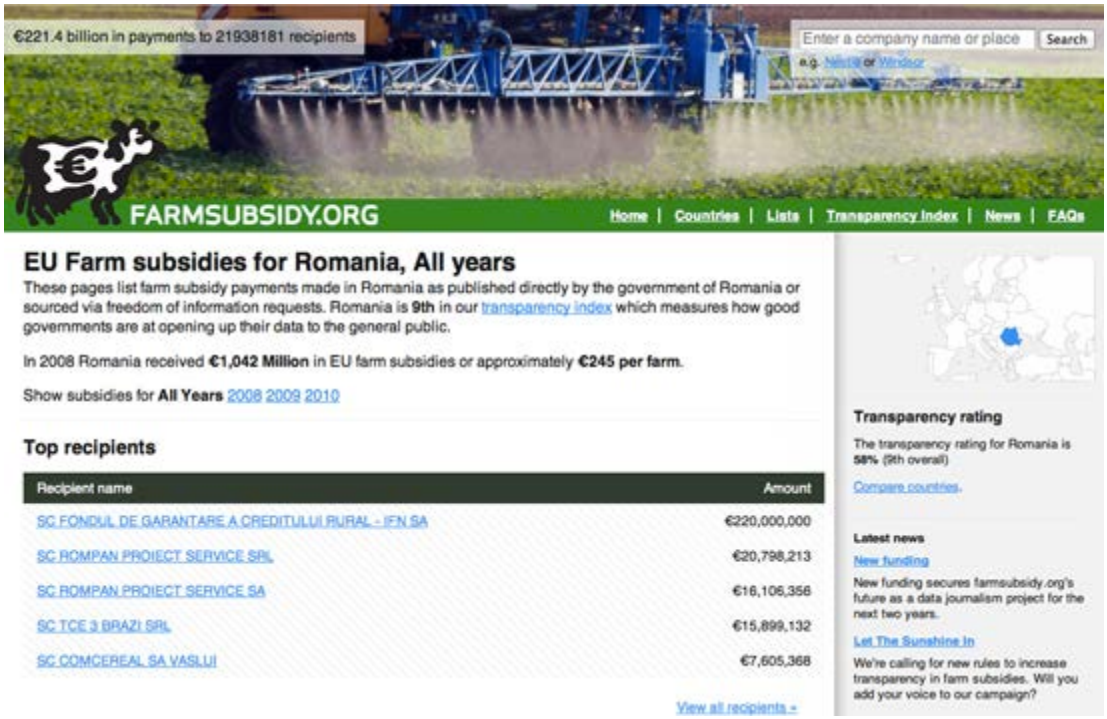


Figure 3. El sitio de los subsidios agrícolas (Farmsubsidy.org)

Por lo que nos asociamos con gente de toda Europa para obtener los datos país por país. Colegas ingleses, suecos y holandeses obtuvieron los datos en 2005. Finlandia, Polonia, Portugal y regiones de España, Eslovenia y otros países también ofrecieron su información. Incluso en Alemania, enemiga del wobbing, logré obtener algunos datos de la provincia del Norte del Rin – Westfalia en 2007. Tuve que recurrir a las cortes para obtener los datos, pero resultó en algunos buenos artículos en la revista Stern y en Stern online.

¿Fue casualidad que Dinamarca y el RU fueran los primeros en dar acceso a sus datos? No necesariamente. Si se mira el cuadro político general, los subsidios agropecuarios en aquel tiempo debían verse en el contexto de las negociaciones de la OMC en las que había presión contra los subsidios agropecuarios. Dinamarca y el RU se cuentan entre los países más liberales de Europa, de modo que bien pudo ser que los vientos políticos soplaran en dirección a una mayor transparencia en esos países.

La historia no se acaba allí; para consultar más episodios y los datos, ver Farm Subsidy.

Conozca sus derechos

Cuando publica datos, ¿debe preocuparse por el copyright y otros derechos en los datos? Aunque debe consultar siempre con su equipo legal, como regla: si está publicado por el estado no tiene porque pedir perdón ni permiso; si es publicado por una organización que no gana dinero vendiendo datos, no tiene mucho de qué preocuparse; si lo publica una organización que obtiene ganancias con la venta de datos, entonces decididamente tiene que pedir permiso.

— Simon Rogers, the Guardian

Estudio de caso 2: efectos colaterales

Todos somos conejillos de Indias en lo que se refiere a tomar medicamentos. Las drogas pueden tener efectos secundarios. Todos sabemos esto: sopesamos los beneficios y riesgos potenciales, y tomamos una decisión. Desgraciadamente, esta a menudo no es una decisión basada en información.

Cuando los adolescentes toman una píldora en contra de los granitos, esperan tener piel suave, no un mal estado de ánimo. Pero esto es precisamente lo que sucedió con una droga, con la que los jóvenes se deprimieron y hasta tuvieron tendencias suicidas por tomarla. El peligro de este efecto secundario en particular –-evidentemente una historia periodística-- no era algo demasiado conocido.

Hay datos sobre efectos secundarios. Los productores tienen que entregar información regularmente a las autoridades de salud acerca de los efectos secundarios observados. Esa información está en manos de las autoridades nacionales y europeas una vez que se permite la venta de la droga.

Nuevamente en este caso se tuvo un primer logro a nivel nacional en Dinamarca. Durante una investigación internacional de un equipo danés, holandés y belga, Holanda también dio la información. Otro ejemplo de salir de ronda con el *wobbing*: nos ayudó mucho poder señalar a las autoridades holandesas que los datos estaban accesibles en Dinamarca.

Pero la historia era cierta: en Europa había gente joven con tendencias suicidas y lamentablemente también hubo suicidios en varios países como resultado del uso de la droga. Periodistas, investigadores y las familias de una joven víctima presionaban duro para obtener acceso a esta información. El Ombudsman Europeo ayudó a presionar por más transparencia en el Ente Europeo de Medicina y parece que tuvo éxito. Por lo que ahora a los periodistas les corresponde obtener los datos y analizar el material a fondo. ¿Somos todos conejillos de Indias, como dijo un investigador, o son buenos los mecanismos de control?

Lecciones: no acepte una negativa cuando de lo que se trata es de transparencia. Sea persistente y siga una historia a lo largo de los años. Las cosas pueden cambiar, permitiendo mejor información con mejor acceso en un momento posterior.

Estudio de caso 3: contrabando de muerte

La historia reciente puede ser muy dolorosa para poblaciones enteras, en particular después de guerras y en tiempos de transición. ¿Entonces cómo pueden obtener datos duros los periodistas para investigar, cuando –por ejemplo- los que se beneficiaron de la última guerra ahora están en el poder? Esta es la tarea que se propuso un equipo de periodistas eslovenos, croatas y bosnios.

El equipo se dispuso a investigar los negocios con armas en la ex Yugoslavia durante el embargo de la ONU a comienzos de la década de 1990. La base del trabajo fueron documentos de investigaciones parlamentarias sobre el tema. Para documentar las rutas de embarque y comprender la estructura del comercio, se debía rastrear el transporte con números de embarcaciones en los puertos y las licencias de los camiones.

Comisiones parlamentarias eslovenas han hecho investigaciones sobre las ganancias obtenidas en las guerras de los Balcanes, pero nunca han llegado a ninguna conclusión. Pero había un rastro extremadamente valioso de documentos y datos desclasificados, incluyendo 6000 páginas que el equipo esloveno obtuvo a través de un pedido de acceso a información.

En este caso los datos debieron extraerse de documentos y bases de datos. Al aumentar los datos con más información, análisis e investigaciones, pudieron determinar numerosas rutas del comercio ilegal de armas.

El equipo tuvo éxito y los resultados son únicos y ya le han significado al equipo su primer premio. Lo que es más importante, la historia es valiosa para toda la región y bien puede ser retomada por periodistas en otros países por los que pasó la carga mortífera.

Lecciones: publique buen material en crudo aunque lo encuentre en lugares inesperados y combínelo con datos existentes de acceso público.

— Brigitte Alfter, Journalismfund.eu

Pedidos de acceso a la información con amigos

Muchos países balcánicos tienen problemas de corrupción gubernamental. La corrupción a menudo es incluso peor cuando se trata de que los gobiernos municipales rindan cuentas en esos países. Durante varios meses un grupo de periodistas serbios vinculados con el Centre for Investigative Reporting de Belgrado han estado cuestionando documentos del año 2009 de más de 30 municipalidades. Antes de eso, casi nada estaba accesible al público. La idea era obtener los registros originales y poner los datos en hojas de cálculo, para hacer controles y comparaciones básicas entre las municipalidades y obtener las cifras máximas y mínimas. Los indicadores básicos eran las cifras presupuestarias, gastos regulares y especiales, salarios de funcionarios, gastos de viaje, número de empleados, gastos de uso de celular, gastos diarios, cifras de compras oficiales, y así siguiendo. Era la primera vez que reporteros pedían esa información.

El resultado fue una base de datos que desnuda numerosas representaciones falsas, prácticas ilegales y casos de corrupción. Una lista de los alcaldes mejor pagos indicaba que unos cuantos de ellos recibían más dinero que el presidente serbio. Muchos otros funcionarios tenían sueldos excesivos, recibiendo muchos de ellos reintegros enormes de

expensas de viaje y por gastos. Los datos de gasto público obtenidos con mucho esfuerzo ayudaron a sacar a luz un enredo oficial. De la base de datos derivaron más de 150 artículos y muchos de ellos fueron reeditados por los medios locales y nacionales en Serbia.

Aprendimos que comparar los registros con datos comparables de entes gubernamentales similares puede sacar a luz desviaciones y echar luz sobre probables hechos de corrupción. Los gastos exagerados e inusuales solo pueden ser detectados por comparación.

— Djordje Padejski, Knight Journalism Fellow, Stanford University

Obtener datos de la red

Ha probado todo y no ha logrado obtener los datos que quiere. Encontró los datos en la red pero lamentablemente no hay opciones de descarga y fracasó en el intento de copiar y pegar. No tema, aún puede haber una manera de obtener los datos. Por ejemplo, puede:

- Obtener datos de APIs (interfaces de programación de aplicaciones) online, tales como interfaces provistas por bases de datos y muchas aplicaciones modernas (incluyendo Twitter, Facebook y otras). Esta es una manera fantástica de acceder a datos oficiales o comerciales, así como datos de redes sociales.
- Extraer datos de PDF. Esto es muy difícil, dado que PDF es un lenguaje para impresoras y no retiene mucha información sobre la estructura de los datos presentados en el documento. Extraer información de PDF va más allá del alcance de este libro, pero hay algunas herramientas y tutoriales que pueden ayudarlo a hacerlo.
- Usar "screen scraping" para obtener datos de sitios de la red. Se trata de extraer contenido estructurado de una página normal de la red con la ayuda de un programa de de recuperación de información o escribiendo una pequeña pieza de software. Si bien este método es muy poderoso y puede ser usado en muchos lugares, requiere comprender un poco cómo funciona la red.

Con todas esas opciones técnicas, no olvide las opciones simples: a menudo vale la pena invertir un poco de tiempo en buscar un archivo con datos que pueden ser interpretados por una computadora o llamar a la institución que tiene los datos que usted quiere.

En este capítulo presentamos un ejemplo muy básico de *scrapear* datos de una página HTML.

¿Qué son los datos procesables por computadora?

Para la mayoría de estos métodos, el objetivo es obtener acceso a datos que puedan ser nterpretados por una computadora. Tales datos son creados para ser procesados por una computadora en vez de ser presentados a un usuario humano. La estructura de estos datos se relaciona con la información contenida en ellos, y no la manera en que será presentada eventualmente. Entre los ejemplos de formatos que son fáciles de interpretar por una

computadora se incluyen CSV, XML, JSON, y los archivos Excel, mientras que formatos como los de documentos Word, páginas HTML, y archivos PDF están más relacionados con la presentación visual de la información. Por ejemplo, PDF es un lenguaje que le habla directamente a su impresora; le interesa la posición de líneas y puntos en una página, en vez de caracteres distinguibles.

====="Scrapear" sitios de la red: ¿Para qué?

Todos lo han hecho: se va a un sitio de la red, uno ve una tabla interesante y trata de copiarla a Excel de modo de poder agregar algunas cifras o guardarla para después. Pero a menudo esto no funciona realmente, o la información que quiere está desparramada en una gran cantidad de sitios. Copiar a mano se puede volver rápidamente muy tedioso, por lo que tiene sentido usar un poco de código para hacerlo.

La ventaja del "scraping" es que se puede hacer prácticamente con cualquier sitio, desde el pronóstico del tiempo hasta el gasto gubernamental, incluso si el sitio no tiene una API para acceso a los datos en crudo.

Lo que se puede y lo que no se puede "scrapear"

Por supuesto, hay límites a lo que se puede_scrapear_. Entre los factores que dificultan *scrapear* en un sitio se incluyen:

- Código HTML mal formateado con poco o nada de información estructural (por ejemplo, sitios oficiales más antiguos).
- Los sistemas de autenticación que se supone impiden el acceso automático (códigos CAPTCHA y exigencia de suscripción paga).
- Sistemas basados en sesiones que usan cookies de navegador para rastrear lo que hace el usuario.
- Falta de listados completos de ítems y ausencia de posibilidades de búsquedas con comodines.
- Bloqueado de acceso por administradores de servidores.

Otro conjunto de limitaciones son las barreras legales: algunos países reconocen los derechos de bases de datos, lo que puede limitar su derecho a reutilizar información que ha sido publicada online. A veces se puede ignorar la licencia y usarla de todos modos, dependiendo de su jurisdicción, puede tener derechos especiales como periodista. No debería haber problema en "scrapear" datos del estado de libre disponibilidad, pero quizás sea mejor cerciorarse antes de publicarlos. Organizaciones comerciales -–y ciertas ONGs-- reaccionan con menos tolerancia y pueden tratar de sostener que usted está “saboteando”

sus sistemas. Otras informaciones pueden violar la privacidad de individuos, y por tanto, violar las leyes de privacidad de datos o la ética profesional.

Emparchar, "Scrapear", compilar, limpiar

El desafío con muchos datos británicos no es lograr obtenerlos, si no ponerlos en un formato que se pueda usar. Se publican muchos datos sobre hospitalidad, los intereses de los parlamentarios fuera de su función pública, lobbys, y más como cosa habitual, pero en formatos difíciles de analizar.

Para algunos datos, la única alternativa es el trabajo duro: unir docenas de archivos Excel, cada uno conteniendo solo una docena de registros, fue la única manera de hacer listas completas de reuniones ministeriales. Para otros datos, "scrapear" la red se demostró increíblemente útil.

Usar un servicio como ScraperWiki para pedir a programadores que produzcan un *scraper* que permita reunir información como el Registro de intereses de parlamentarios, nos ahorró la mitad del trabajo: tuvimos toda la información de los parlamentarios en una hoja, lista para la “larga” tarea de analizarla y expurgarla.

Servicios como éste (o herramientas tales como Outwit Hub) son de inmensa ayuda para periodistas que tratan de compilar datos complicados y que son capaces de programar.

— *James Ball, the Guardian*

Herramientas que lo ayudan a "scrapear"

Hay muchos programas que pueden ser usados para extraer información en masa de un sitio, incluyendo extensiones de navegadores y algunos servicios de la red. Según el navegador que use, herramientas como Readability, que ayuda a extraer texto de una página o DownThemAll, que le permite descargar muchos archivos al mismo tiempo), le ayudarán a automatizar algunas tareas tediosas, mientras que la extensión Scraper de Chrome fue creada explícitamente para extraer tablas de sitios de la red. Extensiones para programadores como FireBug para Firefox, lo mismo ya viene incluido en Chrome, Safari e IE) le permite ver exactamente como está estructurado un sitio y qué comunicaciones se dan entre su navegador y el servidor.

ScraperWiki es un sitio que le permite crear *scrapers* en una cantidad de lenguajes de programación diferentes., incluyendo Python, Ruby y PHP. Si quiere comenzar a *scrapear* sin la complicación de instalar una plataforma de programación en su computadora esta es la manera de hacerlo. Otros servicios de la red, tales como las Hojas de Cálculo de Google y Yahoo! Pipes, también permiten realizar extracciones de otros sitios.

¿Cómo funciona un "Scraper" de la red?

Los "scrapers" de la red por lo general son piezas pequeñas de código escritas en un lenguaje de programación tal como Python, Ruby o PHP. Escoger el lenguaje adecuado depende en gran medida de a qué comunidad tiene acceso: si en su redacción o ciudad hay alguien que ya trabaja con uno de estos lenguajes, entonces tiene sentido adoptar el mismo lenguaje.

Si bien algunas de las herramientas de "scraping" con las que basta cliquear y apuntar mencionadas más arriba pueden ser de ayuda para comenzar, lo verdaderamente complejo a la hora de *scrapear* en un sitio es encontrar las páginas indicadas y los elementos indicados dentro de estas páginas para extraer la información deseada. Estas tareas no tienen que ver con programación, sino con comprender la estructura del sitio y la base de datos.

Al presentar un sitio, su navegador casi siempre usará dos tecnologías, HTTP, para comunicarse con el servidor y pedir recursos específicos, tales como documentos, imágenes o videos; y HTML, el lenguaje en el que se componen los sitios.

La anatomía de una página de la red

Toda página HTML está estructurada como una jerarquía de módulos (que están definidos por etiquetas de HTML). Un módulo grande contiene muchos módulos más pequeños –por ejemplo una tabla que tiene muchas divisiones más pequeñas: filas y celdas. Hay muchos tipos de etiquetas que realizan distintas funciones: algunas producen módulos, otras tablas, imágenes o vínculos. Las etiquetas también pueden tener propiedades adicionales (por ejemplo, pueden ser identificadores únicos y pueden pertenecer a grupos llamados “clases” que hacen posible apuntar a y capturar elementos individuales dentro de un documento). Escoger elementos apropiados de esta manera y extraer su contenido es la clave para escribir un "scraper".

Viendo los elementos en una página de la red, todo puede dividirse en módulos dentro de módulos.

Para "scrapear" en páginas de la red tendrá que aprender un poco acerca de los distintos tipos de elementos que pueden encontrarse en un documento HTML. Por ejemplo, el elemento <table> abarca toda una tabla, que tiene <tr> (table row) elementos para sus filas, que a su vez contienen <td> (table data) para cada celda. El tipo de elemento más común que encontrará es , que puede significar básicamente cualquier bloque de contenido. La manera más fácil de conocer estos elementos es usar la barra de desarrolladores, developer toolbar, de su navegador: le permitirá posicionarse sobre cualquier parte de una página de la red y ver el código correspondiente.

Las etiquetas funcionan como el comienzo y el fin de un libro, marcando el comienzo y el fin de una unidad. Por ejemplo `<em>` *significa el comienzo de un tramo de texto en itálica o destacado y* `</em>` significa el fin de ese tramo. Fácil.

Un ejemplo: "Scraping" de incidentes nucleares con Python

NEWS es el portal de la Agencia Internacional de Energía Atómica (AIEA) que sigue los incidentes de radiación en todo el mundo (y disputa el título máximo del club de los títulos raros). La página tiene listas de incidentes en un sitio simple, tipo blog, que puede ser fácilmente "scrapeado".

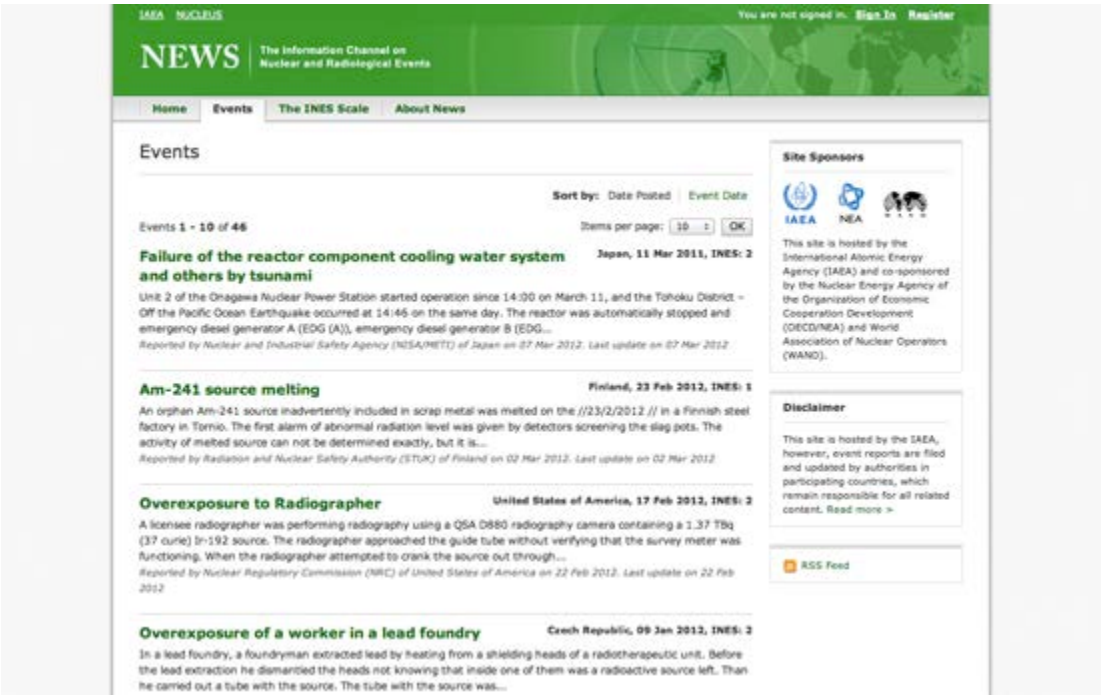


Figure 4. El portal de la Agencia Internacinal de Energía Atómica (IAEA) (news.iaea.org)

Para empezar, cree un nuevo scraper Python en ScraperWiki y se le presentará un área para texto mayormente vacía, salvo por algo de código de soporte. En otra pestaña del navegador abra el http://www-news.iaea.org/EventList.aspx[sitio de AIEA y abra la barra para desarrolladores de su navegador. En la vista de “elementos” trate de encontrar el elemento HTML de uno de los títulos de noticias. La barra para desarrolladores de su navegador le ayudará a relacionar los elementos en la página con el código HTML correspondiente.

Al investigar esta página se revelará que los títulos son elementos `<h4>` dentro de una `<table>`. Cada evento es una fila `<tr>`, que también contiene una descripción y una fecha. Si queremos extraer los títulos de todos los eventos, debiéramos buscar la manera de seleccionar cada fila en la tabla secuencialmente, recopilando todo el texto dentro de los elementos de título.

Para convertir este proceso en código, tenemos que tomar conciencia de todos los pasos. Para tener idea del tipo de pasos requeridos, juguemos un juego simple: en su ventana de

ScraperWiki trate de escribir instrucciones individuales para cada cosa que va a hacer mientras escribe este "scraper", como los pasos de una receta (ponga al comienzo de cada línea un signo de numeral para indicarle a Python que no es un verdadero código de computación). Por ejemplo:

```
# Buscar todas las filas en la tabla
# Unicornio no debe desbordar el lado izquierdo.
```

Trate de ser lo más preciso posible y no suponga que el programa sabe algo acerca de la página que intenta *scrapear*.

Una vez que haya escrito algo de este seudo-código, comparemos esto con el código esencial para nuestro primer *scraper*:

```
import scraperwiki
from lxml import html
```

En esta primera sección estaba importando funcionalidad existente de bibliotecas, recortes de código ya escrito. `Scraperwiki` nos dará la capacidad de descargar sitios de la red, mientras que `lxml` es una herramienta para el análisis estructurado de documentos HTML. Buena noticia: si está escribiendo un scraper con ScraperWiki, estas dos líneas siempre serán las mismas.

```
url = "http://www-news.iaea.org/EventList.aspx"
doc_text = scraperwiki.scrape(url)
doc = html.fromstring(doc_text)
```

A continuación el código hace un nombre (variable): url, y asigna el URL de la página de la AIEA como su valor. Esto le dice al "scraper" que esta cosa existe y que queremos prestarle atención. Nótese que el URL mismo está entre comillas ya que no es parte del código del programa sino un *string*, una secuencia de caracteres.

Entonces usamos la variable del url como entrada de una función, `scraperwiki.scrape`. Una función que producirá una tarea definida, en este caso, descargará una página de la red. Cuando termine, asignará su producto a otra variable, `doc_text`. `doc_text` ahora contendrá el texto del sitio; no la forma visual que ve en su navegador, sino el código fuente, incluyendo todas las etiquetas. Dado que esta forma no es muy fácil de analizar, usaremos otra función, `html.fromstring`, para generar una representación especial, en la que podamos fácilmente referirnos a elementos, el así llamado modelo de documento de objeto o document object model (DOM).

```
for row in doc.cssselect("#tblEvents tr"):
link_in_header = row.cssselect("h4 a").pop()
```

```
event_title = link_in_header.text
print event_title
```

En este paso final, usamos el DOM para encontrar cada fila de nuestra tabla y extraer el título del evento de su encabezado. Se usan dos conceptos nuevos: el riso "for loop" y selección de elemento o "element selection" (`.cssselect`). El "for loop" hace esencialmente lo que implica su nombre; atraviesa una lista de ítems, asignando a cada uno un alias temporal (`row` en este caso) y luego aplicará las instrucciones con sangría para cada ítem.

El otro concepto nuevo, selección de elemento o "element selection", es hacer uso de un lenguaje especial para encontrar elementos en el documento. Los selectores CSS son usados normalmente para agregar información de diseño a elementos HTML y puede ser usado para extraer con precisión un elemento de una página. En este caso (línea 6) estamos seleccionando #tb1Events tr, que hará corresponder cada `<tr>` en el elemento tabla con el ID tb1Events (el signo numeral simplemente significa ID). Nótese que esto producirá una lista de elementos `<tr>`.

Eso puede verse en la siguiente línea (línea 7i), donde estamos aplicando otro selector para encontrar cualquier <a> (que es un hipervínculo) dentro de un `<h4>` (un título). Aquí sólo queremos ver un elemento (solo hay un título por fila), de modo que tenemos que sacarlo del encabezado de la lista creada por nuestro selector con la función `.pop()`.

Nótese que algunos elementos en el DOM contienen texto (es decir, aneder usando la sintaxis `[element].text` que se ve en la línea 8. Finalmente en la línea 9 estamos imprimiendo ese texto a la consola ScraperWiki. Si hace clic en "run" en su "scraper", la ventana más pequeña ahora debiera comenzar a listar los nombres del evento del sitio de la AIEA.

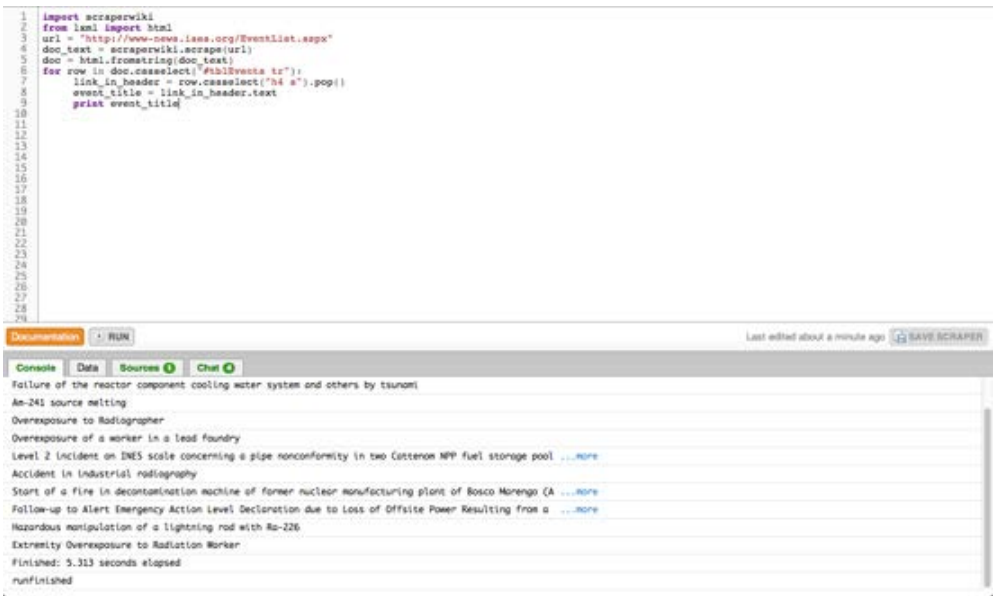


Figure 5. Un scraper en acción (ScraperWiki)

Ahora puede ver un "scraper" básico operando: descarga la página, la transforma a la forma DOM, y luego le permite seleccionar y extraer cierto contenido. Dado este esqueleto, puede tratar de resolver algunos de los problemas que quedan usando la documentación del ScraperWiki y Python:

- ¿Puede encontrar la dirección del vínculo en el título de cada evento?
- ¿Puede seleccionar el pequeño módulo que contiene la fecha y el lugar usando su nombre de clase CSS y extraer el texto del elemento?
- ScraperWiki ofrece una pequeña base de datos para cada scraper, de modo que pueda almacenar los resultados; copie el ejemplo correspondiente de sus docs y adáptelo de modo que guarde los títulos, vínculos y fechas del evento.
- La lista de eventos tiene muchas páginas; ¿puede *scrapear* múltiples páginas para obtener eventos históricos también?

Mientras intenta resolver estos desafíos, investigue un poco el ScraperWiki: hay muchos ejemplos útiles en los "scrapers" existentes; a menudo los datos son bastante interesantes también. De este modo no necesita comenzar su "scraper" de cero: simplemente escoja uno similar, tómelo y adáptelo a su problema.

— *Friedrich Lindenberg, Open Knowledge Foundation*

"Scrapear" en una base de datos pública

Algunos médicos franceses pueden establecer libremente sus honorarios, por lo que uno puede pagar entre € 70 y € 500 por una consulta de 30 minutos con un oncólogo, por ejemplo. Los datos sobre honorarios por ley son públicos, pero la administración solo ofrece una base de datos online difícil de navegar. Para tener una buena visión de los honorarios de los médicos para Le Monde, decidí "scrapear" toda la base de datos.

Ahí comenzó la diversión. De entrada, el formulario de búsqueda era una aplicación Flash que redirigía a una página de resultados HTML vía un pedido POST. Con ayuda de Nicolas Kayser-Bril, nos llevó algo de tiempo descubrir que la aplicación usaba una tercera página como paso “oculto” entre el formulario de búsqueda y la página de resultado. Esta página se usaba en realidad para almacenar un cookie con valores del formulario de búsqueda al que entonces accedía la página de resultados. Hubiese sido difícil imaginarse un proceso más enredado, pero las opciones de la biblioteca cURL en PHP permiten superar fácilmente las vallas, una vez que se sabe cuáles son. Finalmente apoderarnos de la base de datos llevó 10 horas, pero valió la pena.

— *Alexandre Léchenet, Le Monde*

La red como fuente de datos

¿Cómo puede saber más de algo que solo existe en Internet? Esté buscando una dirección de correo electrónico, sitio, imagen o artículo de Wikipedia, en este capítulo haré con usted una recorrida por las herramientas que le dirán más sobre ellos.

Herramientas web

Primero, unos cuantos servicios diferentes que puede usar para descubrir algo más sobre todo un sitio, en vez de una página particular:

Whois

: Si va a whois.domaintools.com/ o simplemente tipea whois seguido de un URL *www.ejemplo.com* en Terminal.app en una Mac puede obtener la información básica de registro de cualquier sitio. En los últimos años algunos dueños han preferido el registro privado, lo que oculta sus detalles, pero en muchos casos verá un nombre, dirección, correo electrónico y número de teléfono de la persona que registró el sitio. También puede ingresar direcciones IP numéricas aquí y obtener datos sobre la organización o el individuo que es dueño del servidor. Esto es especialmente útil cuando trata de encontrar más información sobre un usuario abusivo o malicioso de un servicio, ya que la mayoría de los sitios registran una dirección IP de todo el que accede a ellos.

Blekko

: El motor de búsquedas (Blekko ofrece una cantidad inusual de información sobre las estadísticas internas que reúne sobre sitios mientras recorre la red. Si tipea un nombre de dominio seguido de “/seo”, verá una página de información sobre ese URL. La primera pestaña en Figure 7 le muestra qué otros sitios se vinculan con el dominio por orden de popularidad. Esto puede ser extremadamente útil cuando está tratando de comprender qué tipo de cobertura recibe un sitio y por qué tiene un alto ranking en los resultados de búsquedas de Google, ya que estos se basan en esos vínculos entrantes. Figure 8 le dice qué otros sitios funcionan en la misma máquina. Es común que estafadores y la gente que envía spam se trate de legitimar construyendo múltiples sitios que se ensalzan y vinculan mutuamente. Parecen dominios independientes e incluso pueden tener detalles de registro diferentes, pero a menudo están en el mismo servidor porque eso es mucho más barato. Estas estadísticas le dan una visión de la estructura oculta del sitio que investiga.

Figure 6. El buscador Blekko Blekko.com)

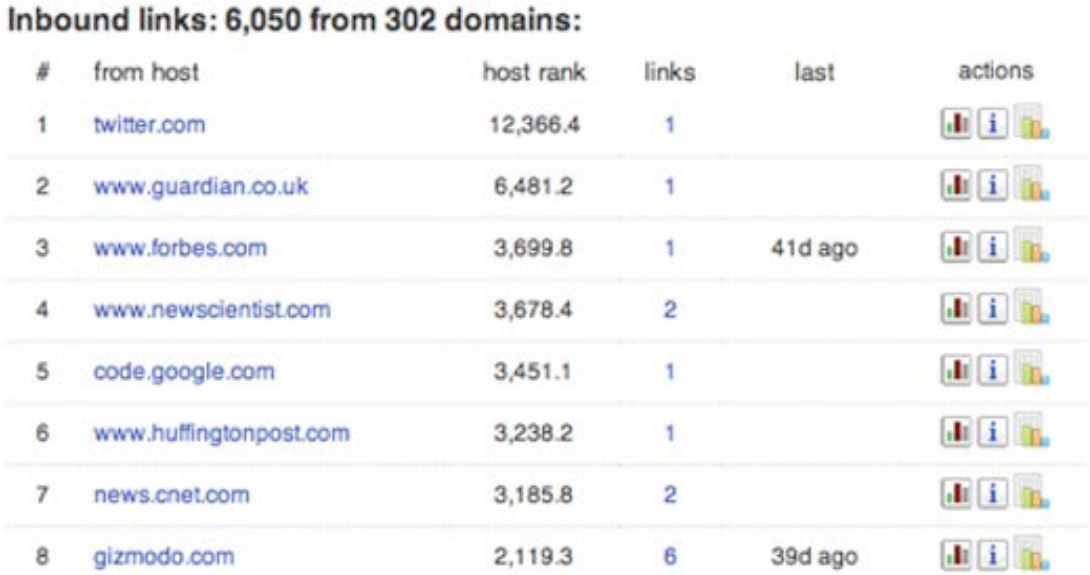


Figure 7. Comprender la popularidad en la red, ¿quién se vincula con quién? La otra pestaña útil es “Estadísticas de Navegación"), especialmente la sección “Co-huesped con”.(Blekko.com)

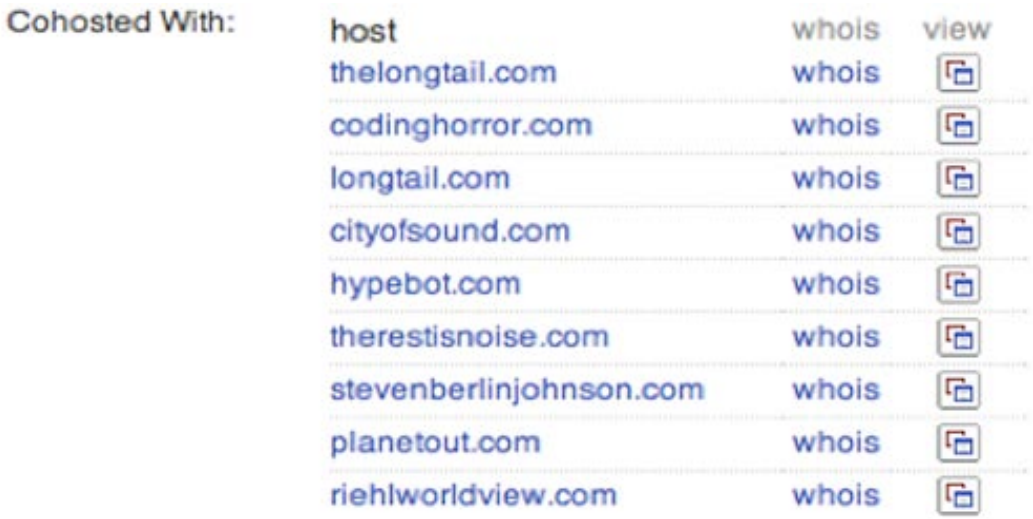


Figure 8. Descubrir spammers y estafadores de la red (Blekko.com)

Compete.com

Al estudiar una muestra representativa de consumidores estadounidenses, Compete.com acumula estadísticas de uso detalladas para la mayoría de los sitios y pone a disposición gratuitamente algunos detalles básicos. Elija la pestaña de Site Profile (Perfil de Sitio) e ingrese un dominio (Figure 9). Entonces verá un gráfico del

tráfico del sitio en el último año, junto con cifras de cuánta gente lo visitó y con qué frecuencia (como en Figure 10). Dado que se basan en muestras los números son solo aproximados, pero yo los encontré razonablemente precisos cuando pude compararlos con la analítica interna. En particular, parecen ser una buena fuente para comparar dos sitios, dado que aunque las cifras absolutas pueden ser equivocadas para ambos, de todos modos es una buena representación de su diferencia relativa en cuanto a popularidad. Pero solo estudian a los consumidores estadounidenses, por lo que los datos serán pobres para los sitios predominantemente internacionales.

Figure 9. El servicio de perfil de Compete (Compete.com)

Figure 10. ¿Qué está de moda? ¿De qué hay demanda?: Lugares calientes de la red (Compete.com)

El buscador de sitios (Site Search) de Google

Un recurso que puede ser extremadamente útil cuando trata de explorar todo el contenido de un dominio particular es ingresar en el buscador los términos “sitio”: palabra clave. Si agrega “site:ejemplo.com” a su frase de búsqueda, Google solo presentará resultados del sitio que ha especificado. Incluso puede afinar aún más la búsqueda incluyendo el prefijo de las páginas que le interesan, por ejemplo, “site: ejemplo.com/páginas/”, y solo verá los resultados que responden a ese patrón. Esto puede ser extremadamente útil cuando trata de encontrar información que los dueños de dominios ofrecen públicamente pero que no desean difundir, de modo que elegir las palabras claves correctas puede permitir descubrir material muy revelador.

Páginas, imágenes y videos en la red

A veces lo que interesa es la actividad que rodea una historia específica, en vez de un sitio entero. Las herramientas que se presentan a continuación le dan distintos ángulos de cómo lee, responde, copia y comparte contenido la gente en la red.

Bit.ly

Siempre recurro a bitly.com cuando quiero saber cómo comparte la gente un vínculo particular. Para usarlo, ingrese el URL que le interesa. Luego haga clic en el vínculo Info Page+. Eso lo lleva a la página de estadísticas completas (aunque puede tener que escoger el vínculo "aggregate bit.ly" primero si ha ingresado en el servicio). Esto le dará una idea de la popularidad de la página, incluyendo actividad en Facebook y Twitter y debajo de eso verá conversaciones públicas respecto del vínculo provistas por backtype.com. Esta combinación de datos de tráfico y conversaciones me resulta muy útil cuando trato de comprender por qué un sitio o página es popular y quiénes son sus fans. Por ejemplo me aportó fuertes evidencias de que la opinión dominante respecto de la relación de Sarah Palin con los delegados de base era equivocada.

Twitter

Al ser el servicio de micro-blogging más usado, es útil parar ver en qué medida la gente comparte y habla acerca de piezas de contenido individuales. Es engañosamente simple descubrir conversaciones públicas sobre un vínculo. Uno simplemente pega el URL en el que está interesado en la ventana de búsqueda y luego posiblemente hace clic en "más tweets" para ver todos los resultados.

Cache de Google

Cuando una página se vuelve polémica los editores la pueden bajar o alterarla sin reconocerlo. Si cree que se está encontrando con este problema, el primer lugar a ir es el cache de Google de la página tal como era cuando hizo su último recorrido. La frecuencia de los recorridos está aumentando constantemente, por lo que tendrá más suerte si intenta esto dentro de las pocas horas posteriores a que se produjeron los supuestos cambios. Ingrese el URL correspondiente en la ventana de búsqueda de Google y luego haga clic en la flecha triple, a la derecha del resultado para esa página. Debiera aparecer una vista gráfica y si tiene suerte habrá un pequeño vínculo de "Cache" arriba. Haga clic allí para ver la toma de Google de la página. Si hay problemas para que cargue, puede cambiar a la página más primitiva, solo de texto, haciendo clic en otro link arriba de la página en cache completa. Usted tendrá que guardar la imagen de la pantalla o copiar y pegar el contenido significativo que encuentre, dado que puede quedar invalidado en cualquier momento por nuevos cambios.

La Wayback Machine (Máquina de Hace Tiempo) del Archivo de Internet

Si necesita saber cómo ha cambiado una página particular en un período de tiempo más largo, como meses o años, el Archivo de Internet tiene un servicio llamado The Wayback Machine que periódicamente hace tomas de las páginas más populares de la red. Vaya al sitio, ingresa el vínculo que quiere buscar y si hay copias, le mostrará un calendario para el momento que quiere examinar. Entonces presentará una versión de la página aproximadamente como era en aquel momento. A menudo le faltará diseño o imágenes, pero por lo general basta para entender cuál era el foco del contenido de la página en ese momento.

Ver el Código Fuente

Es algo un poco improbable, pero los diseñadores a menudo dejan comentarios u otros indicios en el código HTML de cualquier página. Estará en distintos menúes según el navegador que use, pero siempre hay una opción de "view source" (ver código fuente), que le permitirá recorrer el HTML en crudo. No necesita entender lo que significan las partes solo legibles para la máquina, solo esté atento a los tramos de texto que a menudo están desparramados en medio del código. Aunque solo sean referencias de copyright o menciones de los nombres del autor, estos a menudo pueden dar pistas importantes acerca de la creación y el objetivo de la página.

TinEye

A veces uno realmente quiere conocer el origen de una imagen, pero sin un texto claro que lo indique no hay ninguna manera evidente de hacerlo con motores de búsqueda tradicionales como Google. TinEye ofrece un proceso especializado de "búsqueda inversa de imagen", donde uno le da la imagen que tiene y encuentra otras imágenes en la red que se ven muy similares. Debido a que usa reconocimiento de imagen para hacer la búsqueda, funciona incluso cuando una copia ha sido recortada, distorsionada o comprimida. Esto puede ser extremadamente efectivo cuando usted sospecha que una imagen que se presenta como original o nueva no lo es, dado que puede reconducirlo a la verdadera fuente original.

YouTube

Si hace clic en el ícono de estadísticas en el ángulo inferior derecho de cualquier video, puede conseguir información valiosa sobre su público a lo largo del tiempo. Si bien no es completa, es útil para entender aproximadamente quienes son los espectadores, de donde vienen y cuándo.

Correo electrónico

Si está investigando correos electrónicos, a menudo querrá conocer más detalles sobre la identidad y ubicación del que los envió. No hay una buena herramienta disponible para ayudar con esto, pero puede ser muy útil conocer lo básico acerca de los encabezados

ocultos incluidos en todo mensaje de correo electrónico. Estos funcionan como indicadores para el correo y pueden revelar mucho acerca del remitente. En particular, a menudo incluyen la dirección IP de la máquina desde la que fue enviado el correo, parecido a la identidad del que hace una llamada telefónica. Puede entonces usar "whois" con ese número IP para saber qué organización posee esa máquina. Si resulta ser alguien como Comcast o AT&T que proveen conexiones a consumidores, entonces puede visitar MaxMind para obtener su ubicación aproximada.

Para ver estos encabezados en Gmail abra el mensaje y [line-through]*abra*el menú junto a la respuesta arriba a la derecha y elija “Mostrar original”.

Entonces verá una nueva página que revela el contenido oculto. Al comienzo habrá un par de docenas de líneas que son palabras seguidas por una coma. La dirección IP que busca puede estar allí, pero el nombre dependerá de cómo fue enviado el correo. Si se envió desde Hotmail, se llamará `X-Originating-IP:`, pero si fue enviado desde Outlook o Yahoo estará en la primera línea que comienza con `Received:`.

Si investigo la dirección con Whois me dice que está asignado a Virgin Media, un ISP del RU, por lo que uso el servicio de ubicación geográfica de MaxMind para descubrir que viene de mi ciudad, Cambridge. Esto significa que puedo estar razonablemente confiado de que se trata efectivamente de un correo de mis padres y no de impostores.

Tendencias

Si está investigando un tema amplio en vez de un sitio o ítem particular, estas son algunas herramientas que pueden ayudar:

Wikipedia Article Traffic (Tráfico de Artículos de Wikipedia)

: Si le interesa conocer cómo ha variado el interés del público sobre un tema o persona a lo largo del tiempo, puede encontrar cifras de vistas día por día para cualquiera página de Wikipedia en stats.grok.se. Es un sitio un poco tosco, pero le permitirá descubrir la información que necesita revolviendo un poco. Ingrese el nombre que le interesa para tener una visión mensual del tráfico en esa página. Eso le presentará un gráfico que muestra cuántas veces fue vista la página cada día del mes que usted especifique. Desgraciadamente solo se puede ver un mes por vez, por lo que tendrá que seleccionar otro mes y volver a buscar, para ver cambios en períodos más prolongados.

Google Insights

: Puede tener una clara visión de los hábitos de búsquedas del público usando Insights de Google (Figure 11). Ingrese un par de frases de búsquedas comunes, como “Justin Bieber vs Lady Gaga”, y verá un gráfico de sus números relativos de búsquedas con el

paso del tiempo. Hay muchas opciones para refinar su vista de los datos, desde zonas geográficas más reducidas hasta más detalle a medida que pasa el tiempo. Lo único que falta son valores absolutos: solo verá porcentajes relativos, lo que puede ser difícil de interpretar.

Figure 11. Google Insights (Google)

— *Pete Warden, analista de datos y diseñador independiente*

Crowdsourcing en el Datablog de The Guardian

"Crowdsourcing", según Wikipedia, “consiste en externalizar tareas que, tradicionalmente, realizaba un empleado o contratista, a un grupo numeroso de personas o una comunidad, a través de una convocatoria abierta”. Lo que sigue está tomado de una entrevista con Simon Rogers acerca de cómo el Datablog usó "crowdsourcing" para cubrir el escándalo de los gastos de parlamentarios, el uso de drogas y los papeles de Sarah Palin:

A veces uno recibe una tonelada de archivos, estadísticas o informes que es imposible que una persona pueda analizar. También puede conseguir material que es inaccesible o está en un mal formato y no puede hacer demasiado. Es en esto que puede ayudar el "crowdsourcing".

Una cosa que tiene The Guardian es muchos lectores, muchos pares de ojos. Si hay un proyecto interesante en el que necesitamos su ayuda, entonces se lo pedimos. Es lo que hicimos con los Gastos de los parlamentarios. Teníamos 450.000 documentos y poco tiempo para hacer algo. ¿Entonces qué cosa mejor que repartir la tarea entre los lectores?

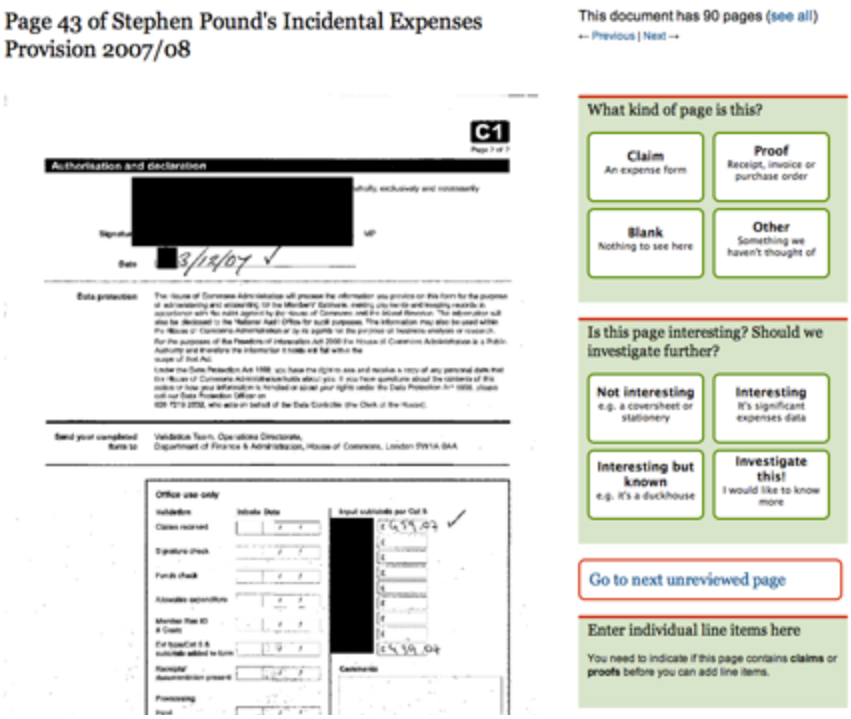


Figure 12. Una copia redactada de los gastos incidentales de Stephen Pound (The Guardian)

El proyecto de los gastos de los parlamentarios generó muchas pistas. Tuvimos más historias que datos. El proyecto fue llamativamente exitoso en términos de tráfico. A la gente realmente le gustó.

Actualmente estamos haciendo algo con MixMag sobre el uso de drogas, que también ha sido fenomenal. Parece que va a ser más grande que la encuesta sobre crímenes en Gran Bretaña en términos de la cantidad de gente que vuelve, lo que es brillante.

Lo que ambos proyectos tienen en común es que se refieren a temas que realmente le importan a la gente, por lo que está dispuesta a dedicarles su tiempo. Mucho del *crowdsourcing* que hemos hecho depende de la ayuda de obsesivos. Con los gastos de los parlamentarios tuvimos una cantidad masiva de tráfico al comienzo y luego bajó. Pero seguimos teniendo gente que lee obsesivamente cada página buscando anomalías e historias. Una persona ha leído 30.000 páginas. Saben muchas cosas.

También usamos "crowdsourcing" con los papeles de Sarah Palin. También en este caso fue de gran ayuda para estudiar la información en crudo en busca de historias.

En términos de generar historias el "crowdsourcing" ha funcionado muy bien. A la gente realmente le gusta e hizo quedar bien a The Guardian. Pero en términos de generar datos no hemos usado el "crowdsourcing" tanto.

Algunos de los proyectos de "crowdsourcing2 que hemos hecho y que funcionaron realmente bien, han sido encuestas a la antigua. Cuando uno le pregunta a la gente acerca de su experiencia, su vida, lo que han hecho, eso funciona muy bien porque la gente no

tiende a inventar en esos casos. Dice lo que siente. Cuando le pedimos a la gente que haga nuestro trabajo por nosotros hay que encontrar una especie de marco para que la gente produzca datos de un modo que resulten confiables.

Respecto de la confiabilidad de los datos, creo que la postura de Old Weather es realmente buena. Consiguen que 10 personas hagan cada entrada, que es una buena manera de asegurarse precisión. Con los gastos de los parlamentarios tratamos de minimizar el riesgo de que los mismos parlamentarios se metieran online a editar sus datos para quedar mejor. Pero no se puede estar permanentemente cuidándose de esto. Sólo se puede estar atento a ciertos URL o si provienen de la zona SW1 de Londres. Así que eso es un poco más difícil. Los datos que sacábamos no eran siempre confiables. Aunque las historias eran muy buenas, no producía números en crudo que pudiéramos usar con certeza.

Si tuviera que dar consejos a quienes aspiran a ser periodistas de datos y que quieren usar el "crowdsourcing" para obtener datos, los alentaría a hacerlo con algo que a la gente realmente le importa y que le seguirá importando cuando deje de producir titulares de primera página. Además, si uno puede crear algo que se parezca a un juego, eso puede ayudar realmente a atraer a la gente. Cuando hicimos la historia de los gastos por segunda vez, fue mucho más como un juego con tareas individuales para que las hiciera la gente. Realmente fue de ayuda dar a la gente tareas específicas. Eso fue importante porque creo que si uno solo le presenta a la gente una montaña de información que tiene que ver y le dice “mire esto”, puede resultar un trabajo duro y poco grato. Por lo que creo que es realmente importante hacer que sea divertido.

— Marianne Bouchart, Data Journalism Blog, interviewing Simon Rogers, the Guardian

Cómo el Datablog usó "crowdsourcing" para cubrir la venta de entradas para las Olimpíadas

Creo que el proyecto de *croudsourcing* que tuvo la mayor respuesta fue un trabajo sobre la subasta de entradas para las Olimpíadas. Miles de personas en el RU trataron de obtener entradas para la Olimpíada de 2012 y hubo mucha indignación porque la gente no las recibió. La gente había hecho pedidos por cientos de libras y se les dijo que no recibirían nada. Pero nadie sabía si eran solo unas pocas personas las que se quejaban ruidosamente mientras la mayoría estaba contenta. Por lo que intentamos encontrar una manera de saberlo.

Decidimos que lo mejor que podíamos hacer realmente, dado que no había buenos datos sobre el tema, era preguntar a la gente. Y pensamos que tendríamos que tratarlo como un tema no demasiado serio, porque no teníamos una muestra representativa.

Creamos un formulario en Google e hicimos preguntas muy específicas. En realidad era un cuestionario largo: preguntaba cuánto era el valor de las entradas que habían pedido, cuánto habían debitado de sus tarjetas de crédito, qué eventos querían ver, este tipo de cosas.

How many Olympic tickets did you get? Here's our readers' results

We asked how you had fared in the London 2012 ticket ballot. Here's our analysis of the information you gave us

Tweet 21 Share 84 reddit this Comments (144)

The London 2012 Olympic stadium: will you be there? Will anyone you know? Photograph: Handout/Getty Images

Posted by James Ball Friday 3 June 2011 11.04 BST guardian.co.uk Article history

A larger | smaller

Sport Olympic tickets · Olympic Games 2012

More from London 2012 Olympics blog on

Sport Olympic tickets · Olympic Games 2012

More blogposts

How much money did London 2012 ticket buyers have to put on the line to stand even a 50:50 chance of getting at least one ticket? If the results submitted by Guardian readers are to be believed, at least £1,000.

Earlier this week, we asked readers of the Guardian London 2012 blog to let us know how their ticket-purchasing attempts had fared. By the end of the day – when this analysis was carried out – we'd had more than 5,000 responses.

Figure 13. ¿Cuántas entradas Olímpicas consiguió?: los resultados de los lectores (The Guardian)

Lo pusimos como una pequeña imagen a la cabeza del sitio y se difundió rápidamente. Creo que esta es una de las cosas claves; no se puede solo pensar "¿Qué es lo que quiero saber para mi historia?". Hay que pensar: "¿Qué me quiere contar la gente ahora?" Y el "crowdsourcing" es exitoso cuando uno descubre de qué quiere hablar la gente. El volumen de respuestas para este proyecto, que es uno de nuestros primeros intentos de "crowdsourcing", fue inmenso. Tuvimos 1.000 respuestas en menos de una hora y 7.000 para el final del día.

Por lo que obviamente, tomamos un poco más seriamente la presentación de los resultados en este momento. Inicialmente no sabíamos cómo nos iba a ir. Por lo que agregamos

algunas consideraciones: los lectores del Guardian pueden tener mayores ingresos que otra gente, la gente que recibió menos de lo esperado podía estar más dispuesta a hablar con nosotros, y así siguiendo.

No sabíamos cuánto valor tendrían los resultados. Terminamos con unos 7.000 registros en los cuales basar nuestro trabajo, y descubrimos que alrededor de la mitad de la gente que pidió entradas no recibió nada. Presentamos todo esto y debido a que tanta gente había participado el día anterior, hubo mucho interés en los resultados.

Pocas semanas más tarde salió el informe oficial y nuestras cifras resultaron llamativamente precisas. Eran casi exactas. Creo que en parte fue por una cuestión de suerte, pero también porque logramos que respondiera tanta gente.

Si uno le pregunta a sus lectores sobre algo así y contestan en los comentarios de la nota, estará limitado en lo que puede hacer con los resultados. De modo que tiene que empezar por pensar: "¿Cuál es la mejor herramienta para lo que quiero saber?" ¿Es un hilo de comentarios? ¿O tengo que crear una aplicación? Y si es crear una aplicación, hay que pensar: "¿Vale la pena la espera? ¿Y se justifican los recursos requeridos para hacer esto?"

En este caso pensamos en los Formularios Google. Si alguien llena el formulario el resultado aparece como una fila en una hoja de cálculo. Esto significa que aunque aún si se estuviera actualizando, aún si siguieran entrando resultados, se podría abrir la hoja de cálculo y ver todos los resultados.

Pude haber tratado de hacer el trabajo en Google, pero lo descargué a Microsoft Excel y luego ordené la información de menor a mayor; también encontré las entradas en las que la gente para decir lo que gastó, había escrito los números como palabras (en vez de colocar los dígitos), y arreglé eso. Decidí excluir lo menos posible. De modo que en vez de solo aceptar las respuestas válidas, traté de arreglar lo que tenía. Algunos habían usado divisas extranjeras, así que las convertí a libras, todo lo cual fue un poco trabajoso.

Pero hice todo el análisis en pocas horas y eliminé las entradas obviamente tontas. Mucha gente decidió decir que no había gastado nada en entradas. Eso es un poco gracioso, pero está bien. Eran menos de cien en más de 7.000 entradas.

También hubo unas pocas docenas de personas que ingresaron cifras demasiado elevadas para tratar de distorsionar los resultados. Cosas como 10.000.000 de libras. Por lo que eso me dejó con un conjunto de datos que podía usar con los principios normales que usamos todos los días. Hice lo que se llama una "tabla dinámica" (pivot table). Hice algunos porcentajes. Ese tipo de cosas.

No teníamos idea del impacto que tendría el proyecto, de modo que trabajé yo solo con el editor del blog de deportes. Juntamos cabezas y pensamos que este podía ser un proyecto divertido. Lo hicimos, de comienzo a fin, en 24 horas. Tuvimos la idea, a la hora del

almuerzo armamos algo, lo pusimos a la cabeza del sitio, vimos que resultaba bastante popular, lo dejamos a la cabeza del sitio el resto del día y presentamos los resultados online a la mañana siguiente.

Decidimos usar Google Docs porque da completo control sobre los resultados. No necesitaba usar las herramientas analíticas de otra gente. Lo puedo trasladar fácilmente a un software de base de datos o a hojas de cálculo. Cuando uno usa el software de consultas de especialistas, a menudo se ve restringido a usar las herramientas de ellos. Si hubiésemos estado pidiendo información muy delicada, quizás hubiésemos dudado de usar Google y pensado en hacer algo "interno". Pero por lo general es muy fácil incorporar Google Forms a una página de The Guardian y para el usuario es prácticamente invisible el hecho de que estamos usando ese formulario. Por lo que es muy conveniente.

En términos de consejos para periodistas de datos que quieren usar el "crowdsourcing", hay que definir cosas muy específicas para consultar a la gente. En lo posible, haga preguntas tipo "multiple choice" (elegir entre opciones fijas). Trate de conseguir datos demográficos básicos de a quién se dirige, de modo de ver si su muestra puede ser distorsionada. Si está pidiendo cantidades y cosas por el estilo, trate de especificar que requiere la información en dígitos, que tienen que usar una moneda específica, y así. Muchos no lo harán, pero cuanto más los guíe en todo, tanto mejor. Y siempre, siempre, agregue una ventana para comentarios porque mucha gente llenará los otros campos pero lo que realmente quiere es darle su opinión sobre el tema. Especialmente si se trata de algo que tiene que ver con los consumidores o un escándalo.

— *Marianne Bouchart, Data Journalism Blog, interviewing James Ball, the Guardian*

Usar y compartir datos: las reglas técnicas legales, la letra chica y la realidad

En esta sección echaremos un rápido vistazo al estado de las leyes relacionadas con datos y bases de datos, y lo que puede hacer para ofrecer sus datos al público usando licencias comunes y herramientas legales. No deje que nada de lo que sigue ahogue su entusiasmo por el periodismo de datos. Las restricciones al manejo de datos por lo general no serán una traba y fácilmente puede asegurarse de que no sean una traba para otros que usen los datos que usted publica.

Para decir lo obvio, obtener datos nunca fue más fácil. Antes de la publicación generalizada de datos en la red, aunque uno hubiera identificado un conjunto de datos que necesitaba, tenía que pedir a quien tuviera una copia que se la pusiera a disposición, lo que posiblemente involucrara el uso del correo o una visita personal. Ahora uno hace que su computadora le pida a la computadora del otro que le envíe una copia. Conceptualmente es

algo similar, pero usted tiene una copia de inmediato y el otro (el creador o editor) no ha hecho nada, y probablemente no tenga idea de que usted descargó una copia.

¿Y qué pasa cuando se trata de descargar datos con un programa (lo que a veces se llama “scrapear”) y condiciones de uso del servicio (en inglés Terms of Service o ToS)? Considere la frase anterior: su navegador es justamente ese tipo de programa. Puede ser que el ToS solo permita acceso con cierto tipo de programa. Si tiene tiempo y dinero ilimitados para gastar en la lectura de tales documentos y quizás para pedir asesoramiento a un abogado, hágalo sin dudar. Pero por lo general trate de no ser un idiota: si su programa causa daño a un sitio, su red puede ver bloqueado el acceso al sitio en cuestión y quizás usted se lo merezca. Ahora hay mucha experiencia respecto de acceder y "scrapear" datos en la red. Si piensa hacer esto, le será provechoso leer los ejemplos que se dan en sitios como ScraperWiki.

Una vez que tiene datos de interés, puede interrogar, desmenuzar, ordenar, visualizar, correlacionar y realizar cualquier tipo de análisis que guste con su copia de los datos. Puede publicar su análisis, citando cualquier dato. La frase hecha “los datos son libres” (en el mismo sentido que la palabra es libre) dice mucho, o quizás sea solo una frase hecha de los que piensan demasiado en las cuestiones legales relacionadas con las bases de datos o en sentido aún más amplio (y retorcido) el aspecto legal del manejo de datos.

¿Qué sucede si, siendo un periodista de datos bueno o que aspira a ser bueno, tiene la intención de publicar no solo su análisis, incluyendo algunos hechos o datos puntuales, sino también los conjuntos de datos/bases de datos que usó –y a los que quizás incorporó más información- al realizar su análisis? O quizás solo está curando datos y no ha hecho ningún análisis (eso es bueno: el mundo necesita curadores de datos). Si usted está usando datos recopilados por algún otro ente, podría haber alguna complicación. (Si su base de datos ha sido armada totalmente por usted, de todos modos lea el siguiente párrafo como motivación para las prácticas de compartir información que aparecen en el párrafo posterior).

Si usted está familiarizado con el modo en que el copyright limita el trabajo creativo –si el titular del copyright no ha dado permiso para usar un trabajo (o el trabajo está en el dominio público o su uso puede estar cubierto por excepciones y limitaciones tal como el uso leal) y usted usa –distribuye, realiza, etc.- el trabajo de todos modos, el titular del copyright podría obligarlo a interrumpirlo. Aunque los datos son libres, los conjuntos de datos pueden ser restringidos de modo muy similar, aunque hay más variaciones en las leyes relevantes que en el caso del copyright aplicado a obras creativas. En síntesis, una base de datos puede estar sujeta a copyright, como obra creativa. En muchas jurisdicciones, por “el sudor de la frente”, simplemente armar una base de datos, incluso de modo no creativo, hace que la base de datos esté sujeta a copyright. En Estados Unidos en particular, tiende a exigirse un mínimo mayor de creatividad para que haya derecho de autor (Feist v. Rural, un

caso sobre una guía telefónica, es el caso clásico estadounidense si quiere buscarlo). Pero en algunas jurisdicciones también hay "derechos de base de datos" que restringen el uso de bases de datos, como cosa distinta al copyright (aunque hay mucha superposición en términos de lo que está cubierto, en particular donde los umbrales de creatividad para la existencia de copyright son prácticamente inexistentes). Los más conocidos de estos son los derechos de base de datos *sui generis* de la Unión Europea. De nuevo, especialmente si se encuentra en Europa, quizás quiera asegurarse de que tiene autorización antes de publicar una base de datos de otra entidad.

Obviamente tales restricciones no son la mejor manera de promover un ecosistema de periodismo basado en datos (tampoco es algo bueno para la sociedad en general; científicos sociales y otros le dijeron a la UE que no lo serían antes de la aparición de los derechos *sui generis*, y estudios realizados desde su aparición han demostrado que tenían razón). Afortunadamente como editor de una base de datos usted puede eliminar tales restricciones para el uso de la base de datos (suponiendo que no contiene elementos sobre los que usted no tiene autorización para otorgar permiso), esencialmente otorgando permiso por adelantado. Puede hacer esto publicando su base de datos bajo una licencia pública o una dedicatoria al dominio público, del mismo modo que muchos programadores difunden sus códigos bajo una licencia libre y de libre acceso, de modo que otros puedan utilizar su código (dado que el periodismo basado en datos a menudo involucra código, no solo datos, por supuesto que usted debe autorizar el uso de su código también, de modo que su colección de datos y su análisis sean reproducibles). Hay muchos motivos para dar libre acceso a sus datos. Por ejemplo, su público podría crear nuevas visualizaciones o aplicaciones con los mismos y con las que usted puede crear un vínculo, como hace The Guardian con su grupo en Flickr de visualización de datos. Sus conjuntos de datos pueden combinarse con otros conjuntos de datos para que usted y sus lectores tengan una mejor visión de un tema. Las cosas que hacen otros con sus datos pueden darle pistas para nuevas historias, o ideas para historias, o ideas para otros proyectos basados en datos. Y sin duda le dará prestigio.

Figure 14. Distintivos de datos abiertos (Open Knowledge Foundation)

Cuando uno advierte que difundir trabajos bajo licencias públicas es una necesidad, la cuestión pasa a ser: ¿cuál licencia? Esa pregunta complicada frecuentemente será respondida por el proyecto o la comunidad en cuyo trabajo usted basa el suyo, o al que espera poder contribuir con su trabajo: use la licencia que ellos usan. Si necesita investigar más a fondo, empiece por el conjunto de licencias que son libres y abiertas, es decir, que autorizan a cualquiera a darle cualquier uso (puede requerirse tanto libertad de atribución como de compartir). La Definición de Conocimiento Abierto , en español http://es.wikipedia.org/wiki/Conocimiento_abierto, significa para todo otro conocimiento, incluyendo las bases de datos, lo mismo que la Definición de Software Libre y la Definición de Código Libre significan para el software: define lo que hace que una obra sea de libre acceso y lo que las licencias de libre acceso permiten hacer a los usuarios.

Puede visitar el sitio de Open Knowledge Definition para ver el actual conjunto de licencias, algunas definiciones en español en Creative Commons). En síntesis, básicamente hay 3 clases de licencias abiertas:

Dominio Público
: Estas también sirven como licencias de máxima permisividad; no hay condiciones impuestas al uso de la obra.

Licencias permisivas o sólo de atribución
: Reconocer la autoría es la única condición sustancial de estas licencias.

Licencias copyleft, recíprocas o de compartir por igual
: Estas también requieren que si se publican obras modificadas, sean compartidas bajo la misma licencia.

Si usted está usando un conjunto de datos publicados por otro bajo una licencia abierta, considere el párrafo anterior como una breve guía respecto de cómo debe cumplir las condiciones de esa licencia abierta. Las licencias más comunes de Creative Commons, Open data Commons y varios gobiernos por lo general van acompañadas de una síntesis que le permitirá ver fácilmente cuáles son las condiciones sustanciales requeridas. Comúnmente la licencia se presentará en una página de la red de la que puede descargarse un conjunto de datos (o de donde pueden ser "scrapeados", ya que, por supuestos, las páginas de la red pueden contener conjuntos de datos) o en un lugar conspicuo dentro del conjunto de datos mismos, según el formato. Esto es lo que usted debiera hacer también cuando autoriza el acceso a sus conjuntos de datos.

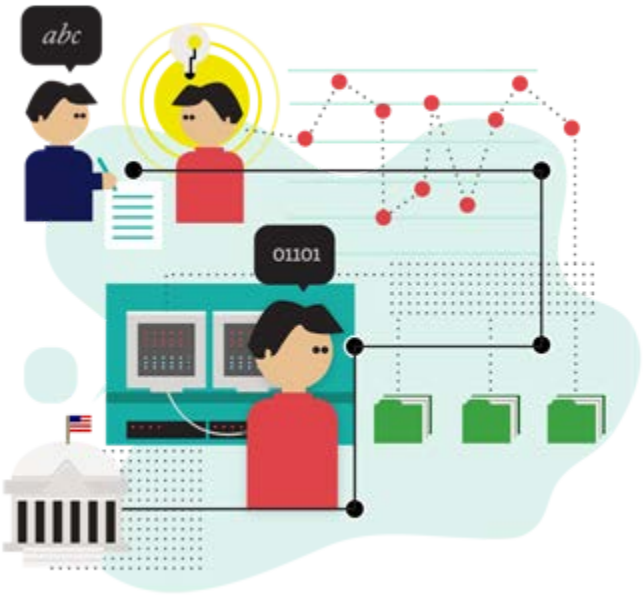
Volviendo al comienzo, ¿qué pasa si el conjunto de datos que necesita no está disponibles online aún o hay algún tipo de control sobre los mismos? Considere la posibilidad de pedir acceso no solo para usted, sino que los datos se abran al uso de todo el mundo. Usted puede dar algunas indicaciones de algunas de las grandes cosas que podrían suceder con esos datos si así se hiciera.

El tema de compartir datos con todo el mundo podría llevar a la cuestión de que algunos conjuntos de datos pueden afectar derechos de privacidad y otras consideraciones y regulaciones. Por cierto, el hecho de que el carácter abierto de la información elimina muchas barreras técnicas y de copyright,, o del tipo del copyright no significa que no haya que cumplir otras leyes. Pero, en el caso de que su sentido común le indique que existe la necesidad de investigar esa cuestión, tenga en cuenta que esto siempre fue así y que hay tremendos recursos y en algunos casos medidas de protección para periodistas.

¡Buena suerte! Pero probablemente necesite la suerte mucho más para otros aspectos de su proyecto que para enfrentar los (escasos) riesgos legales.

— Mike Linksvayer, Creative Commons

Entender los datos



Una vez que tiene sus datos, ¿qué hace con ellos? ¿Qué debe buscar? ¿Qué herramientas debe usar? Esta sección comienza con algunas ideas acerca de cómo mejorar su conocimiento del manejo de datos, consejos para trabajar con cifras y estadísticas, y cosas a tener en cuenta cuando trabaja con conjuntos de datos desordenados, imperfectos y a menudo poco documentados. Podemos luego aprender a obtener historias de los datos, cuáles son las herramientas preferidas de los periodistas de datos, y cómo usar la visualización de datos para que ayude a entender el tópico en cuestión.

Qué contiene este capítulo?

- Aprenda a manejar datos con 3 pasos simples
- Consejos para trabajar con cifras en las noticias
- Pasos básicos para trabajar con datos
- La pieza de pan de £ 32
- Empiece por los datos, termine con una historia
- Historias basadas en datos
- Los periodistas de datos debaten sobre sus herramientas preferidas
- Usar visualizaciones para descubrir cosas en los datos

Aprenda a manejar datos con 3 pasos simples

Así como la alfabetización refiere a “la capacidad de leer para conocer, escribir de modo coherente y pensar críticamente acerca de material impreso”, la alfabetización en materia de datos es la capacidad de manejar datos para conocer, producir coherentemente y pensar críticamente acerca de datos. La alfabetización en materia de datos incluye la alfabetización estadística, pero también comprende cómo trabajar con grandes conjuntos de datos, cómo fueron producidos, como relacionar varios conjuntos de datos y como interpretarlos.

Figure 1. Cavar en los datos (photo by JDHancock)

Poynter News University ofrece clases de matemática para periodistas que ayudan a dominar conceptos tales como cambios porcentuales y promedios. Es interesante que estos conceptos se enseñen simultáneamente cerca de las oficinas de Poynter, en escuelas de Florida a estudiantes de quinto grado (10-11 años), como lo atestigua la currícula.

Que los periodistas necesiten ayuda con temas matemáticos normalmente vistos antes de la escuela secundaria muestra lo lejos que están las redacciones de saber manejar datos. Esto es un problema. ¿Cómo puede una periodista hacer uso de una cantidad de cifras sobre cambio climático si no sabe lo que significa un intervalo de confianza? ¿Cómo puede un periodista de datos escribir una historia sobre distribución del ingreso si no sabe la diferencia entre media y mediana?

Una periodista por cierto no necesita tener un título en estadística para ser más eficiente en el manejo de los datos. Enfrentada a las cifras, unos pocos trucos simples pueden ayudarla a armar una historia mucho mejor. Como dice el profesor del Instituto Max Planck, Gerd Gigerenzer, tener mejores herramientas no permitirá hacer mejor periodismo si éstas no son utilizadas con visión. Aunque no tenga ningún conocimiento de matemática o estadísticas, puede convertirse fácilmente en una periodista de datos experimentada haciendo 3 preguntas muy simples.

1. ¿Cómo se obtuvieron los datos?

Sorprendente crecimiento del PBI

La manera más fácil de darse aires con datos espectaculares es fabricarlos. Suena obvio, pero datos tan comúnmente comentados como las cifras del PBI bien pueden ser falsos. El ex embajador británico Craig Murray informa en su libro, Asesinato en Samarcanda, que las tasas de crecimiento en Uzbekistán están sujetas a intensas negociaciones entre el gobierno local y entes internacionales. Dicho de otro modo, no tienen nada que ver con la economía local.

El PBI es usado como el principal indicador porque los gobiernos tienen que controlar su principal fuente de ingresos: el IVA. Cuando un gobierno no se financia con el IVA, o cuando no informa públicamente de su presupuesto, no tiene motivos para recoger datos sobre el PBI y le vendrá mejor inventarlos.

El crimen siempre está en aumento

“El crimen en España creció un 3%”, escribe El País. Bruselas es presa de un aumento del crimen de extranjeros ilegales y drogadictos, escribe RTL. Este tipo de de informes basados en estadísticas recogidas por la policía es común, pero no nos dice gran cosa sobre la violencia.

Podemos confiar en que dentro de la Unión Europea los datos no son falsificados. Pero el personal policial responde a incentivos. Cuando el desempeño está ligado a la tasa de esclarecimiento, por ejemplo, los policías tienen un incentivo para informar lo más posible de incidentes que no requieren investigación. Uno de tales crímenes es el de fumar marihuana. Esto explica por qué los crímenes relacionados con las drogas en Francia se multiplicaron por 4 en los últimos 15 años, mientras que el consumo se mantuvo constante.

Qué se puede hacer

Cuando dude de la credibilidad de una cifra, verifíquela, tal como lo haría si se tratara de una declaración de un político. En el caso uzbeco, una llamada a alguien que haya vivido allí un tiempo basta ("¿Es cierto que el país es 3 veces más rico que en 1995, como muestran las cifras oficiales?").

Para los datos policiales, los sociólogos a menudo realizan estudios de victmización, en los que preguntan a la gente si es víctima de crímenes. Estos estudios son mucho menos volátiles que los datos policiales. Quizás ese sea el motivo por el que no se los destaca en los medios.

Otros tests permiten evaluar la credibilidad de los datos, tales como la ley de Benford, pero ninguno de ellos suplanta su pensamiento crítico.

2. ¿Qué se puede aprender de ello?

El riesgo de esclerosis múltiple aumenta al doble cuando se trabaja de noche

Sin duda cualquier alemana que no esté loca dejaría de trabajar de noche luego de leer este titular. Pero el artículo no nos dice cuál es el riesgo realmente.

Tome 1000 alemanes. Solo uno tendrá EM. Si todos estos 1000 alemanes trabajaran de noche, el número de pacientes de EM se iría a 2. El riesgo adicional de tener EM trabajando de noche es 1 en 1000, no 100%. Sin duda esta información es más útil al ponderar si aceptar un empleo.

En promedio, 1 de cada 15 europeos es totalmente analfabeto

Este titular asusta. Además es cierto. Entre los 500 millones de europeos, 36 millones probablemente no saben leer. Agreguemos que 36 millones también tienen menos de 7 años; datos de Eurostat.

Cuando escriba sobre un promedio, siempre piense: ¿Un promedio de qué? ¿La población de referencia es homogénea? Los patrones de distribución desigual explican por qué la mayoría de la gente maneja mejor que el promedio, por ejemplo. Mucha gente tiene cero o solo un accidente en toda su vida. Unos pocos conductores irresponsables tienen muchos, lo que hace que el número promedio de accidentes sea mucho más elevado de lo que es la experiencia de la mayoría de la gente. Lo mismo vale para la distribución del ingreso: la mayoría de la gente gana menos que el promedio.

Qué puede hacer

Siempre tome en cuenta la distribución y la tasa base. Verificar el media y la mediana así como la moda (el valor más frecuente en la distribución) le ayuda a interpretar los datos. Conocer el orden de magnitud hace más fácil contextualizar, como en el ejemplo de EM. Finalmente, informar en base a frecuencias naturales (1 de cada 100) es mucho más fácil de entender para los lectores que usar porcentuales (1%).

3. ¿En qué medida es confiable la información?

El problema del tamaño de la muestra

“80% insatisfecho con el sistema judicial”, dice una encuesta de la que se informa en el Diario de Navarra con sede en Zaragoza. ¿Cómo se puede extrapolar de 800 encuestados a 46.000.000 de españoles? Sin duda esto es poco serio.

Cuando se investiga una gran población (más de unos pocos miles) rara vez se necesita más que un millar de encuestados para lograr un margen de error de menos del 3%. Significa que si fuera a rehacer la encuesta con una muestra totalmente distinta, 19 veces de 20 las respuestas que recibiría estarían dentro del intervalo de 3 puntos porcentuales del valor encontrado, comparado con lo que hubiera sucedido si entrevistaba a todas las personas.

Tomar té reduce el riesgo de infarto

Los artículos acerca de los beneficios de tomar té son comunes. Este artículo breve en Die Welt que dice que el té reduce el riesgo de infarto del miocardio no es la excepción. Si bien los efectos del té son estudiados seriamente por algunos, muchas piezas de investigación no toman en cuenta factores de estilo de vida, tales como dieta, ocupación, o deportes.

En la mayoría de los países, el té es la bebida de las clases altas preocupadas por la salud. Si los investigadores no toman en cuenta los factores de estilo de vida en sus estudios sobre el té, no nos dicen más que “los ricos son más sanos y probablemente toman té”.

Lo que puede hacer

La matemática que es la base de las correlaciones y los márgenes de error en los estudios sobre el té es por cierto correcta, al menos la mayoría de las veces. Pero si los investigadores no buscan correlaciones (por ejemplo, tomar té se correlaciona con hacer deporte), sus resultados son de escaso valor. Como periodista, tiene poco sentido cuestionar los resultados numéricos de un estudio, tales como el tamaño de la muestra, a menos que haya

serias dudas al respecto. Sin embargo, es fácil de ver si los investigadores no tomaron en cuenta elementos relevantes de información.

— *Nicolas Kayser-Bril, Journalism++*

Consejos para trabajar con cifras en las noticias

- El mejor consejo para manejar datos es que lo disfrute. Los datos pueden parecer algo intimidantes. Pero si se deja intimidar no llegará a nada. Trátelos como algo para jugar y explorar y a menudo entregarán secretos e historias con sorprendente facilidad. De modo que manéjelos de manera simple, como lo hace con otras evidencias, sin temor ni parcialidad. En particular, piense en esto como un ejercicio de su imaginación. Sea creativo pensando en las historias alternativas que podrían ser coherentes con los datos y los explican mejor, luego póngalas a prueba con más evidencias. "¿Qué otra historia podría explicar esto?", puede ser una buena pregunta para pensar cómo esta cifra evidentemente grande o equivocada, esta clara prueba de esto o aquello, podría no ser nada por el estilo.
- No confunda el escepticismo respecto de los datos con cinismo. El escepticismo es bueno; el cinismo simplemente es darse por vencido. Si cree en el periodismo de datos (y probablemente es así o no estaría leyendo este libro), entonces debe creer que los datos tienen algo mucho mejor que ofrecer que las mentiras de caricatura o los datos de titulares impactantes. Los datos a menudo nos dan conocimiento profundo, si se los usa cuidadosamente. No necesitamos ser cínicos ni ingenuos, sino estar alertas.
- Si le digo que se bebe más durante la recesión, podría decirme que se debe a que todos están deprimidos. Si le digo que se bebe menos, podría decirme que es porque nadie tiene plata. Dicho de otro modo, lo que digan los datos no incide en la interpretación que usted esté decidido a hacer, a saber, que las cosas están muy mal no importa lo que suceda con la bebida. Si aumenta, es malo; si se reduce, es malo. La cuestión aquí es que si usted cree en los datos, trate de dejar que hablen antes de imponerles su propio estado de ánimo, creencias o expectativas. Hay tantos datos que a menudo podría encontrar confirmación de sus creencias previas si busca un poco. Dicho de otro modo, el periodismo de datos, al menos para mí, agrega poco valor si usted no tiene la mente abierta. Es solo objetivo en la medida que usted lo hace objetivo y no en virtud de que se basa en números.
- La incertidumbre no es problema. Asociamos las cifras con la autoridad y la certidumbre. Muy a menudo la respuesta es que no hay respuesta, o la respuesta es la mejor que tenemos pero no es para nada precisa. Creo que debemos decir estas cosas. Si eso suena como una buena manera de matar una historia, sostendría que es una gran manera de generar nuevos interrogantes. Del mismo modo, a menudo puede haber más de un modo legítimo de ordenar los datos. Los números no tienen que ser ciertos o falsos.
- La investigación es una historia. La historia de cómo intentó descubrir algo, al avanzar de un elemento de evidencia a otro, puede ser excelente periodismo y esto se aplica especialmente a la evidencia de los datos, donde rara vez basta con una cifra. Distintas fuentes dan nuevos ángulos de interpretación, nuevas ideas y una comprensión enriquecida. Me pregunto si estamos demasiado preocupados por ganar autoridad y darle la respuesta a la gente, hasta el punto de que desaprovechamos un recurso, que es mostrar nuestra investigación.

- Las mejores preguntas son las de siempre: ¿eso realmente es un número grande? ¿De dónde salió? ¿Está seguro de que cuenta lo que usted cree que cuenta? Estos por lo

> general son solo incentivos para mirar lo que rodea a los datos, las cosas que quedaron de lado por mirar un solo número, las complicaciones de la vida real, la amplia gama de otras comparaciones posibles con relación al tiempo, el grupo o la geografía; en síntesis, el contexto.

— *Michael Blastland, freelance journalist*

Pasos básicos para trabajar con datos

Hay al menos 3 conceptos clave que tiene que entender cuando comience un proyecto de datos:

- Los pedidos de datos deben comenzar con una lista de preguntas que quiere contestar
- Los datos a menudo vienen sucios y hay que limpiarlos
- Los datos pueden tener aspectos sin documentar

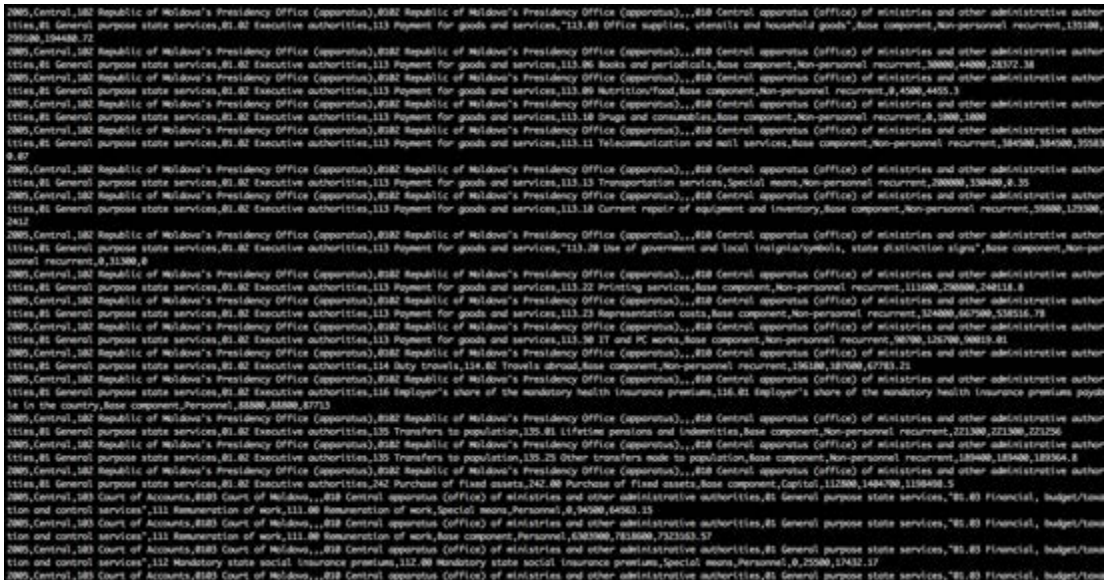


Figure 2. Datos desordenados

Sepa para qué preguntas quiere respuestas

En muchos sentidos, trabajar con datos es como entrevistar una fuente en vivo. Usted le hace preguntas a los datos y logra que revelen las respuestas. Pero así como una fuente solo puede dar respuestas respecto de las cuales tiene información, un conjunto de datos solo puede responder preguntas para las que tiene los registros adecuados y las variables correspondientes. Esto significa que usted debe considerar cuidadosamente qué preguntas quiere responder antes de obtener datos. Básicamente se trabaja hacia atrás. Primero, la lista de afirmaciones basadas en datos que quieres presentar en su historia. Luego decida qué variables y evidencias tendrá que obtener y analizar para hacer esas afirmaciones.

Considere un ejemplo que tiene que ver con los informes de crímenes locales. Digamos que quiere escribir una historia que analice los patrones del crimen en su ciudad, y las cosas que quiere decir tienen que ver con la hora del día y los días de la semana en los que es más común que se den ciertos tipos de crímenes, así como en qué zonas de la ciudad se concentran varias categorías de crímenes.

Tendría que advertir que su pedido de datos tiene que incluir la fecha y el momento en que cada crimen fue informado, el tipo de crimen (asesinato, robo, asalto, etc.), así como la dirección de donde se dio el crimen. De modo que fecha, hora, categoría de crimen y dirección son las variables mínimas que necesita para responder a esas preguntas.

Pero sea consciente de que hay una cantidad de preguntas potencialmente interesantes que este conjunto de datos de 4 variables no puede responder, como la raza y el género de las víctimas, o el valor total de la propiedad robada, o qué agentes son más productivos en cuanto a lograr arrestos. Además quizás solo pueda conseguir los registros de cierto período, como los últimos 3 años, lo que quiere decir que no podría saber si los patrones del crimen han cambiado respecto de un período más prolongado. Esas preguntas pueden quedar por fuera del plan de su historia, y eso no es problema. Pero usted no querrá meterse en su análisis de los datos y de pronto advertir que necesita saber qué porcentaje de los crímenes en distintas zonas de la ciudad son resueltos con arrestos.

Una lección aquí es que a menudo es buena idea pedir todas las variables y registros en la base de datos, en vez del subconjunto que respondería a las preguntas para la historia inmediata. (de hecho, obtener todos los datos puede ser más barato que obtener un subconjunto, si tiene que pagar a la agencia por la programación necesaria para producir el subconjunto.) Siempre puede armar el subconjunto de datos por su cuenta y tener acceso al conjunto de datos completo le permitirá responder nuevas preguntas que pueden surgir durante su trabajo e incluso producir nuevas ideas para la continuación de la historia. Puede ser que las leyes de confidencialidad u otras políticas signifiquen que algunas variables, tales como las identidades de las víctimas o los nombres de informantes confidenciales, no puedan difundirse. Pero incluso una base de datos parcial es mejor que nada, siempre que usted entienda qué preguntas puede o no contestar la base de datos.

Limpiar datos no normalizados

Uno de los mayores problemas en el trabajo con bases de datos es que a menudo usará datos para análisis que han sido recogidos por motivos burocráticos. El problema es que el nivel exigido de precisión es bastante diferente.

Por ejemplo, una función clave de un sistema de bases de datos para la justicia penal es asegurarse que el acusado Jones sea traído de la cárcel al juzgado para estar frente la juez

Smith en el momento de su audiencia. Para ese propósito no importa mucho si la fecha de nacimiento de Jones es incorrecta, o que esté mal escrito el nombre de la calle en la que vive o siquiera si la inicial de su segundo nombre sea equivocada. En general el sistema aún puede usar este registro imperfecto para llevar a Jones al juzgado de Smith a la hora indicada.

Pero tales errores pueden complicar los esfuerzos del periodista por descubrir patrones en la base de datos. Por ese motivo, la primera gran tarea que debe encarar cuando obtiene un nuevo conjunto de datos es examinar hasta donde tiene problemas y solucionarlos. Una manera rápida de buscar datos no normalizados es crear tablas de frecuencias de las variables por categoría, las que uno esperaría que tengan un número relativamente pequeño de valores diferentes. (Cuando use Excel, por ejemplo, usted puede usar Filtros o Tablas Dinámicas para cada variable categórica.)

Tomemos por caso “género”, un ejemplo simple. Usted puede descubrir que el campo de Género incluye cualquier mezcla de valores como estos: Masculino, Femenino, M, F, 1, 0, MASCULINO, FEMENINO, etc., incluyendo errores de escritura como “Femeno”. Para hacer un análisis de género adecuado debe estandarizar – quizás se decida por M y F- y luego cambiar todas las variaciones para que coincidan con los estándares. Otra base de datos común con este tipo de problemas es la de los registros financieros de campañas electorales de Estados Unidos, donde en el campo de Ocupación puede dar las distintas variantes de abogado en inglés (Lawyer, Attorney, Atty, Counsel, Trial lawyer y muchas otras) además de los errores de escritura; nuevamente el truco es estandarizar los títulos de ocupación en una lista de posibilidades más corta.

La limpieza de los datos se vuelve aún más problemática cuando se trabaja con nombres. ¿“Joseph T. Smith”, “Joseph Smith”, “J.T. Smith”, “Jos. Smith” y “Joe Smith” son todos la misma persona? Quizás haya que mirar otras variables como dirección o fecha de nacimiento, o hacer una investigación aún más profunda en otros registros, para decidir. Pero herramientas como Google Refine pueden hacer que la limpieza y estandarización sean más rápidas y menos tediosas.

Datos sucios

Gracias a las leyes de registro público por lo general fuertes en Estados Unidos, obtener datos aquí no es un problema tan grande como en muchos otros países. Pero una vez que los tenemos, aún quedan los problemas de trabajar con datos que han sido recogidos por motivos burocráticos y no con propósitos analíticos. Los datos a menudo vienen “sucios”, con valores que no están estandarizados. Varias veces he recibido datos que no se corresponden con el supuesto diagrama del archivo y el diccionario de datos que los acompañan. Algunas entidades insistirán en darle los datos en formatos poco útiles como

.PDF, que tienen que ser convertidos. Problemas como estos hacen que uno lo aprecie cuando ocasionalmente recibe un conjunto de datos sin complicaciones.

— Steve Doig, Walter Cronkite School of Journalism, Arizona State University

Los datos pueden tener aspectos no documentados

La Piedra de Roseta de cualquier base de datos es el llamado diccionario de datos. Comúnmente, este archivo (puede ser texto PDF o incluso una hoja de cálculo) le dirá cómo está formateado el archivo de los datos (texto delimitado, texto de ancho fijo, Excel, dBase, etc.), el orden de las variables, los nombres de cada variable y el tipo de datos de cada variable (hilo de texto, entero, decimal, etc.). Usted usará esta información para que lo ayude a importar adecuadamente el archivo de datos al software de análisis que piensa usar (Excel, Access, SPSS, Fusion Tables, distintas variantes de SQL, etc.)

El otro elemento clave de un diccionario de datos es una explicación de los códigos que puedan usar variables particulares. Por ejemplo, género puede estar codificado de tal modo que “1=Masculino” y “0=Femenino”. Los crímenes pueden estar codificados de acuerdo a los números estatutarios de su jurisdicción para cada tipo de crimen. Los registros de tratamientos hospitalarios pueden usar cualquiera de los cientos de códigos de 5 dígitos existentes para el diagnóstico de las enfermedades por las que está tratando a un paciente. Sin el diccionario de datos, estos conjuntos de datos serían difíciles o incluso imposibles de analizar adecuadamente.

Pero incluso contando con un diccionario de datos puede haber problemas. Un ejemplo de tales problemas es lo que le sucedió a periodistas del Miami Herald en Florida hace algunos años, cuando estaban haciendo el análisis de los variados castigos que distintos jueces estaban imponiendo a gente arrestada por manejar ebria e intoxicada. Los periodistas obtuvieron los registros de condenas del sistema judicial y analizaron las cifras con las 3 variables distintas de castigos en el diccionario de datos: cantidad de tiempo en prisión, cantidad de tiempo detenido y cantidad de multa. Las cifras variaban bastante entre los jueces, dando a los periodistas evidencias para una historia acerca de que algunos jueces eran duros y otros más permisivos.

Pero para todos los jueces, en alrededor del 1-2 por ciento de los casos no había tiempo de prisión, ni de detención, ni multa. Por lo que el cuadro que mostraba los patrones de condenas de cada juez incluía una cantidad pequeña de casos como “Sin castigo”, casi como una nota al margen. Cuando la historia y el cuadro se publicaron, los jueces aullaron de indignación, diciendo que el Herald los acusaba de violar una ley estatal que exige que cualquiera condenado por manejar borracho sufra castigo.

Por lo que los periodistas volvieron a la oficina del empleado de la corte que les había dado el archivo de datos y le preguntaron qué era lo que había causado el error. Se les dijo que los casos en cuestión involucraban a acusados indigentes que eran arrestados por primera vez. Normalmente se les hubiera impuesto una multa pero no tenían dinero. Por lo que los jueces los condenaban a servicios comunitarios, tales como limpiar la basura en los caminos. Resultó que la ley que requería el castigo había sido aprobada después de que fuera creada la estructura de la base de datos. Por lo que los empleados de la corte sabían que en los datos los ceros en las 3 variables de prisión-detención-multa significaban servicio comunitario. Sin embargo, esto no aparecía en el diccionario de datos y por tanto el Herald tuvo que publicar la correspondiente rectificación.

La lección en este caso es que siempre hay que preguntar al ente que le da los datos si hay elementos no documentados de los mismos, lo que podría significar códigos nuevos que no están incluidos en el diccionario de datos, cambios en el ordenamiento del archivo, o cualquier otra cosa. Además examine siempre los resultados de su análisis y pregúntese: "¿Esto tiene sentido?" Los periodistas del Herald estaban armando el cuadro apurados por el plazo de entrega y estaban tan concentrados en los niveles de castigo promedio de cada juez, que no prestaron atención a los pocos casos que parecían no tener castigo. Debieron haberse preguntado si tenía sentido que todos los jueces aparentemente estuvieran violando la ley estatal, aunque más no fuera en mínima medida.

— Steve Doig, Walter Cronkite School of Journalism, Arizona State University

Datos mezclados, ocultos y faltantes

Recuerdo una situación graciosa en la que tratamos de acceder a los datos de Hungría sobre subsidios agropecuarios de la UE: estaban todos allí, pero en un documento PDF excesivamente pesado y mezclado con datos sobre subsidios agropecuarios nacionales. Nuestros programadores tuvieron que trabajar horas antes de que los datos pudieran ser utilizados.

También tuvimos una experiencia bastante interesante con datos sobre subsidios de pesca de la UE, que los entes nacionales encargados de los pagos de los 27 estados miembros están obligados a dar a conocer. Esto está tomado de un informe que escribimos sobre el tema: "En el Reino Unido, por ejemplo, el formato de los datos varía de páginas de búsqueda HTML muy fáciles de usar hasta resúmenes en PDF o incluso listas de receptores en formatos variados disimulados al final de declaraciones de prensa. Todo esto de un solo estado miembro. Mientras tanto, en Alemania y Bulgaria se publican listas vacías. Tienen los encabezados apropiados, pero sin datos".

— Brigitte Alfter, Journalismfund.eu

La pieza de pan de £ 32

Una historia para el Wales On Sunday acerca de cuánto gasta el gobierno galés en órdenes para productos libre de gluten contenía un titular que indicaba que pagaba £32 por una pieza de pan. Pero en realidad eran 11 piezas que costaban £2,82 cada una.

Los datos, tomados de una respuesta por escrito de la Legislatura Galesa y un informe estadístico del Servicio de Salud de Gales, estaban presentados con el formato del costo por cada ítem de las órdenes. Sin embargo en el diccionario de datos no daban ninguna definición adicional de lo que podría ser un ítem de orden o cómo podría definirlo una columna de cantidades por separado.

La suposición era que se refería a un ítem individual –es decir, una pieza de pan- en vez de lo que era en realidad, varias piezas.

Nadie, ni la gente que dio la respuesta por escrito ni la oficina de prensa, plantearon la cuestión de la cantidad hasta el lunes posterior a la publicación de la historia.

Por lo que no debe dar por supuesto que las notas de soporte de los datos oficiales ayudarán a explicar qué información se presenta, o que la gente responsable de los datos advertirá que la información no es clara, incluso cuando usted les presente una suposición equivocada.

Por lo general los diarios quieren cosas que produzcan buenos titulares, de modo que, a menos que algo contradiga evidentemente una interpretación, por lo general es más fácil aceptar lo que permite hacer un buen titular y no investigar demasiado, con el riesgo de que se hunda la historia, especialmente a la hora del cierre.

Figure 3. Las órdenes de pan libre de gluten le cuestan a los contribuyentes (WalesOnline)

Pero los periodistas tienen la responsabilidad de verificar las afirmaciones ridículas, aunque signifique que esto hace caer la noticia.

— Claire Miller, WalesOnline

Empiece por los datos, termine con una historia

Para atraer a sus lectores tiene que poder darles una cifra en los titulares que los haga prestar atención. Casi se debiera poder leer la historia sin tener que saber que se basa en un conjunto de datos. Hágala interesante y recuerde quién es su público.

Un ejemplo de esto puede encontrarse en un proyecto del Bureau of Investigative Journalism que utiliza el Sistema de Transparencia Financiera de la Comisión de la UE. La historia se construyó con el conjunto de datos teniendo en mente interrogantes específicos.

Investigamos en los datos con términos clave tales como “coctel”, “golf” y “días de descanso”. Esto nos permitió establecer lo que la Comisión había gastado en estos ítems y esto planteó numerosas preguntas e historias para seguir.

Pero los términos clave no siempre le dan lo que quiere; a veces tiene que tomarse un respiro y pensar qué es realmente lo que busca. Durante este proyecto también queríamos descubrir cuánto gastan los comisionados en viajes en jet privado pero como el conjunto de datos no contenía la frase “jet privado” tuvimos que obtener el nombre de sus proveedores de viajes por otros medios. Una vez que tuvimos el nombre del proveedor de servicios de la Comisión, “Abelag”, pudimos buscar en los datos cuánto se estaba gastando en servicios provistos por Abelag.

Con este enfoque teníamos un objetivo claramente definido para investigar con los datos: encontrar una cifra que pudiera proveer un titular; el colorido de la noticia siguió a ello.

Otro enfoque es comenzar con una lista negra y buscar exclusiones. ¿Una manera fácil de encontrar historias en los datos es saber qué cosas no debiera encontrar allí! Un buen ejemplo de cómo esto puede funcionar es ilustrado por el proyecto en colaboración de Fondos Estructurales de la UE, entre el Financial Times y el Bureau of Investigative Journalism.

Investigamos los datos basándonos en las reglas de la Comisión respecto de qué compañías y asociaciones no deben recibir fondos estructurales. Un ejemplo era el gasto en tabaco y productores de tabaco.

Investigando los datos con los nombres de las compañías, productores y cultivadores de tabaco, encontramos datos que revelaron que British American Tabaco estaba recibiendo € 1.500.000 para una planta en Alemania.

Dado que esa financiación violaba las normas de gastos de la Comisión, fue una manera rápida de encontrar una historia en los datos.

Nunca se sabe lo que uno puede encontrar en un conjunto de datos, así que eche una mirada. Hay que ser bastante audaz y este enfoque funciona mejor por lo general cuando se trata de identificar características evidentes que se verán con el filtrado (los mayores, los extremos, los más comunes, etc.).

— *Caelainn Barr, Citywire*

Historias basadas en datos

El periodismo de datos a veces puede dar la impresión que principalmente se trata de la presentación de los datos, tales como visualizaciones que son instrumentos poderosos que permiten comprender rápidamente algún aspecto de las cifras, o bases de datos interactivas que permiten a los individuos buscar lugares como su propia calle o un hospital. Todo esto puede ser muy valioso, pero al igual que otras formas de periodismo, el periodismo de datos también debe ser sobre historias. ¿Qué tipos de historias pueden encontrarse en los datos? Basándome en mi experiencia en la BBC he armado una lista o "tipología" de distintos tipos de historias basadas en datos.

Creo que es útil tener en cuenta esta lista, no solo cuando analiza datos, sino también en la fase previa, cuando los está buscando (sean datos a disposición del público o los que exigen presentar pedidos de acceso a la información).

Medición

: La historia simple; contar o hacer el total: "Los consejos municipales de todo el país gastaron un total de $x miles de millones en broches de papel el año pasado". Pero a menudo es difícil saber si eso es mucho o poco. Para eso se necesita contexto, lo que puede ser aportado por:

Proporción

: "El año pasado los consejos municipales gastaron 2/3 de su presupuesto de librería en broches de papel"

Comparación interna

: "Los consejos municipales gastan más en broches para papel que en proveer comidas para personas mayores".

Comparación externa

> “El gasto de los consejos en broches de papel el año pasado fue el doble del presupuesto de la nación de ayuda a otros países”.

También hay otras maneras de explorar los datos de un modo contextual o comparativo:

Cambio a lo largo del tiempo

> “El gasto de los consejos en broches para papel se ha triplicado en los últimos 4 años”.

“Tablas comparativas”

> Estas a menudo son geográficas o por institución, y debe asegurarse de que la base de comparación sea justa (por ejemplo, que tome en cuenta el tamaño de la población local). “El Consejo de Borsetshire gasta más en broches para papel por cada miembro del personal que cualquier otra municipalidad, con una tasa de 4 veces el promedio nacional”.

O puede dividir los temas de los datos en grupos:

Análisis por categorías

> “Los consejos dirigidos por el Partido Violeta gastan 50% más en broches de papel que los controlados por el Partido Amarillo”.

O puede relacionar los factores numéricamente:

Asociación

> “Los consejos dirigidos por políticos que han recibido aportes de campaña de compañías de productos de librería gastan más en broches de papel, con el gasto aumentando en promedio £ 100 por cada libra aportada en la campaña”.

Pero, por supuesto, recuerde que correlación y causa no son la misma cosa.

De modo que si está investigando el gasto en broches de papel, ¿está obteniendo también las siguientes cifras?

- Gasto total para dar contexto
- Referencias geográficas/ históricas/de otro tipo para poder dar datos comparativos
- Los datos adicionales que necesita para asegurarse de que las comparaciones son justas, tales como el tamaño de la población.
- Otros datos que podrían facilitar un análisis interesante o con los cuales comparar o relacionar el gasto.

— *Martin Rosenbaum, BBC*

Historias basadas en datos

El periodismo de datos a veces puede dar la impresión que principalmente se trata de la presentación de los datos, tales como visualizaciones que son instrumentos poderosos que

permiten comprender rápidamente algún aspecto de las cifras, o bases de datos interactivas que permiten a los individuos buscar lugares como su propia calle o un hospital. Todo esto puede ser muy valioso, pero al igual que otras formas de periodismo, el periodismo de datos también debe ser sobre historias. ¿Qué tipos de historias pueden encontrarse en los datos? Basándome en mi experiencia en la BBC he armado una lista o "tipología" de distintos tipos de historias basadas en datos.

Creo que es útil tener en cuenta esta lista, no solo cuando analiza datos, sino también en la fase previa, cuando los está buscando (sean datos a disposición del público o los que exigen presentar pedidos de acceso a la información).

Medición

> La historia simple; contar o hacer el total: "Los consejos municipales de todo el país gastaron un total de $x miles de millones en broches de papel el año pasado". Pero a menudo es difícil saber si eso es mucho o poco. Para eso se necesita contexto, lo que puede ser aportado por:

Proporción

> "El año pasado los consejos municipales gastaron 2/3 de su presupuesto de librería en broches de papel"

Comparación interna

> "Los consejos municipales gastan más en broches para papel que en proveer comidas para personas mayores".

Comparación externa

> "El gasto de los consejos en broches de papel el año pasado fue el doble del presupuesto de la nación de ayuda a otros países".

También hay otras maneras de explorar los datos de un modo contextual o comparativo:

Cambio a lo largo del tiempo

> "El gasto de los consejos en broches para papel se ha triplicado en los últimos 4 años".

"Tablas comparativas"

> Estas a menudo son geográficas o por institución, y debe asegurarse de que la base de comparación sea justa (por ejemplo, que tome en cuenta el tamaño de la población local). "El Consejo de Borsetshire gasta más en broches para papel por cada miembro del personal que cualquier otra municipalidad, con una tasa de 4 veces el promedio nacional".

O puede dividir los temas de los datos en grupos:

Análisis por categorías

> "Los consejos dirigidos por el Partido Violeta gastan 50% más en broches de papel que los controlados por el Partido Amarillo".

O puede relacionar los factores numéricamente:

Asociación

> "Los consejos dirigidos por políticos que han recibido aportes de campaña de compañías de productos de librería gastan más en broches de papel, con el gasto aumentando en promedio £ 100 por cada libra aportada en la campaña".

Pero, por supuesto, recuerde que correlación y causa no son la misma cosa.

De modo que si está investigando el gasto en broches de papel, ¿está obteniendo también las siguientes cifras?

- Gasto total para dar contexto
- Referencias geográficas/ históricas/de otro tipo para poder dar datos comparativos
- Los datos adicionales que necesita para asegurarse de que las comparaciones son justas, tales como el tamaño de la población.
- Otros datos que podrían facilitar un análisis interesante o con los cuales comparar o relacionar el gasto.

— *Martin Rosenbaum, BBC*

Los periodistas de datos debaten sobre sus herramientas preferidas

Sssssss. Es el sonido de sus datos descomprimiéndose al abrirse su envoltorio al vacío. ¿Y ahora qué? ¿Qué busca? ¿Y qué herramientas usa? Pedimos a periodistas de datos que nos contaran un poco de cómo trabajan con datos. Esto es lo que nos dijeron:

En el Datablog de The Guardian nos gusta interactuar con nuestros lectores y permitirles replicar nuestro periodismo de datos rápidamente significa que pueden desarrollar el trabajo que hacemos y a veces ver cosas que se nos pasaron. Por lo que cuanto más intuitivas son las herramientas de datos mejor. Tratamos de elegir herramientas que cualquiera pueda manejar sin tener que aprender un lenguaje de programación o que requieran fuerte capacitación a un costo elevado.

Por este motivo actualmente usamos mucho productos de Google. Todos los conjuntos de datos que ordenamos y difundimos aparecen como Google Fusion Tables, lo que significa que gente que tenga una cuenta de Google puede descargar los datos, importarlos a su propia cuenta y hacer sus propios cuadros, ordenar los datos y crear tablas comparativas, o pueden importar los datos a la herramienta que prefieran.

Para mapear los datos usamosa Google Fusion Tables. Cuando creamos mapas de calor en Fusion, compartimos nuestros archivos KML de modo que los lectores puedan descargar y crear sus propios mapas de calor –quizás agregando más capas de datos al mapa original del Datablog. El otro aspecto positivo de estas herramientas de Google es que funcionan con

las muchas plataformas que usan nuestros lectores para acceder al blog, incluyendo PC, celulares y tabletas.

Además de las de Google Spreadsheets y Google Fusion Tables, usamos otras dos herramientas en nuestro trabajo cotidiano. La primera es Tableau, para visualizar conjuntos de datos multidimensionales; y la segunda es ManyEyes, para un análisis rápido de datos. Ninguna de estas herramientas es perfecta, por lo que seguimos buscando mejores herramientas de visualización que nuestros lectores puedan disfrutar.

The Guardian — Lisa Evans

¿Llegaré a ser programador alguna vez? ¡Es muy improbable! Por cierto que no creo que todos los periodistas tengan que saber programar. Pero sí creo que es muy valioso que todos tengan una conciencia general de qué cosas son posibles y cómo hablar con programadores.

Si está recién comenzando, camine, no corra. Tiene que persuadir a sus colegas y editores que trabajar con datos le puede permitir conseguir historias que de otro modo no tendría y que valen la pena. Cuando adviertan el valor de este enfoque, puede comenzar a hacer historias y proyectos más complejos.

Mi consejo es aprender Excel y hacer algunas historias simples primero. Comience por cosas pequeñas y vaya recorriendo el camino hasta el análisis y mapeo de bases de datos. Se puede hacer tanto en Excel; es una herramienta extremadamente poderosa y la mayoría de la gente no usa siquiera una mínima parte de su funcionalidad. Si puede haga un curso de Excel para periodistas, tales como los que ofrece el Centre for Investigative Journalism.

Con respecto a interpretar datos: no lo tome a la ligera. Tiene que ser detallista. Preste atención a los detalles y cuestione sus resultados. Tome notas de cómo procesa los datos y guarde una copia de los datos originales. Es fácil cometer un error. Siempre hago mi análisis 2 o 3 veces prácticamente desde cero. Incluso mejor sería conseguir que su editor u otra persona analice los datos por su cuenta y compare los resultados.

Financial Times — Cynthia O'Murchu

La capacidad de escribir, instalar y ejecutar software complejo tan rápido como un periodista puede escribir una historia es algo bastante nuevo. Antes llevaba mucho más tiempo. Las cosas cambiaron gracias al desarrollo de bases de desarrollo rápido de código abierto: Django y Ruby on Rails; ambos se conocieron a mediados de la década del 2000.

Django, que está construido sobre el lenguaje de programación Python, fue desarrollado por Adrian Holovaty y un equipo que trabajaba en una redacción, el Lawrence Journal-World en Lawrence, Kansas. Ruby on Rails fue desarrollado en Chicago por David Heinemeier Hansson y 37Signals, una compañía de aplicaciones para la red.

Si bien estas plataformas tienen enfoques diferentes del “patrón MVC”, ambas son excelentes y hacen posible crear aplicaciones para la red rápidamente, incluso muy

complejas. Eliminan parte del trabajo rudimentario en la creación de una aplicación. Cosas como crear y buscar ítems de la base de datos, y hacer corresponder URL con códigos específicos en una aplicación, están incorporados a esas plataformas, por lo que los diseñadores no necesitan escribir programas o hacer cosas básicas como esas.

El desarrollo de servicios de provisión de espacio en servidores rápidos de la red como los Amazon Web Services eliminaron parte de lo que hacía del lanzamiento de una aplicación un proceso lento.

Aparte de eso, usamos herramientas bastante estándar para el trabajo con datos: Google Refine y Microsoft Excel para limpiar los datos; SPSS y R para hacer estadísticas; ArcGIS y QGIS para hacer GIS; Git para el manejo de códigos fuente; TextMate, Vim y Sublime Text para escribir código; y una mezcla de MySql, PostgreSQL y SQL Server para bases de datos. Creamos nuestra propia plataforma de JavaScript llamada "Glass" que nos ayuda a crear aplicaciones para usuarios pesadas en JavaScript muy rápidamente.

ProPublica — Scott Klein

A veces la mejor herramienta es la más simple, es fácil subestimar el poder de una planilla de cálculo. Pero usar una planilla de cálculo en los tiempos en que todo funcionaba con DOS me permitió entender una fórmula compleja del acuerdo de asociación de los dueños de los Texas Rangers, cuando George W. Bush era uno de los propietarios claves. Una planilla de cálculo me permite descubrir datos importantes o errores en cálculos. Puedo escribir líneas de código en algún lenguaje de programación (script) para limpieza, normalización y más. Es un elemento básico del set de herramientas del periodista de datos.

Dicho eso, mis herramientas favoritas son aún más poderosas: SPSS para análisis estadístico y mapear programas que me permiten ver patrones geográficos.

The Seattle Times — Cheryl Phillips

Soy fanático de Python. Es un lenguaje de programación de código abierto maravilloso que es fácil de leer y escribir (por ejemplo, no hay que escribir un punto y coma después de cada línea). Lo que es más importante, Python tiene una base tremenda de usuarios y por tanto tiene plugins (llamados paquetes) para todo lo que uno necesite.

Considero que Django es algo que los periodistas de datos rara vez necesitan. Es una plataforma basada en Python para aplicaciones en la red, es decir una herramienta para crear aplicaciones grandes en la red con bases de datos. Decididamente es demasiado pesado para infografías interactivas pequeñas.

También uso QGis, que es una herramienta de código abierto con una gran variedad de funciones GIS, que son necesarias para periodistas de datos que de vez en cuanto tienen que manejar datos geográficos. Si necesita convertir datos geo-espaciales de un formato a otro, entonces QGis es lo que necesita. Puede manejar casi cualquier formato de geo-datos que exista (Shapefiles, KML, GeoJSON, etc.). Si necesita recortar unas cuantas regiones, QGis

también puede hacerlo. Además hay una inmensa comunidad en torno de QGis, por lo que hay toneladas de recursos como tutoriales en la red.

R fue creada principalmente como herramienta de visualización científica. Es difícil encontrar un método de visualización o técnica de manejo de datos que no esté incorporado a R. R es un universo en sí mismo, la meca del análisis visual de datos. Una contra es que hay que aprender otro lenguaje de programación, ya que R tiene su propio lenguaje. Pero una vez que superó los primeros pasos en la curva de aprendizaje, no hay herramienta más poderosa que R. Los periodistas de datos capacitados pueden usar R para analizar conjuntos de datos inmensos que extienden los limites de Excel (por ejemplo, si tiene una tabla con un millón de filas).

Lo realmente lindo de R es que se puede tener un "protocolo" exacto de lo que está haciendo con los datos durante todo el proceso, desde la lectura de un archivo CSV a generar cuadros. Si los datos cambian puede regenerar el cuadro usando un clic. Si alguien tiene curiosidad respecto de la integridad de su cuadro, puede mostrarle la fuente exacta, lo que permite a cualquiera recrear el mismo cuadro por su cuenta (o quizás encontrar los errores que usted cometió).

NumPy + MatPlotLIb es una manera de hacer lo mismo en Python. Es una opción si ya está capacitado en Python. De hecho, NumPy y MatPlotLIb son dos ejemplos de paquetes de Python. Pueden ser usados para análisis y visualización de datos y los dos se limitan a visualizaciones estáticas. No pueden usarse para crear cuadros interactivos con consejos sobre el manejo de herramientas y cosas más avanzadas.

Yo no uso MapBox, pero supe que es una gran herramienta si se quiere presentar mapas más sofisticados basados en OpenStreetMap. Permite por ejemplo adecuar los estilos del mapa (colores, etiquetas, etc.). También hay un acompañante de MapBox, llamado Leaflet. Es básicamente una biblioteca de JavaScript de más alto nivel para mapear que le permite pasar de un proveedor de mapas a otro fácilmente (OSM, MapBox, Google Maps, Bing, etc.).

RaphaelJS es una biblioteca de visualización más bien de bajo nivel que le permite trabajar con elementos primitivos (como círculos, líneas, texto) y animarlos, agregar interacciones, etc. No contiene nada parecido a un cuadro de barras listo para usar, por lo que usted mismo tiene que dibujar un conjunto de rectángulos.

Sin embargo, lo bueno de Raphael es que todo lo que crea funciona también en Internet Explorer. Eso no sucede con muchas otras bibliotecas de visualización (asombrosas) como D3. Lamentablemente, tantos usuarios siguen usando IE y ninguna redacción puede darse le lujo de ignorar al 30% de sus usuarios.

Además de RaphaelJS, también está la opción de crear una alternativa en Flash para IE. Es básicamente lo que está haciendo el New York Times. Eso significa que tiene que desarrollar cada aplicación dos veces.

Aún no estoy convencido de cuál es el "mejor" proceso para crear visualizaciones para IE y navegadores modernos. A menudo resulta que las aplicaciones creadas con RapahelJS funcionan muy lentas en IE, alrededor de 10 veces más lentas que con Flash usando navegadores modernos. Por lo que las alternativas en Flash pueden ser mejor opción si quiere ofrecer visualizaciones animadas de alta calidad para todos los usuarios.

Open Knowledge Foundation — Gregor Aisch

Mi herramienta preferida es Excel, que puede manejar la mayoría de los problemas CAR (periodismo asistido por computadoras) y tiene las ventajas de ser fácil de aprender y estar disponible para la mayoría de los periodistas. Cuando necesito fusionar tablas, comúnmente uso Access, pero luego exporto la tabla fusionada de nuevo a Excel para más trabajo. Uso el ArcMap de ESRI para análisis geográficos; es poderoso y es utilizado por las agencias que recopilan datos geo-codificados. TextWrangler es muy bueno para examinar datos de texto con diseños y delimitadores complicados, y puede hacer búsqueda y reemplazo sofisticada con expresiones regulares. Cuando se necesita técnicas estadística, como regresión lineal, uso SPSS; tiene un menú para señalar y cliquear fácil de usar. Para trabajos realmente pesados, como las tareas con conjuntos de datos que tienen millones de registros que necesitan un importante filtrado y transformaciones de variables programadas, uso software SAS.

Walter Cronkite School of Journalism — Steve Doig

Entre nuestras herramientas preferidas se incluyen Python y Django para hackear, scrapear y jugar con datos; y PostGIS, QGIS y las herramientas de MapBow para crear mapas locos en la red. R y MumPy + MatPlotLib actualmente disputan la supremacía como nuestro equipo de trabajo para análisis de datos exploratorio, aunque últimamente nuestra herramienta de datos preferida es de nuestra propia cosecha: CSVKit. Hacemos casi todo en la nube.

Chicago Tribune — Brian Boyer

En La Nación usamos:

- Excel para limpiar, organizar y analizar datos,
- Google Spreadsheets para edición y conexión con servicios tales como Google Fusion Tables y la Junar Open Data Platform,
- Junar para compartir nuestros datos e incrustarlos en nuestros artículos y actualizaciones del blog,
- Tableau Public para nuestras visualizaciones de datos interactivas,
- Qlikview, una herramienta de inteligencia para empresas muy rápida para analizar y filtrar conjuntos de datos grandes,

- NitroPDF para convertir PDF a archivos de texto y Excel,
- Google Fusion Tables para visualizaciones de mapas.

La Nacion (Argentina) — Angélica Peralta Ramos

Como comunidad de base sin preferencias técnicas, en Transparency Hackers usamos muchas herramientas y lenguajes de programación diferentes. Cada miembro tiene su propio conjunto de preferencias y esta gran variedad es al mismo tiempo nuestro punto fuerte y nuestra debilidad. Algunos estamos construyendo una "Versión de Linux para Hackers de Transparencia", que podamos iniciar en cualquier parte para hackear datos. Este recurso tiene algunas herramientas y bibliotecas interesantes para manejar datos como Refine, RStudio y OpenOffice Calc (por lo general una herramienta poco usada por la gente que conoce del tema, pero realmente útil para cosas rápidas/pequeñas). También hemos estado usando ScraperWiki mucho para hacer prototipos rápidamente y guardar resultados de datos online.

Hay muchas herramientas que nos gustan para visualizaciones de datos y gráficos. Python y NumPy son bastante poderosas. Alguna gente de la comunidad ha estado jugando con R, pero en definitiva las bibliotecas para ploteado de gráficos, como D3, Flot, y RaphaelJS es lo que se termina usando en la mayoría de nuestros proyectos. Finalmente, hemos estado experimentando mucho con mapeado, y Tilemill ha sido una herramienta muy interesante para este trabajo.

Transparência Hacker — Pedro Markun

Usar visualizaciones para descubrir cosas en los datos

La visualización es crítica para el análisis de datos. Aporta una primera línea de ataque, revelando estructuras intrincadas en datos que no pueden ser absorbidas de otro modo. Descubrimos efectos inimaginados y cuestionamos aquellos que han sido imaginados.

Hobart Press) — William S. Cleveland (de Visualizing Data

Los datos por sí mismos, que consisten de bits y bytes almacenados en un archivo en el disco rígido de una computadora, son invisibles. Para poder verlos y encontrarles sentido, necesitamos visualizarlos. En esta sección voy a usar el término visualizar en un sentido más amplio, que incluye incluso representaciones textuales puras de datos. Por ejemplo, simplemente cargar un conjunto de datos en un software de planilla de cálculo puede considerarse una visualización de datos. Los datos invisibles de pronto se convierten en una "imagen" visible en nuestra pantalla. Por tanto, la pregunta no debe ser si los periodistas necesitan visualizar los datos o no, sino qué tipo de visualización puede ser la más útil en cada situación.

Dicho de otro modo: ¿cuándo tiene sentido ir más allá de la visualización en tablas? La respuesta más simple es: casi siempre. Las tablas por sí solas decididamente no bastan para darnos una visión general de un conjunto de datos. Y las tablas por sí solas no nos permiten identificar inmediatamente patrones dentro de los datos. El ejemplo más común aquí son los patrones geográficos que solo pueden observarse al visualizar datos en un mapa. Pero también hay otros tipos de patrones, que veremos luego en esta sección.

Usar visualización de datos para descubrir información clarificadora

No es realista esperar que herramientas y técnicas de visualización de datos disparen una andanada de historias listas para usar a partir de los conjuntos de datos. No hay reglas ni "protocolos" que nos garanticen que tendremos una historia. En cambio, creo que tiene más sentido buscar "percepciones", que un buen periodista puede incorporar a historias.

Cada nueva visualización puede darnos percepciones sobre nuestros datos. Parte de esa información reveladora puede ser conocida ya (pero quizás aún no demostrada), mientras que otros aspectos pueden resultarnos completamente nuevos o incluso sorprendentes. Algunas cosas nuevas que percibimos podrían significar el comienzo de una historia, mientras que otras podrían ser simplemente el resultado de errores en los datos, que es más probable que encontremos visualizando los datos.

Para hacer más efectiva la búsqueda de nuevas percepciones en los datos, me resulta de gran ayuda el proceso representado en Figure 4 (y descripto en el resto de esta sección).

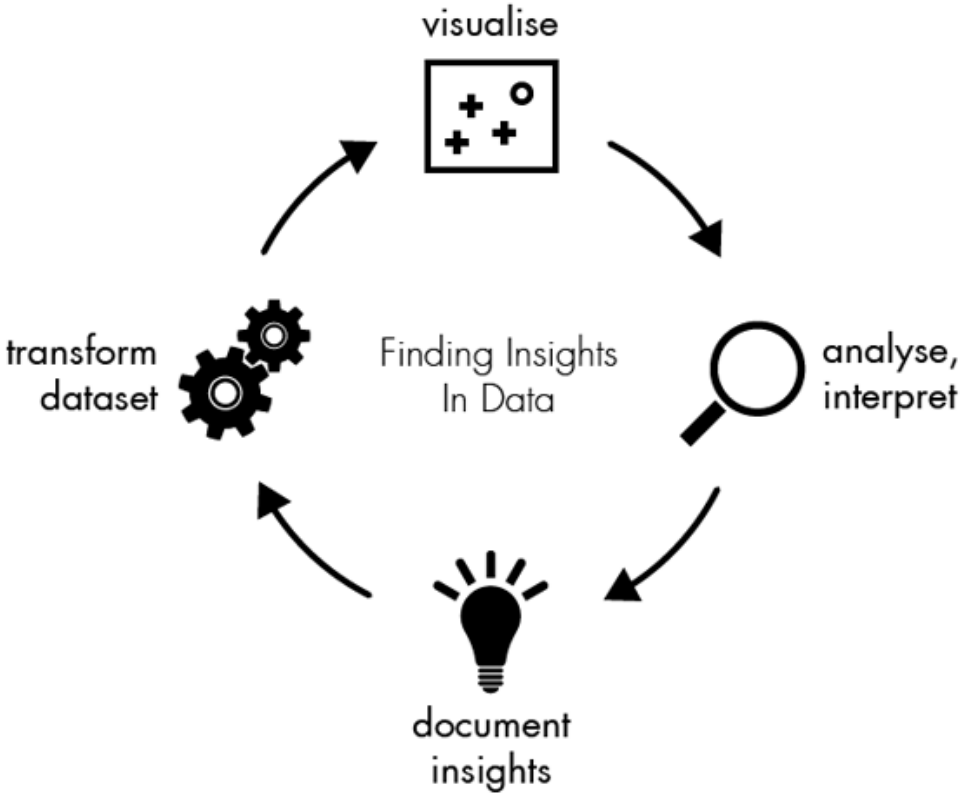
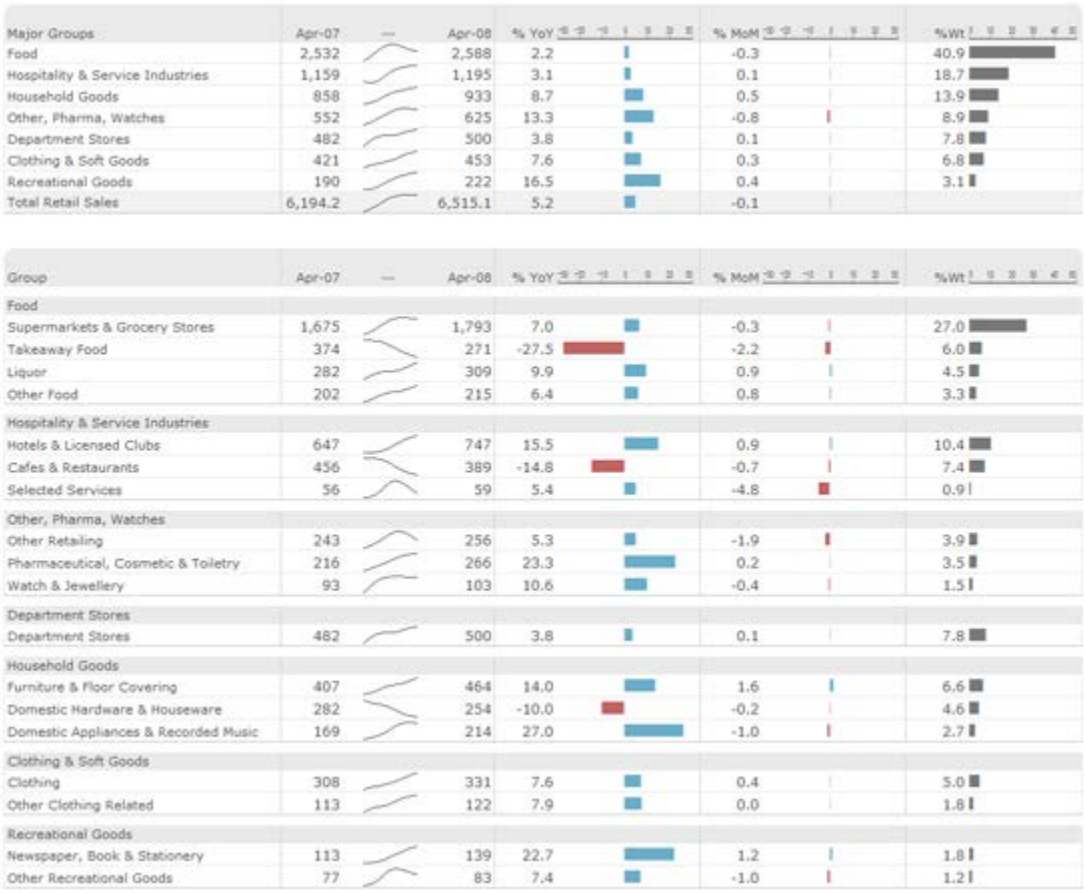


Figure 4. Información reveladora en datos; una visualización (Gregor Aisch)

Aprenda a visualizar datos

La visualización ofrece una perspectiva particular sobre el conjunto de datos. Usted puede visualizar datos de muchas maneras diferentes.

Las tablas son muy poderosas cuando se trata de un número relativamente pequeño de puntos. Muestran etiquetas y montos del modo más estructurado y organizado y revelan su potencial plenamente cuando se las combina con la capacidad de ordenar y filtrar los datos. Adicionalmente, Edward Tufte sugirió incluir pequeños gráficos dentro de columnas de tablas, por ejemplo, una barra por fila o una pequeña línea de cuadro (desde entonces conocida también como sparkline). Pero aún así, y tal como ya dijimos, las tablas claramente tienen limitaciones. Son muy buenas para mostrar cuestiones unidimensionales, como los primeros 10, pero son muy pobres cuando se trata de comparar múltiples dimensiones simultáneamente (por ejemplo, población por país a lo largo del tiempo).



Major Groups	Apr-07	Apr-08	% YoY	% MoM	%Wt
Food	2,532	2,588	2.2	-0.3	40.9
Hospitality & Service Industries	1,159	1,195	3.1	0.1	18.7
Household Goods	858	933	8.7	0.5	13.9
Other, Pharma, Watches	552	625	13.3	-0.8	8.9
Department Stores	482	500	3.8	0.1	7.8
Clothing & Soft Goods	421	453	7.6	0.3	6.8
Recreational Goods	190	222	16.5	0.4	3.1
Total Retail Sales	6,194.2	6,515.1	5.2	-0.1	

Group	Apr-07	Apr-08	% YoY	% MoM	%Wt
Food					
Supermarkets & Grocery Stores	1,675	1,793	7.0	-0.3	27.0
Takeaway Food	374	271	-27.5	-2.2	6.0
Liquor	282	309	9.9	0.9	4.5
Other Food	202	215	6.4	0.8	3.3
Hospitality & Service Industries					
Hotels & Licensed Clubs	647	747	15.5	0.9	10.4
Cafes & Restaurants	456	389	-14.8	-0.7	7.4
Selected Services	56	59	5.4	-4.8	0.9
Other, Pharma, Watches					
Other Retailing	243	256	5.3	-1.9	3.9
Pharmaceutical, Cosmetic & Toiletry	216	266	23.3	0.2	3.5
Watch & Jewellery	93	103	10.6	-0.4	1.5
Department Stores					
Department Stores	482	500	3.8	0.1	7.8
Household Goods					
Furniture & Floor Covering	407	464	14.0	1.6	6.6
Domestic Hardware & Houseware	282	254	-10.0	-0.2	4.6
Domestic Appliances & Recorded Music	169	214	27.0	-1.0	2.7
Clothing & Soft Goods					
Clothing	308	331	7.6	0.4	5.0
Other Clothing Related	113	122	7.9	0.0	1.8
Recreational Goods					
Newspaper, Book & Stationery	113	139	22.7	1.2	1.8
Other Recreational Goods	77	83	7.4	-1.0	1.2

Figure 5. Consejos de Tufte: sparklines (Gregor Aisch)

Los cuadros, en general, le permiten vincular dimensiones de sus datos con propiedades visuales de formas geométricas. Mucho se ha escrito sobre la efectividad de las propiedades visuales individuales, y la versión más breve de todo ello es: el color es difícil, la posición es todo. En un diagrama de dispersión, por ejemplo, se relaciona dos dimensiones con las

posiciones x- e y-. Incluso se puede presentar una tercera dimensión relacionada con el color o el tamaño de los símbolos presentados. Los cuadros lineales son especialmente adecuados para mostrar evoluciones temporales, mientras que los cuadros de barras son perfectos para comparar datos de categorías. Se puede apilar elementos de cuadros. Si desea comparar un pequeño número de grupos de sus datos, presentar múltiples instancias del mismo gráfico es una forma muy poderosa de hacerlo (también conocido como múltiplos pequeños). En todos los cuadros se puede usar distintos tipos de escalas para explorar aspectos diferentes de los datos (por ejemplo, lineal o escala logarítmica).

De hecho la mayor parte de los datos que manejamos están relacionados de algún modo con gente real. El poder de los mapas es que reconectan los datos con nuestro mundo físico. Imagine un conjunto de datos de incidentes criminales ubicados geográficamente. Lo crucial es ver dónde suceden los crímenes. Además los mapas pueden revelar relaciones geográficas dentro de los datos (por ejemplo, una tendencia de norte a sur, o de zonas urbanas a rurales).

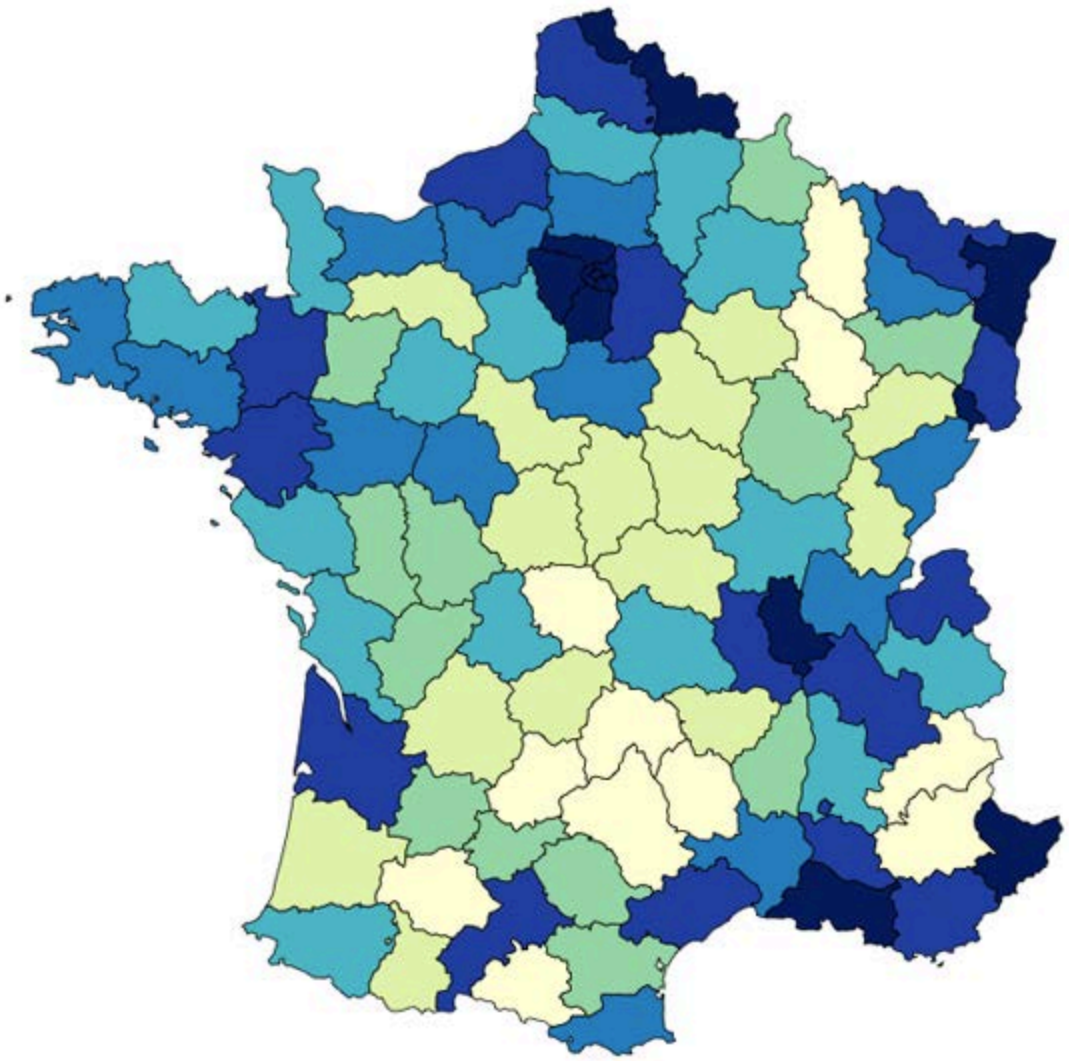


Figure 6. Mapa coroplético (Gregor Aisch)

Hablando de relaciones, el cuarto tipo más importante de visualización es el gráfico. Los gráficos sirven para mostrar las interconexiones (bordes) de sus puntos de datos (nodos). La posición de los nodos se calcula entonces por algoritmos de diagrama de gráficos más o menos complejos que nos permiten ver inmediatamente la estructura dentro de la red. El truco de la visualización por gráficos en general es encontrar el modo adecuado para modelar la red misma. No todos los conjuntos de datos incluyen ya relaciones y aunque las incluyan puede no ser el aspecto más interesante. A veces el periodista tiene que definir los bordes entre nodos. Un ejemplo perfecto de esto es el Gráfico Social del Senado de EE.UU., cuyos bordes conectan senadores que votaron lo mismo en más del 65% de los casos.

Analice e intérprete lo que ve

Una vez que haya visualizado sus datos, el siguiente paso es aprender algo del cuadro que creó. Podría preguntarse:

- ¿Qué puedo ver en esta imagen? ¿Es lo que esperaba?
- ¿Hay patrones interesantes?
- ¿Qué significa esto en el contexto de los datos?

A veces puede terminar con una visualización que, pese a su belleza, puede no decirle nada de interés de sus datos. Pero casi siempre hay algo que puede aprender de cualquier visualización, por trivial que sea.

Documente sus percepciones y sus pasos

Si piensa en este proceso como un viaje a través del conjunto de datos, la documentación es su diario de viaje. Dirá a dónde viajó, que ha visto allí y cómo tomó sus decisiones para sus siguientes pasos. Incluso puede comenzar con su documentación antes de echar su primera mirada a los datos.

En la mayoría de los casos cuando comenzamos a trabajar con un conjunto de datos que no hemos visto previamente, ya estamos llenos de expectativas y supuestos sobre los datos. Por lo general hay un motivo por el que estamos interesados en el conjunto de datos que estamos mirando. Es buena idea comenzar la documentación escribiendo estos pensamientos iniciales. Esto nos ayuda a identificar nuestros prejuicios y reduce el riesgo de malas interpretaciones de los datos encontrando simplemente lo que queríamos encontrar originalmente.

Realmente creo que la documentación es el paso más importante del proceso, y es también el que somos más proclives a dejar de lado. Como verá en el ejemplo que viene a continuación, el proceso descripto involucra mucha planificación y manejo de datos. Mirar

un conjunto de 15 cuadros que ha creado puede ser muy confuso, especialmente al transcurrir algún tiempo. De hecho esos cuadros solo son valiosos (para usted o cualquier persona a la que quiera comunicar lo que descubrió) si se los presenta en el contexto en el que fueron creados. Por tanto debe tomarse algún tiempo para hacer notas sobre cosas como:

- ¿Por qué creé este cuadro?
- ¿Qué he hecho con los datos para crearlo?
- ¿Qué me dice este cuadro?

Transforme los datos

Naturalmente con las nuevas cosas que percibió con la última visualización, puede tener una idea de lo que quiere ver a continuación. Puede haber encontrado algún patrón interesante en el conjunto de datos que ahora quiere inspeccionar con más detalle.

Las posibles transformaciones:

Acercamiento (zoom)
: Para ver cierto detalle en la visualización

Agregación
: Combinar muchos puntos de datos en un solo grupo.

Filtrado
: Eliminar (temporariamente) puntos de datos que no son de nuestro mayor interés

Eliminación de datos atípicos
: Eliminar puntos individuales que no son representativos del 99% del conjunto de datos.

Situémonos en el caso de que usted ha visualizado un gráfico y lo que surgió no fue más que un enredo de nodos conectados por cientos de bordes (un resultado muy común cuando se visualiza lo que se llama redes densamente conectadas). Un paso de transformación común sería filtrar algunos bordes. Si, por ejemplo, los bordes representan flujos de dinero de países donantes a países receptores, podríamos eliminar todos los flujos menores a cierto monto.

Qué herramientas usar

La cuestión de las herramientas no es fácil. Toda herramienta de visualización de datos disponible es buena para algo. La visualización y el manejo de los datos debe ser fácil y barato. Si cambiar los parámetros de las visualizaciones le lleva horas, no va a experimentar

demasiado. Eso no quiere decir necesariamente que no deba aprender cómo usar la herramienta. Pero una vez que aprendió, debiera ser realmente eficiente.

A menudo hay que tener mucho criterio para elegir una herramienta que cubra tanto las cuestiones del manejo de los datos como la visualización de datos. Separar las tareas en distintas herramientas significa que tiene que importar y exportar datos muy a menudo. Esta es una breve lista de algunas herramientas de visualización y manejo de datos:

- Planillas de cálculo como LibreOffice, Excel o Google Docs
- Plataformas de programación estadística como R (r-project.org) o Pandas (pandas-pydata.org)
- Sistemas de Información Geográfica (GIS) como Quantum GIS, ARcGIS, o GRASS
- Biblitoecas de Visualización como d3.js (mbostock.github.com/d3), Prefuse (prefuse.org) o Flare (flare.prefuse.org)
- Herramientas de manejo de datos como Google Refine o Datawrangler
- Software para crear visualizaciones como ManyEyes o Tableau Public (tableausoftware.com/products/public)
-

Las visualizaciones de muestra en la siguiente sección fueron creadas usando R, que es el cortaplumas suizo de la visualización de datos (científica).

Un ejemplo: encontrarle sentido a los datos sobre contribuciones electorales

Veamos la base de datos de las Finanzas de la Campaña Presidencial de Estados Unidos, que contiene alrededor de 450.000 aportes a candidatos presidenciales estadounidenses. El archivo CSV es de 60 megabytes y demasiado grande para manejar fácilmente en un programa como Excel.

En el primer paso escribiré explícitamente mis supuestos iniciales respecto del conjunto de datos sobre contribuciones para las campañas electorales:

- Obama recibe la mayor suma en contribuciones (dado que es el presidente y tiene la mayor popularidad)
- La cantidad de contribuciones aumenta al acercarse la fecha de las elecciones.
- Obama recibe más contribuciones pequeñas que los candidatos republicanos

Para responder a la primera pregunta, tenemos que transformar los datos. En vez de cada contribución individual, necesitamos sumar el total de lo aportado a cada candidato. Luego

de visualizar los resultados en una tabla ordenada, confirmamos nuestro supuesto de que Obama obtendría la mayor cantidad de dinero:

Candidato	**Monto ($)**
Obama, Barack	72.453.620,39
Romney, Mitt	50.372.334,87
Perry, Rick	18.529.490,47
Paul, Ron	11.844.361,96
Cain, Herman	7.010.445,99
Gingrich, Newt	6.311.193,03
Pawlenty, Timothy	4.202.769,03
Huntsman, Jon	2.955.726,98
Bachmann, Michelle	2.607.916,06
Santorum, Rick	1.413.552,45
Johnson, Gary Earl	413.276,89
Roemer, Charles E. *Buddy* III	291.218,80
McCotter, Thaddeus G	37.030,00

Si bien esta tabla muestra los montos mínimo y máximo y el orden, no dice demasiado acerca de los patrones subyacentes al ranking de los candidatos. Figure 7 es otra vista de los datos, un tipo de cuadro conocido como “cuadro de puntos”, en el que podemos ver todo lo que aparece en la tabla más los patrones dentro del campo. Por ejemplo, el cuadro de puntos nos permite comparar inmediatamente la distancia entre Obama y Romney y Romney y Perry, sin tener que restar valores. (Nota: este cuadro de puntos fue creado usando R. Puede encontrar vínculos con el código fuente al final de este capítulo).

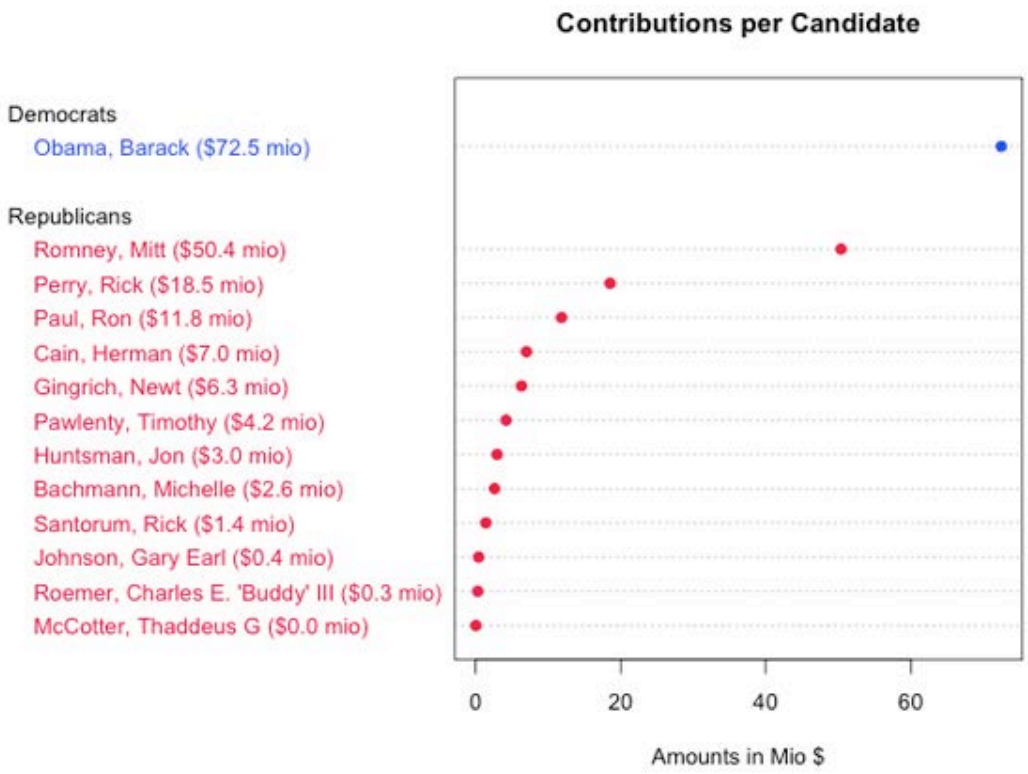


Figure 7. Visualizaciones para descubrir patrones subyacentes (Gregor Aisch)

Ahora procedamos con un cuadro más grande del conjunto de datos. Como primer paso, visualicé todos los montos aportados a lo largo del tiempo en una sola vista. Podemos ver que casi todas las contribuciones son muy, muy pequeñas comparado con 3 casos salientes. Una investigación más a fondo revela que estas contribuciones inmensas provienen del "Fondo para la Victoria de Obama 2012" (también conocido como SuperPAC) y se hicieron el 9 de junio (US$ 450.000), septiembre 29 (US$ 1.500.000) y diciembre 30 (US$ 1,900.000).

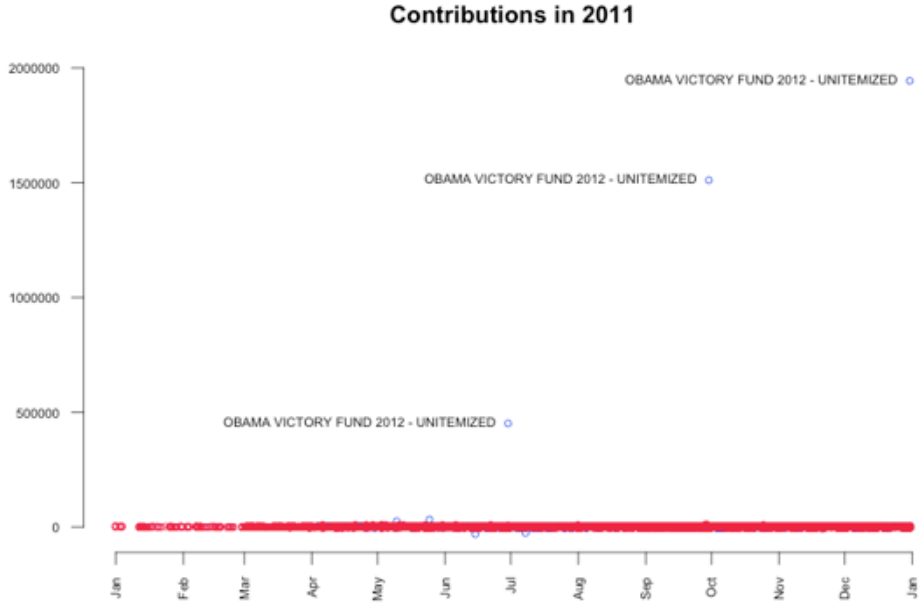


Figure 8. 3 casos salientes (Gregor Aisch)

Si bien las contribuciones de Súper PACs por si solas son sin duda la historia más importante en los datos, podría ser interesante mirar más allá. La cuestión ahora es que estas grandes contribuciones perturban nuestra visión de las contribuciones más pequeñas que provienen de individuos, por lo que vamos a quitarlas de los datos. Esta transformación se conoce comúnmente como eliminación de datos atípicos. Luego de visualizar nuevamente, podemos ver que la mayoría de las donaciones están dentro del rango de entre US$ 5.000 y US$ 10.000.

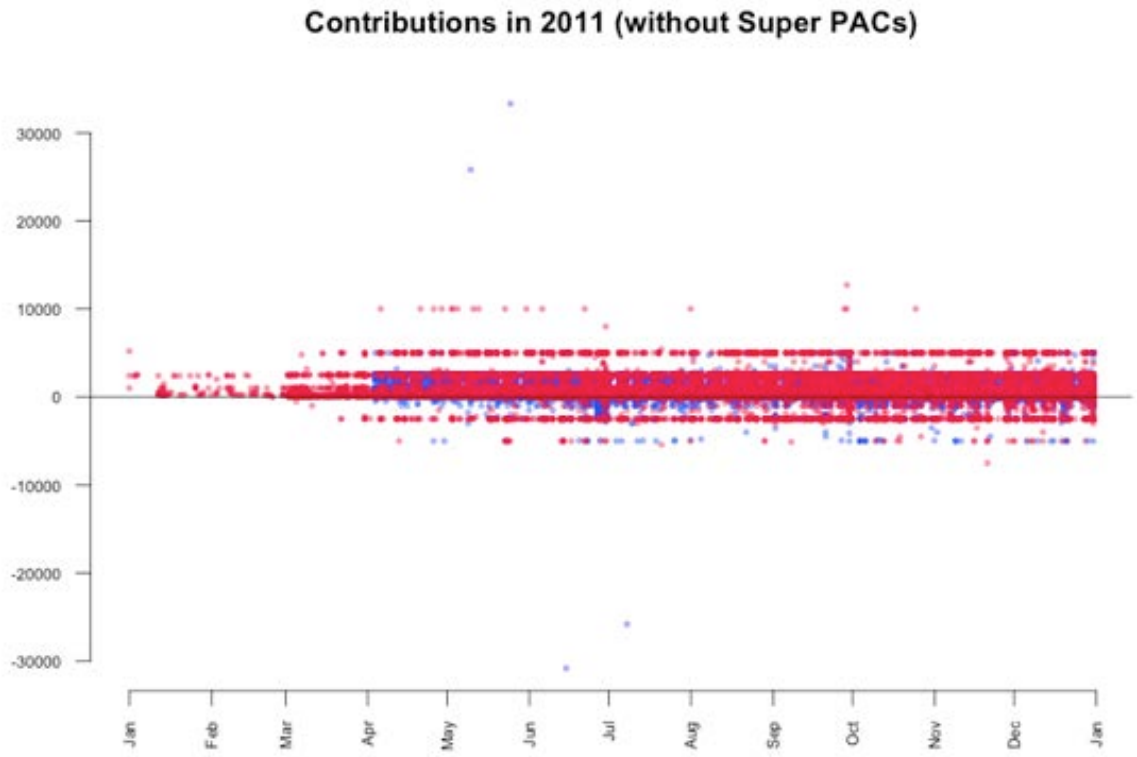


Figure 9. Eliminar datos atípicos (Gregor Aisch)

De acuerdo al límite a las contribuciones establecidos por FECA (autoridad electoral), no se permite a los individuos donar más de US$ 2500 a cada candidato. Como podemos ver en el gráfico, hay numerosas donaciones por encima de ese límite. En particular, nos llaman la atención dos grandes contribuciones en mayo. Parece que son compensadas por montos negativos (reembolsos) en junio y julio. Una investigación más a fondo de los datos revela las siguientes transacciones:

- El 10 de mayo, *Stephen James Davis*, de San Francisco, empleado en Banneker Partners (abogados), ha donado **US$ 25.800** a Obama.
- El 25 de mayo, *Cynthia Murphy*, de Little Rock, empleada en el Murphy Group (relaciones públicas), ha donado **US$ 33.300** a Obama
- El 15 de junio el monto de **US$ 30.800** fue devuelto a *Cynthia Murphy*, lo que redujo el monto donado a US$ 2500.
- El 8 de julio, se devolvió el monto de **US$ 25.800** a *Stephen James Davis*, lo que redujo el monto donado a US$ 0.

¿Qué tienen de interesantes estas cifras? Los US$ 30.800 devueltos a Cynthia Murphy equivalen al monto máximo que pueden dar individuos a comités nacionales de partidos al año. Quizás quería combinar ambas donaciones en una transacción, que fue rechazada. Los US$ 25.800 devueltos a Stephen James Davis posiblemente equivalen a los US$ 30.800 menos US$ 5000 (el límite de aportes a cualquier otro comité político).

Otra cosa interesante descubierta en el último gráfico es un patrón lineal horizontal de contribuciones para candidatos republicanos por US$ 5000 y -US$ 2500. Para verlos con más detalle, visualicé solo las donaciones a republicanos. El gráfico resultante es un gran ejemplo de patrones en datos que serían invisibles sin visualización de datos.

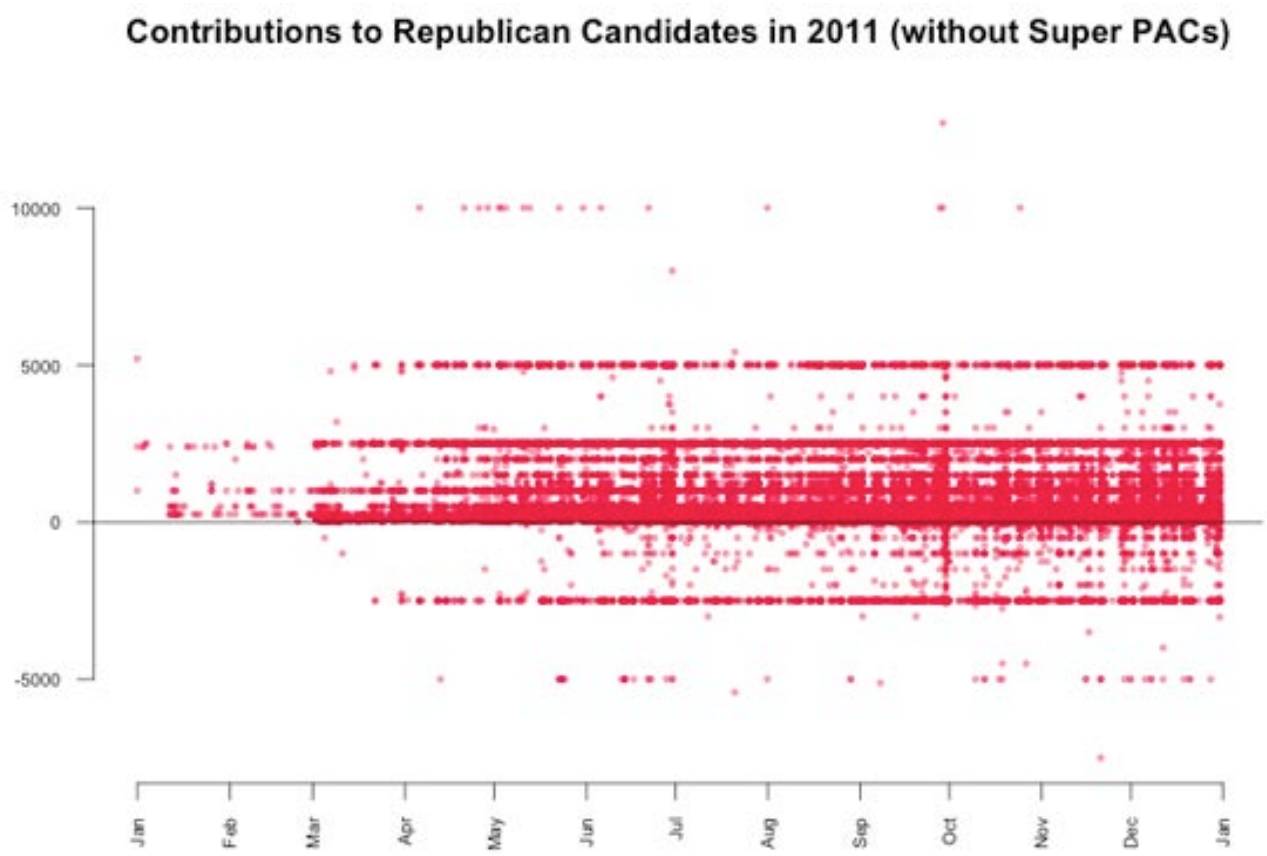


Figure 10. Eliminación de datos atípicos 2 (Gregor Aisch)

Lo que podemos ver es que hay muchas donaciones de US$ 5000 a candidatos republicanos. De hecho, un análisis de los datos da que hay 1243 de estas donaciones, que es solo el 0,3% del número total de donaciones, pero debido a que esas donaciones se reparten de modo parejo en el tiempo, la línea aparece. Lo interesante de la línea es que las donaciones de individuos estaban limitadas a US$ 2500. En consecuencia cada dólar que superó ese límite fue devuelto a los donantes, lo que resulta en la segunda línea de –US$ 2500. En contraste, las contribuciones a Barack Obama no muestran un patrón similar.

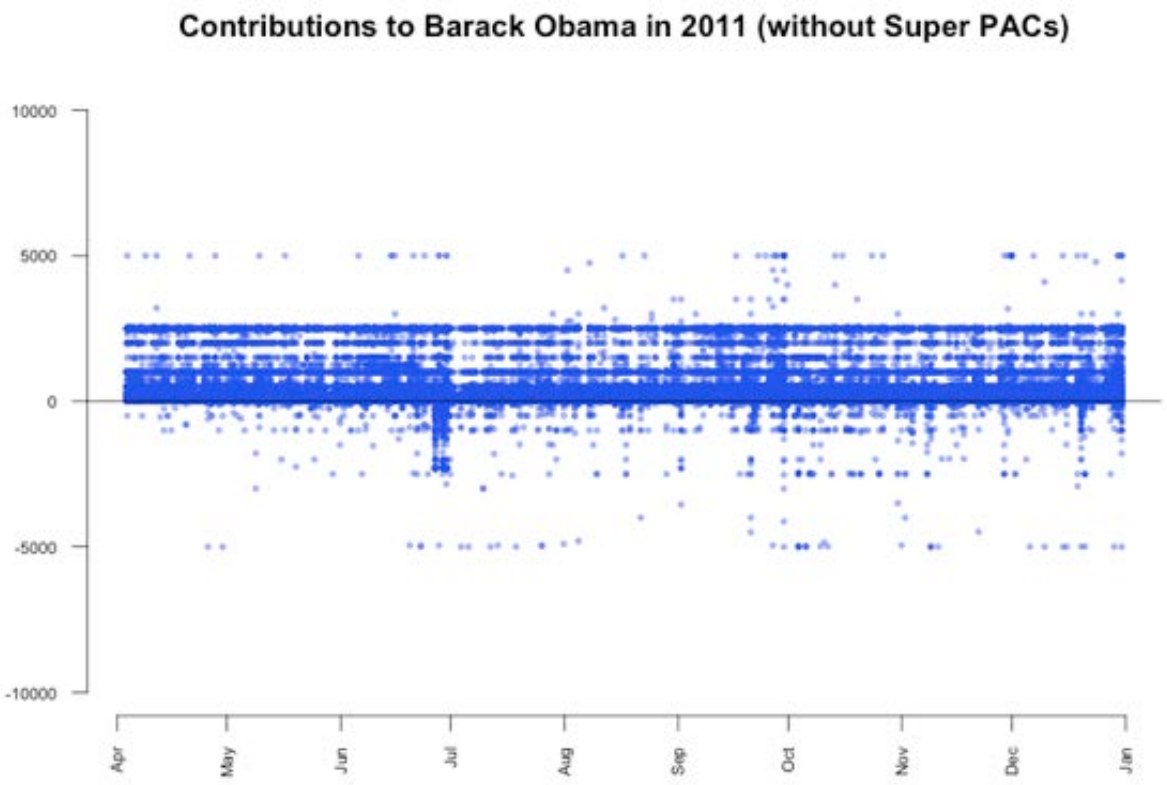


Figure 11. Eliminación de datos atípicos 3 (Gregor Aisch)

Por lo que podría ser interesante averiguar por qué miles de donantes republicanos no advirtieron los límites para donaciones de individuos. Para analizar más en profundidad el tema, podemos ver el número total de donaciones de US$ 5000 por candidato.

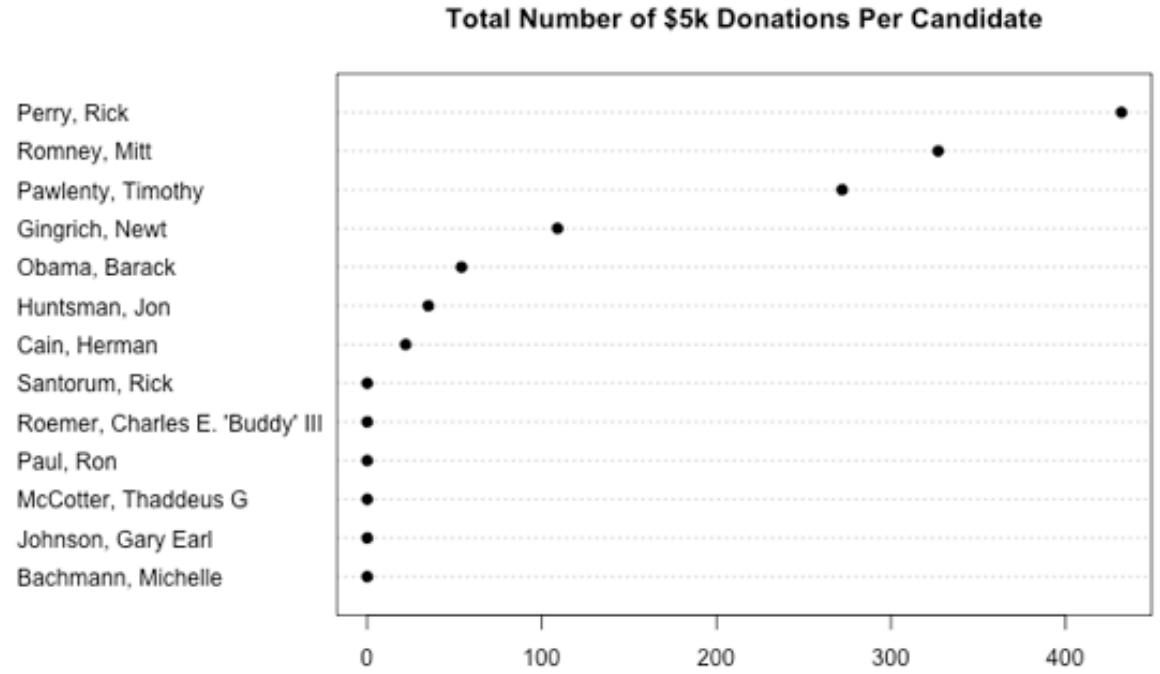


Figure 12. Donaciones por candidato (Gregor Aisch)

Por supuesto que esta es una visión distorsionada dado que no considera los montos totales de donaciones recibidas por cada candidato. El siguiente gráfico muestra el porcentaje de donaciones de US$ 5000 por candidato.

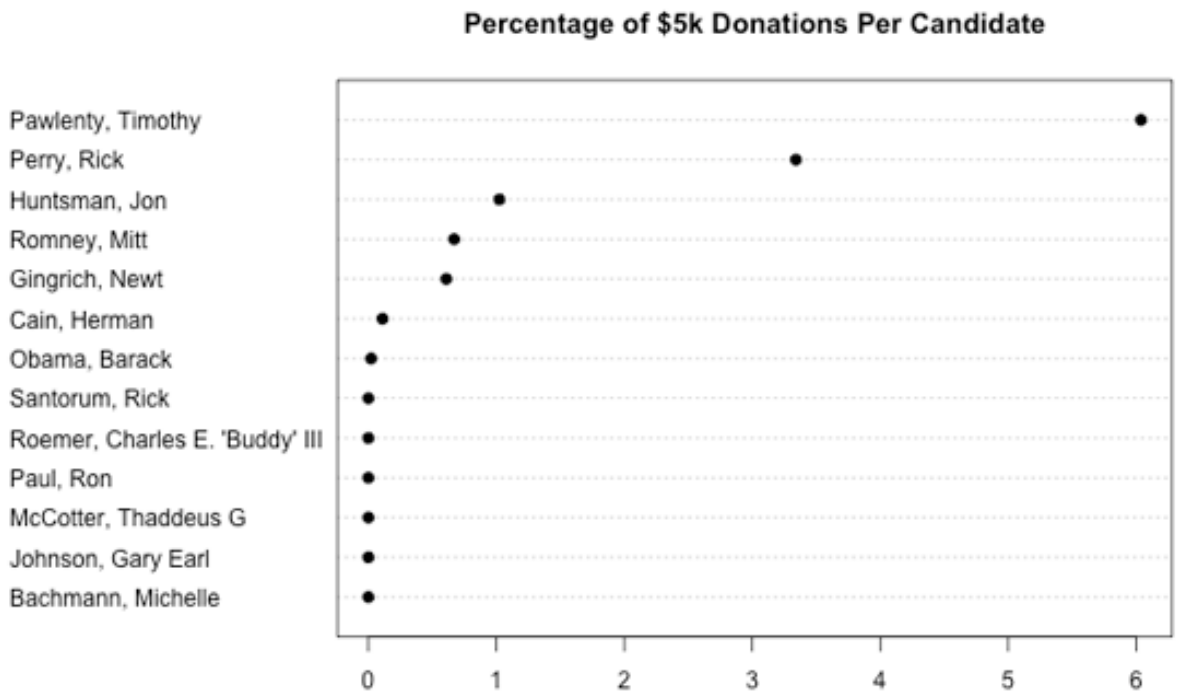


Figure 13. ¿De dónde viene la plata del senador?: donaciones por candidato (Gregor Aisch)

Qué aprender de esto

A menudo tal análisis visual de un nuevo conjunto de datos se vive como un viaje excitante a un país desconocido. Uno comienza como un extranjero contando solo con los datos y sus supuestos, pero con cada paso que da, con cada cuadro que produce, percibe cosas nuevas sobre el tópico. Basado en esas percepciones, toma decisiones respecto de sus siguientes pasos y que cuestiones ameritan una mayor investigación. Como habrá visto en este capítulo, este proceso de visualizar, analizar y transformar datos podría repetirse casi al infinito.

Consiga el código fuente

Todos los cuadros que se muestran en este capítulo fueron creados usando el maravilloso y poderoso software R. Creado principalmente como herramienta de visualización científica, es difícil encontrar alguna técnica de visualización o manejo de datos que no esté incorporada a R. Para aquellos interesados en saber cómo visualizar y manejar datos

usando R, a continuación aparecen los códigos fuente para los cuadros generados en este capítulo:

- Cuadro de puntos: contribuciones por candidato
- Gráfico: todas las contribuciones a lo largo del tiempo
- Gráfico: contribuciones por comités autorizados

Hay también una gran variedad de libros y tutoriales disponibles.

— *Gregor Aisch, Open Knowledge Foundation*

Difundir datos

Una vez que analizó bien sus datos y decidió que hay algo interesante sobre lo cual escribir, ¿cómo puede difundirlos al público? Esta sección se inicia con breves anécdotas acerca de cómo conocidos periodistas presentaron los datos a sus lectores: desde infografías, pasando por plataformas de datos, hasta *links* para descarga. Luego analizamos con más detenimiento cómo crear nuevas aplicaciones y los detalles de la visualización de datos. Finalmente analizamos lo que puede hacer para que su proyecto le resulte atractivo al público.

Qué contiene este capítulo?

- Presentar datos al público
- Cómo crear una aplicación de noticias
- Aplicaciones de noticias en ProPublica
- La visualización como el caballo de tiro del periodismo de datos
- El uso de visualizaciones para narrar historias
- Cuadros diferentes dicen cosas diferentes
- Selección de herramientas "Hágalo Ud. mismo" para hacer sus propias visualizaciones de datos.
- Cómo presentamos los datos en el Verdens Gang
- Los datos públicos se vuelven sociales
- Interactuar con la audiencia en torno a sus datos

Presentar datos al público

Hay muchas maneras diferentes de presentar los datos al público, desde publicar conjuntos de datos en crudo con historias, hasta crear hermosas visualizaciones y aplicaciones interactivas en la web. Pedimos consejos a periodistas de datos con experiencia sobre cómo presentar datos al público.

Visualizar o no visualizar

Hay momentos en que los datos pueden contar la historia mejor que palabras o fotos y es por eso que términos como “aplicación de noticias” y “visualización de datos” han adquirido el status de palabras clave en tantas redacciones en los últimos tiempos. También promueve el interés la gran cosecha de nuevas herramientas y tecnologías (a menudo gratuitas) destinadas a ayudar incluso al periodista menos dotado técnicamente a convertir datos en una presentación visual de una historia.

Herramientas como Google Fusion Tables, Many Eyes, Tableau, Dipity, y otras hacen más fácil que nunca crear mapas, cuadros, gráficos o incluso aplicaciones con datos que hasta aquí eran el dominio de especialistas. Siendo las barreras al ingreso ahora apenas un lomo de burro, la cuestión para los periodistas ahora no es tanto si pueden convertir sus conjuntos de datos en una visualización sino si les conviene hacerlo. Una mala visualización de datos es peor en muchos sentidos que ninguna visualización.

— *Aron Pilhofer, New York Times*

El uso de gráficos con movimiento

Con un guión ajustado, animaciones bien cronometradas y explicaciones claras, los gráficos con movimiento pueden dar vida a números o ideas complejas, orientando a su público. Las videoconferencias de Hans Rosling son un buen ejemplo de cómo los datos pueden narrar una historia en la pantalla. Concuerde o no usted con su metodología, yo también creo que el índice Shoe-throwers de The Economist es un buen ejemplo del uso del video para contar una historia basada en números. Usted no presentaría este gráfico como una imagen estática o al menos no debería hacerlo. Suceden demasiadas cosas en la presentación. Pero habiendo llegado paso a paso, a uno le queda la comprensión de cómo y por qué llegaron a este índice. Con gráficos con movimiento y cortos animados, puede reforzar lo que el público está escuchando. Los recursos visuales explicativos con la voz *en off* ofrecen una manera poderosa y memorable de contar una historia.

— *Lulu Pinney, freelance infographic designer*

Contarle al mundo

Nuestro flujo de trabajo por lo general comienza con Excel. Es una manera fácil de descubrir si hay algo interesante en los datos. Si tenemos la sensación de que hay algo, entonces vamos a la mesa de noticias. Tenemos suerte de estar ubicados junto a la principal mesa de noticias de The Guardian. Entonces analizamos cómo visualizarlo o mostrarlo en la página. Luego escribimos el texto que lo acompaña. Cuando escribo por lo general tengo una versión reducida de la planilla de cálculo junto al editor de texto. A menudo hago análisis parciales mientras escribo, para encontrar cosas interesantes. Publico en el blog y me dedico un rato a hacer tweets al respecto, escribiendo a distintas personas y asegurándome de que tiene links a todos los lugares indicados.

La mitad del tráfico de algunas de las cosas que subimos al blog vienen de Twitter y Facebook. Estamos bastante orgullosos de que el tiempo promedio dedicado a un artículo en nuestro Datablog es de 6 minutos, comparado con un promedio de 1 minuto para el resto del sitio de The Guardian. Seis minutos es bastante bueno y el tiempo en la página es una de las métricas claves al analizar nuestro tráfico.

Esto también ayuda a convencer a nuestros colegas acerca del valor de lo que estamos haciendo. Eso y las grandes historias basadas en datos en las que hemos trabajado que todos los demás en la redacción conocen: COINS, WikiLeaks y los disturbios en el Reino Unido. Para los datos de gasto de COINS, tuvimos 5-6 periodistas especializados trabajando en The Guardian para dar sus puntos de vista sobre los datos cuando fueron difundidos por el gobierno del Reino Unido. También tuvimos otro equipo de 5-6 periodistas cuando el gobierno difundió los datos de gastos por encima de las £25000 libras, incluyendo periodistas muy conocidas como Polly Curtis. WikiLeaks también obviamente fue muy importante, con muchas historias sobre Irak y Afganistán. Los disturbios también fueron bastante importantes, con más de 550.000 vistas en 2 días.

Pero no se trata solo de las visitas de corto plazo: también tiene que ver con ser una fuente confiable de información útil. Tratamos de ser el lugar donde usted puede obtener información buena y significativa sobre los temas que cubrimos.

— Simon Rogers, the Guardian

Publicar los datos

A menudo publicamos los datos en nuestro sitio en una visualización y de una forma que permite la fácil descarga del conjunto de datos. Nuestros lectores pueden explorar los datos detrás de las historias interactuando en las visualizaciones o usando los datos mismos de otros modos. ¿Por qué es importante esto? Aumenta la transparencia de The Seattle Times. Mostramos a los lectores los mismos datos que usamos para sacar poderosas conclusiones.

¿Y quién las usa? Nuestros críticos sin duda, así como todos los que simplemente están interesados en la historia y todas sus ramificaciones. Al dar acceso a los datos también podemos recibir comentarios de estos mismos críticos y los lectores en general respecto de lo que no vimos y qué más podemos explorar, todas cosas valiosas para un periodismo que importa.

— *Cheryl Phillips, The Seattle Times*

Dar acceso a sus datos

Dar a los consumidores de noticias acceso fácil a los datos que usamos para nuestro trabajo es lo correcto por varios motivos. Los lectores pueden asegurarse de que no estamos torturando a los datos para llegar a conclusiones forzadas. Dar acceso a nuestros datos continúa la tradición de las ciencias sociales de permitir que investigadores reproduzcan nuestro trabajo. Alentar a los lectores a estudiar los datos puede generar ideas que lleven a la continuación de las historias. Finalmente, los lectores interesados en sus datos van a ser proclives a volver una y otra vez.

— *Steve Doig, Walter Cronkite School of Journalism, Arizona State University*

Crear una plataforma de datos abierta

En La Nación publicar datos con libre acceso es una parte integral de nuestras actividades periodísticas de datos. En la argentina no existe una Ley de Acceso a la Información ni un portal nacional de datos, por lo que nos parece importante proveer a nuestros lectores el acceso a los datos que usamos en nuestras historias.

Por tanto publicamos datos estructurados en crudo a través de nuestra plataforma integrada Junar así como en Google Spreadsheets. Explícitamente autorizamos y alentamos a terceros a reutilizar nuestros datos y explicamos un poco acerca de cómo hacer esto con documentación y tutoriales en video.

Lo que es más, presentamos algunos de estos conjuntos de datos y visualizaciones en nuestro blog NACION DATA. Hacemos esto para evangelizar sobre nuestros datos y herramientas de edición de datos en la Argentina y mostrar a otros cómo reunimos nuestros datos, como los usamos y cómo pueden reutilizarlos.

Desde que lanzamos la plataforma en febrero de 2012, hemos recibido sugerencias e ideas para conjuntos de datos, principalmente de académicos e investigadores, así como estudiantes de universidades que se muestran muy agradecidos cada vez que contestamos con una solución o un conjunto de datos específico. Hay gente que conoce y comenta nuestros datos en Tableau y varias veces hemos sido el ítem más comentado y visto en el

sitio. En 2011 tuvimos 7 de las 100 visualizaciones más vistas.

— Angélica Peralta Ramos, La Nación (Argentina)

Humanizar los datos

Al ampliarse el conocimiento del debate en torno a los grandes conjuntos de datos, una parte importante ha estado notoriamente ausente: el elemento humano. Mientras muchos pensamos en los datos como números disociados, flotando en el vacío, en realidad son mediciones de cosas tangibles (y a menudo humanas). Los datos están unidos a la vida de gente real y cuando abordamos los números, debemos considerar los sistemas del mundo real de los que provienen.

Tomemos, por ejemplo, los datos de ubicación que se están recogiendo en este momento de cientos de millones de celulares y dispositivos móviles. Es fácil pensar en estos datos (cifras que representan latitud, longitud y tiempo) como "descarga digital", pero en realidad son datos destilados de momentos de nuestras narrativas personales. Si bien pueden ser secos y clínicos cuando se leen en una planilla de cálculo, cuando permitimos a la gente incorporar sus propios datos a un mapa y reproducirlos, experimentan una especie de replay de la memoria que es poderoso y humano.

En este momento, los datos de localización son utilizados por muchos diseñadores de aplicaciones, grandes marcas y anunciantes. Mientras las segundas (empresas de telecomunicaciones y administradores de dispositivos) son dueños y almacenan los datos, el primero en esta ecuación –usted- no tiene acceso ni control de esta información. En el grupo de Investigación y Desarrollo del NYTimes, hemos lanzado un proyecto prototipo llamado OpenPaths para permitir al público explorar sus propios datos de locación y experimentar el concepto de propiedad de los datos. Al fin de cuentas, la gente debiera tener control de estas cifras tan estrechamente ligadas a sus propias vidas y experiencias.

Los periodistas tienen un rol muy importante en sacar a luz esta humanidad inherente a los datos. Al hacerlo, tienen el poder de cambiar la comprensión del público tanto de los datos como de los sistemas de los que emergieron los números.

— Jer Thorp, Data Artist in Residence: New York Times R&D Group

Datos abiertos, fuentes abiertas, noticias abiertas

El 2012 bien pudo ser el año de las noticias abiertas. Está en el centro de nuestra ideología editorial y es un mensaje clave de nuestra marca actualmente. En medio de todo esto, está claro de que necesitamos un proceso abierto para el periodismo basado en datos. Este proceso no solo debe ser alimentado de datos abiertos, sino también facilitado por

herramientas abiertas. Para fin de año esperamos poder acompañar cada visualización que publicamos con acceso tanto a los datos como al código con el que se construyó.

Muchas de las herramientas usadas en la visualización hoy son de fuente cerrada. Otras vienen con licencias restrictivas que prohíben el uso de datos derivados. Las bibliotecas de fuente abierta existentes a menudo resuelven un problema bien pero no ofrecen una metodología más amplia. De conjunto esto dificulta a la gente apoyarse en el trabajo de los demás. Esto cierra conversaciones en vez de abrirlas. Con este fin, estamos desarrollando una cantidad de herramientas abiertas para narrar historias interactivas. El Miso Project (@themisoproject) es un ejemplo.

Estamos analizando este trabajo con una cantidad de organizaciones de medios. Se requiere de la participación de la comunidad para realizar plenamente el potencial del software de código abierto. Si tenemos éxito introducirá una dinámica fundamentalmente diferente con nuestros lectores. Las contribuciones pueden ir más allá del comentario a bifurcar nuestro trabajo, solucionar problemas o re-utilizar datos de maneras inesperadas.

— *Alastair Dant, the Guardian*

Agregue un link de descarga

En los últimos años trabajé con unos cuantos gigabytes de datos para proyectos o artículos, desde el escaneado de tablas escritas a máquina de la década del ‘60 hasta los 1,5 gigabytes de cables publicados por WikiLeaks. Siempre ha sido difícil convencer a los editores de publicar sistemáticamente los datos en formato abierto y accesible. Para superar el problema, agregué links para “Descargar los datos” dentro de los artículos, apuntando a los archivos que los contenían o los Google Docs relevantes. El interés de potenciales reutilizadores coincidía con lo que vemos en los programas promovidos por el Estado (es decir, muy, pero muy escaso). Sin embargo, las pocas instancias de reutilización aportaron nuevas visiones o promovieron conversaciones que bien valen los pocos minutos extra por proyecto.

— *Nicolas Kayser-Bril, Journalism++*

Conozca su alcance

Hay una gran diferencia entre hackear por diversión y hacer ingeniería de sistemas buscando escala y buen desempeño. Asegúrese de asociarse con gente que tenga las capacidades apropiadas para su proyecto. No olvide el diseño. La facilidad de uso, la experiencia del usuario y el diseño de la presentación pueden afectar mucho el éxito de su proyecto.

— *Chrys Wu, Hacks/Hackers*

Cómo crear una aplicación de noticias

Son ventanas que muestran los datos en los que se apoya la historia. Pueden ser bases de datos abiertas a búsquedas, visualizaciones elegantes, o algo totalmente distinto. Pero no importa la forma que asuman, las aplicaciones alientan a los lectores a interactuar con los datos en un contexto que es significativo para ellos: investigar tendencias criminalísticas en su zona, verificar los antecedentes de su médico local o analizar las contribuciones políticas de su candidato.

Más que infografías de alta tecnología, las mejores aplicaciones de noticias son productos durables. Tienen vida por fuera del ciclo de las noticias, ayudando a menudo a los lectores a resolver problemas del mundo real, o respondiendo preguntas de un modo tan útil como novedoso que se convierten en recursos perdurables. Cuando periodistas de ProPublica quisieron explorar en qué medida eran seguras las clínicas de diálisis de riñón estadounidenses, crearon una aplicación que ayudaba a los usuarios a verificar si las instalaciones en su ciudad eran seguras. Proveer un servicio tan importante y relevante crea una relación con los usuarios que va mucho más allá de lo que una historia narrativa puede hacer por sí sola.

Allí está el desafío y la promesa de crear aplicaciones de noticias que son lo último en materia tecnológica: crear algo de valor duradero. Sea usted un diseñador o un gerente, cualquier discusión acerca de crear una gran aplicación debe comenzar con una mentalidad de desarrollo de un producto: mantenerse enfocado en el usuario y trabajar para lograr el mayor impacto con su inversión. Por lo que, antes de comenzar a crear una aplicación, es bueno hacerse tres preguntas, que se abordan en las siguientes secciones.

Figure 1. Monitor de instalaciones para diálisis (ProPublica)

¿Cuál es mi público y cuáles son sus necesidades?

Las aplicaciones de noticias no sirven a la historia por la historia misma, sirven al usuario. Según el proyecto, el usuario puede ser un paciente de diálisis que quiere conocer los antecedentes de su clínica o incluso una dueña de casa que no conoce el riesgo de terremoto cerca de su hogar. No importa quién sea, toda discusión sobre la creación de una aplicación de noticias, como cualquier buen producto, debe empezar por la gente que la va a usar.

Una sola aplicación puede servir a muchos usuarios. Por ejemplo, un proyecto llamado Curbwise, creado por el Omaha (Nebraska) World-Herald le sirve a propietarios de casas que creen que les están cobrando impuestos excesivos, a residentes curiosos interesados en los valores de propiedades cercanas y trabajadores inmobiliarios que buscan seguir las tendencias de las ventas recientes. En cada uno de esos casos, la aplicación responde a necesidades específicas que hacen que los usuarios vuelvan.

Los propietarios de casas, por ejemplo, podrían necesitar ayuda para reunir información sobre propiedades próximas de modo de poder demostrar que sus impuestos son injustamente elevados. Reunir esa información exige tiempo y es complicado, un problema que Curbwise resuelve para sus usuarios compilando un informe fácil de usar de toda la información que necesitan para cuestionar los impuestos a sus propiedades ante las autoridades municipales. Curbwise vende ese informe por US$ 20 y la gente lo paga porque le resuelve un problema real de sus vidas.

Sea que su aplicación resuelva un problema del mundo real como Curbwise o acompañe la narrativa de una historia con visualizaciones interesantes, siempre sea consciente de la gente que la usará. Concéntrese en diseñar y crear los componentes basados en sus necesidades.

¿Cuánto tiempo debo dedicar a esto?

Los programadores en la redacción son como agua en el desierto: muy buscados y escasos. Crear aplicaciones de noticias significa equilibrar las necesidades diarias de una redacción con los compromisos de largo plazo que se necesita para crear productos realmente buenos.

Digamos que su editor le viene con una idea: el Consejo Municipal va a votar la semana entrante si demoler o no varias propiedades históricas en su ciudad. Sugiere crear una aplicación simple que le permita a los usuarios ver los edificios en un mapa.

Como programador, usted tiene unas pocas opciones. Puede flexionar su músculo de ingeniero de sistemas creando un mapa fabuloso usando software especialmente desarrollado para el caso. O puede usar herramientas existentes como las Google Fusion Tables o bibliotecas de mapeado de código abierto y terminar el trabajo en un par de horas.

La primera opción le dará una mejor aplicación; pero la segunda puede darle más tiempo para crear otra cosa con mayores probabilidades de tener un impacto duradero.

El hecho de que una historia sea apta para crear una aplicación compleja y hermosa no significa que tenga que crearla. Es crítico saber medir las prioridades. La cuestión es recordar que toda aplicación que usted cree tiene un costo: a saber, otra aplicación potencialmente más impactante en la que pudo haber estado trabajando.

¿Cómo puedo llevar la cosa al siguiente nivel?

Crear aplicaciones de noticias sofisticadas puede exigir mucho tiempo y ser costoso. Por eso siempre se justifica preguntar cuál será el rédito. ¿Cómo se convierte una aplicación maravillosa pero que produce solo un impacto momentáneo en algo especial y duradero?

Crear un proyecto duradero que trascienda el ciclo de las noticias es una manera de hacerlo. Otra manera es crear una herramienta que le ahorre tiempo en el futuro (y haciéndolo con código abierto) o aplicar un sistema de medición avanzada a su aplicación para saber más de su público.

Muchas organizaciones crean mapas en base al censo para mostrar los cambios demográficos en sus ciudades. Pero cuando el equipo de aplicaciones interactivas del Chicago Tribune hizo el suyo, llevó las cosas al siguiente nivel desarrollando herramientas y técnicas para crear esos mapas rápidamente, y que luego pusieron a disposición de otras organizaciones.

En mi lugar de empleo, el Center for Investigative Reporting, unimos una base de datos simple en la que se podía hacer búsquedas, con una plataforma de búsqueda fina que nos permitió saber, entre otras cosas, cuántos usuarios valoran los hallazgos fortuitos y la exploración en nuestras aplicaciones.

A riesgo de parecer que lo único que le importa es la plata, siempre piense en términos de ganancias sobre la inversión. Resuelva un problema genérico; cree una nueva manera de atraer a los usuarios; ofrezca partes de su trabajo con código abierto; use sistemas de medición para saber más acerca de sus usuarios; o incluso descubra cómo puede generar ingresos con partes de su aplicación, como lo hace Curbwise.

En síntesis

La creación de aplicaciones de noticias ha recorrido un largo camino en muy poco tiempo. Las aplicaciones 1.0 eran muy parecidas a infografías 2.0, visualizaciones de datos interactivas, mezcladas con bases de datos en las que se podía hacer búsquedas, dirigidas primordialmente a sostener la narrativa de la historia. Ahora muchas de esas aplicaciones

pueden ser diseñadas por periodistas incluso cuando están apurados por plazos de entrega usando herramientas de código abierto, lo que deja a los programadores libres para pensar en cosas más importantes.

Las aplicaciones 2.0, que es hacia donde se dirige el sector, tienen que ver con combinar la narración y los puntos fuertes del periodismo como servicio público con el desarrollo de productos y los conocimientos tecnológicos. El resultado, sin duda, será una explosión de innovación en torno a maneras de hacer que los datos sean relevantes, interesantes y especialmente útiles para nuestro público y, al mismo tiempo, esperemos que ayude al periodismo a hacer esto mismo.

— Chase Davis, Center for Investigative Reporting

Aplicaciones de noticias en ProPublica

Una aplicación es una gran base de datos interactiva que narra una historia noticiosa. Piense en ella como lo haría con cualquier otra pieza de periodismo. Simplemente usa software en vez de palabras e imágenes.

Al mostrar a cada lector datos que son específicos a él, una aplicación puede ayudar a cada lector a comprender una historia de un modo que sea personalmente significativo. Puede ayudar a un lector a comprender su relación personal con un fenómeno nacional amplio y ayudarlo a relacionar lo que sabe con lo que no sabe y por tanto alentar una comprensión profunda de conceptos abstractos.

Tendemos a crear aplicaciones de noticias cuando tenemos un conjunto de datos (o creemos que podemos adquirir un conjunto de datos) que sea de alcance nacional y a la vez lo suficientemente granular como para exponer detalles significativos.

Una aplicación debiera narrar una historia, y al igual que cualquier buena historia, necesita un titular, una firma, un encabezado y una síntesis que presente el contenido. Algunos de estos conceptos pueden ser difíciles de distinguir en una pieza de software interactivo, pero están allí si uno lo estudia atentamente.

Además, una aplicación debiera ser generadora de más historias y más informes. Las mejores aplicaciones de ProPublica han sido usadas como base para historias locales.

Por ejemplo, tomemos el caso de nuestra aplicación Dollars for Docs. Rastreaba pagos de compañías farmacéuticas por millones de dólares a médicos para que hicieran consultoría, dieran conferencias y otras cosas por el estilo. La aplicación que creamos permite a los lectores hacer una búsqueda sobre su propio médico y ver los pagos que recibió. Periodistas de otras organizaciones también usaron los datos. Más de 125 organizaciones de noticias

locales, incluyendo el Boston Globe, Chicago Tribune y St. Louis Post-Dispatch hicieron investigaciones sobre médicos locales basados en datos de Dollars for Docs.

Unas cuantas de estas historias locales fueron resultado de asociaciones formales, pero la mayoría se hicieron de modo independiente, en algunos casos no tuvimos demasiado conocimiento –si es que supimos algo - de que se estaba trabajando en la historia hasta que apareció. Como organización pequeña pero de alcance nacional, este tipo de repercusión es crucial para nosotros. No podemos tener conocimiento de lo que sucede en 125 ciudades, pero si nuestros datos pueden ayudar a periodistas que tienen conocimiento local a narrar historias con impacto, estamos cumpliendo nuestra misión.

Una de mis aplicaciones favoritas es Mapping L.A. de Los Ángeles Times, que comenzó como un mapa de varios barrios de esa ciudad con datos del público y que hasta su aparición no tenían límites aceptados por todos. Luego del primer proyecto con aportes del público (crowdsourcing) el Times pudo usar los barrios como un gran dispositivo de base para hacer informes de datos: cosas como la tasa de criminalidad por barrio, calidad de las escuelas por barrio, etc., que antes no podía hacer. De modo que Mapping L.A. no solo es a la vez genérico y específico, es generador de proyectos y cuenta las historias de la propia gente.

Los recursos necesarios para crear una aplicación son muy variados. The New York Times tiene docenas de personas trabajando en aplicaciones y gráficos interactivos. Pero Talking Points Memo hizo un seguidor de encuestas políticas de última generación con 2 empleados, ninguno de los cuales tenía título en ciencias de la computación.

Al igual que la mayoría de los programadores que trabajan en redacciones, seguimos una metodología Agile modificada para crear nuestras aplicaciones. Iteramos rápidamente y mostramos borradores a la otra gente de la redacción con la que trabajamos. Es de la mayor importancia el hecho de que trabajamos en estrecho contacto con periodistas y leemos sus borradores, incluso los muy iniciales. Trabajamos mucho más como periodistas que como programadores tradicionales. Además de escribir código, llamamos a las fuentes, reunimos información y acumulamos experiencia. Sería difícil hacer una buena aplicación usando material que no entendemos.

¿Por qué debieran interesarse las redacciones en producir aplicaciones basadas en datos? Tres razones: es excelente periodismo, es inmensamente popular –los contenidos más populares de ProPublica son aplicaciones de noticias- y si no lo hacemos, otro lo hará. Piense en todas las exclusivas que nos perderíamos. Lo que es más importante, las redacciones debieran saber que pueden hacerlo también. Es más fácil de lo que parece.

— Scott Klein, ProPublica

La visualización como el caballo de tiro del periodismo de datos

Antes de lanzarse a tratar de armar cuadros o mapas con sus datos, tómese un minuto para pensar acerca de los muchos roles que los elementos gráficos estáticos e interactivos tienen en su trabajo periodístico.

En la fase de buscar la información, las visualizaciones pueden:

- Ayudarlo a identificar temas y cuestiones para el resto de su tarea.
- Identificar cosas fuera de lugar: buenas historias o quizás errores en sus datos.
- Ayudarlo a encontrar ejemplos típicos.
- Mostrar baches en sus informes.

Las visualizaciones también tienen múltiples roles en la edición. Pueden:

- Ilustrar un argumento de una historia de un modo más convincente.
- Quitar información técnica innecesaria de la prosa.
- En particular cuando son interactivos y permiten la exploración, ofrecen transparencia respecto de su proceso de información a sus lectores.

Estos roles sugieren que debiera comenzar temprano y a menudo con visualizaciones en sus informes, sea o no que comience con datos o registros electrónicos. No lo considere un paso por separado, algo a considerar una vez que la historia en gran medida ya esté escrita. Permita que este trabajo ayude a guiar su tarea periodística.

Comenzar a veces significa simplemente poner las notas que ya tomó en formato visual. Considere el gráfico en la Figura 6-2, que se publicó en el Washington Post en 2006.

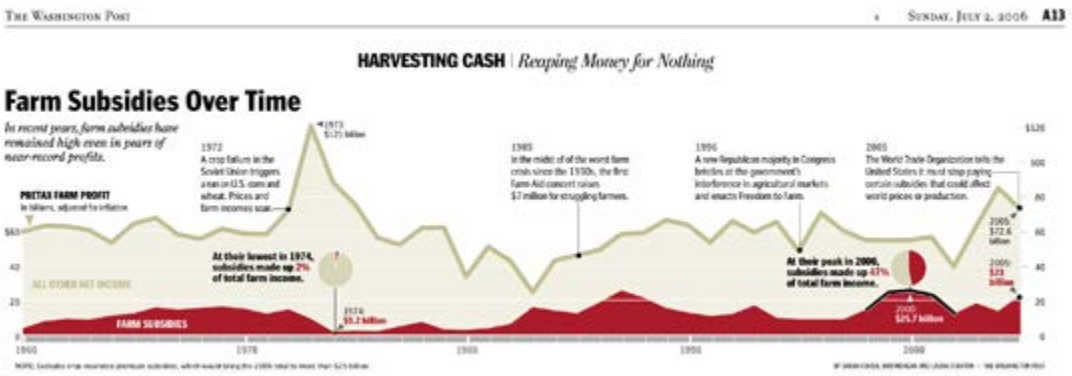


Figure 2. Subsidios agropecuarios a lo largo del tiempo (Washington Post)

Muestra la porción del ingreso agropecuario asociado con subsidios y eventos claves en los últimos 45 años, y fue creado a lo largo de una serie de meses. Encontrar datos que pudieran utilizarse para largos períodos de tiempo con definiciones y significados similares fue un desafío. Investigar todas las alzas y bajas nos ayudó a tener presente el contexto mientras hacíamos el resto de nuestro trabajo. También significó que la tarea estuvo prácticamente acabada antes de que se escribieran las historias.

A continuación, algunos consejos sobre el uso de visualizaciones para comenzar a explorar sus conjuntos de datos.

Consejo 1: Use pequeños múltiplos para orientarse rápidamente en un conjunto de datos grande

Usé esta técnica en el Washington Post cuando seguimos una pista de que la administración de George W. Bush estaba otorgando subsidios por motivos políticos y no de fondo. La mayoría de estos programas de ayuda se guían por fórmulas y otros han sido financiados desde hace años, por lo que estábamos curiosos por ver si pudiéramos encontrar un patrón analizando casi 1500 casos diferentes discrecionales.

Creé un gráfico para cada programa, con puntos rojos indicando un año con elecciones presidenciales y puntos verdes indicando elecciones parlamentarias. El problema: sí, había un salto en los seis meses antes de la elección presidencial en varios de estos programas – los puntos rojos con los números pico junto a ellos- pero es el año electoral equivocado. El patrón apareció de modo sistemático durante la elección presidencial del 2000 entre Al Gore y George W. Bush, no la elección de 2004.

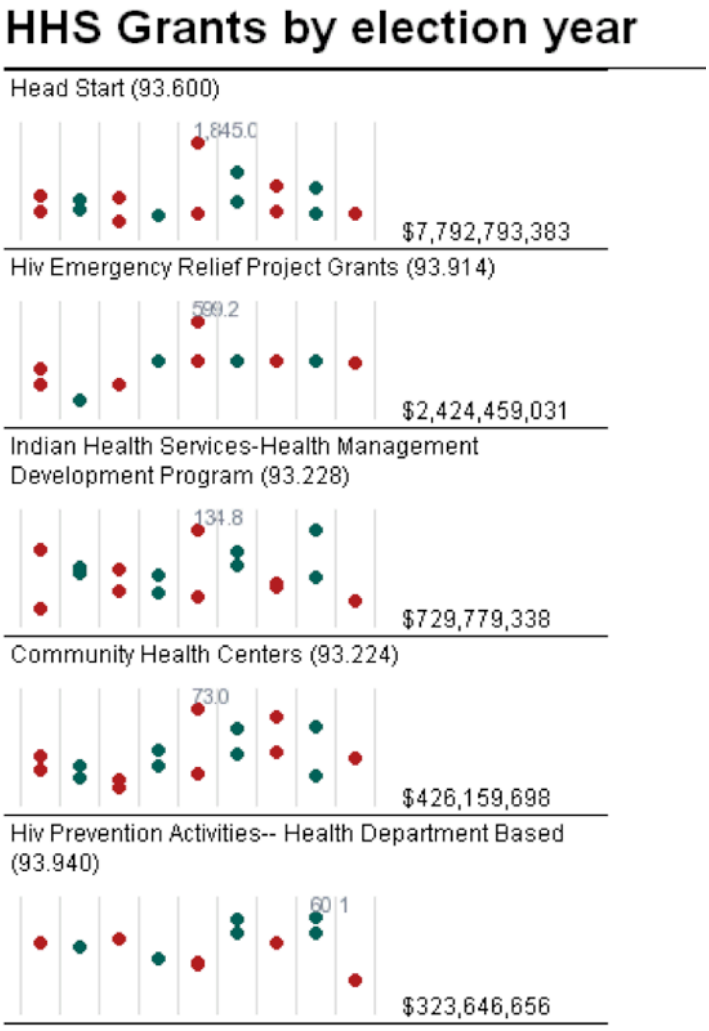


Figure 3. Subsidios HHS: los sparklines ayudan a encontrar historias (Washington Post)

Esto fue realmente fácil de ver en una serie de gráficos en vez de una tabla numérica, y un formulario interactivo nos permitió verificar varios tipos de subsidios, regiones y entes. Los mapas con pequeños múltiplos pueden ser una manera un modo de mostrar tiempo y lugar en una imagen estática que es fácil de comparar, a veces incluso más fácil que la versión interactiva.

Este ejemplo fue creado con un programa breve escrito en PHP, pero ahora es mucho más fácil de hacer con Excel 2007 y los _sparklines_de 2010. Edward Tufte, el experto en visualización, inventó estos "gráficos intensos, simples, como palabras" para transmitir información con una sola mirada basados en un conjunto de datos grandes. Ahora se los ve en todas partes, desde los pequeños gráficos bajo las cotizaciones de la bolsa hasta los records de triunfos y derrotas en deportes.

Consejo 2: Mire sus datos del derecho y del revés

Cuando trata de entender una historia o un conjunto de datos, no hay una manera equivocada de mirar; inténtelo de todas las maneras que se le ocurren y tendrán muchas perspectivas distintas. Si está informando sobre criminalidad, podría ver un conjunto de cuadros con cambios en los crímenes violentos en un año; otro podría indicar el cambio porcentual; otro podría ser una comparación con otras ciudades, y otro podría ser de cambios en el tiempo. Use cifras crudas, porcentajes e índices.

Mírelos en distintas escalas. Trate de seguir la regla de que el eje de las x debe estar en cero. Luego viole esa regla y vea si encuentra más cosas. Pruebe con logaritmos y raíces cuadradas para datos con distribuciones extrañas.

Tenga en mente las investigaciones hechas con percepciones visuales. Los experimentos de William Cleveland mostraron que los ojos ven cambios en una imagen cuando la inclinación promedio es de alrededor de 45 grados. Esto sugiere que hay que ignorar las admoniciones de que siempre se debe comenzar desde cero y en cambio trabajar pensando en los gráficos que permitan ver más cosas. Otras investigaciones sobre epidemiología han sugerido que se puede encontrar un nivel determinado como delimitador para su cuadro. Cada uno de estos modos permite ver los datos de modo diferente. Cuando ya no le dicen nada nuevo sabe que acabó su tarea.

Consejo 3: No dé nada por supuesto

Ahora que ha mirado sus datos de distintos modos, probablemente habrá encontrado registros que no parecen correctos: puede no haber entendido lo que significaban o hay

algunos casos fuera de lo común que parecen errores de tipeo o hay tendencias que parecen invertir las cosas.

Si quiere publicar algo basado en sus primeras exploraciones o en una visualización, tiene que resolver estas cuestiones y no dar nada por supuesto. Son historias interesantes o errores; desafíos interesantes a las verdades sabidas o confusiones.

No es inusual que gobiernos municipales den planillas de cálculo llenas de errores, y es también fácil confundirse con la jerga oficial en un conjunto de datos.

Primero, vuelva a mirar su trabajo. ¿Ha leído la documentación, sus advertencias, y existe el problema en la versión original de los datos? Si todo lo hecho por usted parece estar bien, entonces es hora de tomar el teléfono. Tendrá que conseguir resolverlo si quiere usarlo, por lo que mejor ponerse ya mismo en marcha.

Dicho esto, no todo error es importante. En los registros de finanzas de campañas electorales, es común que haya varios cientos de códigos postales que no existen en una base de datos de 100.000 registros. Siempre que no sean todos en la misma ciudad o estén relacionados con un mismo candidato, el registro ocasional equivocado simplemente no importa.

La pregunta que debe hacerse: ¿si fueran a usar esto, los lectores tendrían una visión acertada en lo esencial de lo que dicen los datos?

Consejo 4: Evite obsesionarse con la precisión

La contracara de no hacer suficientes preguntas es obsesionarse con la precisión antes de que importe. Sus gráficos exploratorios debieran ser correctos en general, pero no se preocupe si tiene varios niveles de redondeo, si no suman exactamente 100 por ciento o si le faltan datos de 1 o 2 años en 20 años. Esto es parte del proceso exploratorio. Aún así verá las grandes tendencias y sabrá lo que tiene que buscar antes de que llegue el momento de publicar.

De hecho, podría considerar eliminar las marcas y los indicadores de escala, como en los cuadros de más arriba, para tener una mejor visión del sentido general de los datos.

Consejo 5: Cree cronologías de casos y eventos

Al comienzo de cualquier historia compleja, comience a crear cronologías de eventos y casos claves. Puede usar Excel, un documento en Word, o una herramienta especial como TimeFlow para la tarea, pero en algún punto encontrará un conjunto de datos que puede

usar como base de referencia. Releerlo periódicamente le mostrará qué baches tiene en su informe que deben cubrirse.

Consejo 6: Reúnase desde el comienzo y a menudo con el departamento gráfico

Intercambie ideas respecto de gráficos posibles con los ilustradores y diagramadores de su redacción. Ellos tendrán buenas alternativas para ver sus datos, sugerencias de cómo podría funcionar interactivamente, y saben cómo conectar datos e historias. Le hará mucho más fácil su tarea si sabe desde el comienzo qué es lo que tiene que buscar o si puede alertar a su equipo de que no es posible realizar determinado gráfico cuando no logra obtener los datos necesarios.

Consejos para la publicación de datos

Puede haber pasado solo unos pocos días o unas pocas horas en su exploración, o puede haber tardado meses en reunir la información para su historia. Pero cuando se acerca el momento de publicarla, hay dos aspectos que se vuelven importantes.

¿Se acuerda de ese año que le faltó en sus exploraciones iniciales? De pronto ya no puede avanzar más sin esos datos. ¿Todos los datos con problemas que ignoró en sus informes? Ahora vuelven como fantasmas. La razón es que no se puede simplemente esquivar los problemas. Se tiene todo lo que se necesita para un gráfico o no se lo tiene, y no hay solución intermedia.

El esfuerzo de recolección de los datos tiene que coincidir con lo que requiere el gráfico interactivo:: No hay modo de ocultarse en un gráfico interactivo. Si realmente va a hacer que sus lectores puedan explorar los datos de cualquier manera que quieran, entonces cada elemento de los datos tiene que ser lo que dice ser. Los usuarios pueden encontrar cualquier error en cualquier momento, y eso podría afectarlo por meses o años. Si está creando su propia base de datos, tiene que prever la corrección de errores, el control de datos y la edición del texto de toda la base de datos. Si está usando archivos oficiales, debe decidir cuánto los va a controlar y qué piensa hacer cuando encuentre el inevitable error.

Diseñe pensando en dos tipos de lectores

> El gráfico –sea un elemento interactivo que se presenta solo o una visualización estática que acompaña su artículo- debe satisfacer a dos tipos diferentes de lectores. Debe ser fácil de entender de un vistazo, pero lo suficientemente complejo como para ofrecer algo interesante a la gente que quiere ir más allá. Si lo hace interactivo, asegúrese de que sus lectores obtengan algo más que una sola cifra o número.

Transmita una idea y luego simplifique

> Asegúrese de que haya una cosa que quiere que la gente vea. Decida cuál es la impresión general que quiere que tenga el lector y haga que todo lo demás desaparezca. En muchos casos, esto significa eliminar información aún cuando Internet le permita proveer todo. A menos que su objetivo principal sea la transparencia en su actividad periodística, la mayor parte de los detalles que ha recogido en su línea de tiempo y cronología simplemente no son demasiado importantes. En un gráfico estático serán intimidantes. En un gráfico interactivo serán aburridos.

— Sarah Cohen, Duke University

El uso de visualizaciones para narrar historias

La visualización de datos amerita su consideración por varios motivos. No solo puede ser llamativamente hermosa y atraer la atención –recurso social valioso para compartir y atraer a los lectores- también aprovecha una ventaja cognitiva poderosa: la mitad del cerebro humano está dedicado a procesar información visual. Cuando se presenta a un usuario un gráfico informativo, se está llegando a él a través de la vía de banda más ancha de acceso a la mente. Una visualización de datos bien diseñada puede ofrecer a los que la ven una impresión inmediata y profunda, e ir al grando de la cuestión sin enredarse con todo lo que hay en una historia compleja.

Pero a diferencia de otros medios visuales –tales como la fotografía y el video- la visualización de datos también está enraizada en hechos mensurables. Aunque atractiva estéticamente, tiene menos carga emocional, está más interesada en echar luz que calor. En una era de medios con foco estrecho que a menudo están hechos a medida de públicos con puntos de vista particulares, la visualización de datos (y el periodismo de datos en general) ofrece la oportunidad tentadora de narrar historias orientadas principalmente por los hechos y no el fanatismo.

Lo que es más, al igual que otras formas de periodismo narrativo, la visualización de datos puede ser efectiva tanto para presentar noticias nuevas –transmitiendo rápidamente nueva información al estilo de la ubicación de un accidente y el número de víctimas- como artículos de fondo, donde puede profundizar en un tema y ofrecer una nueva perspectiva, ayudándolo a ver algo familiar de un modo completamente nuevo.

Ver lo familiar de un modo nuevo

De hecho, la capacidad de las visualizaciones de datos de cuestionar las verdades aceptadas es ejemplificada por un gráfico interactivo publicado por The New York Times a fines de

2009, un año después de que comenzara la crisis económica global. Con la tasa de desempleo nacional de Estados Unidos en torno 9 %, los usuarios podían analizar la población del país con varios filtros demográficos y educativos, para ver lo dramáticos que eran los cambios en las tasas. Resultó que la tasa iba, de menos del 4% para mujeres de edad media con títulos avanzados, hasta casi la mitad de todos los jóvenes negros que no habían terminado la escuela secundaria, y además esta disparidad no era nada nuevo: dato subrayado por líneas de fiebre que mostraban los valores históricos para cada uno de esto grupos.

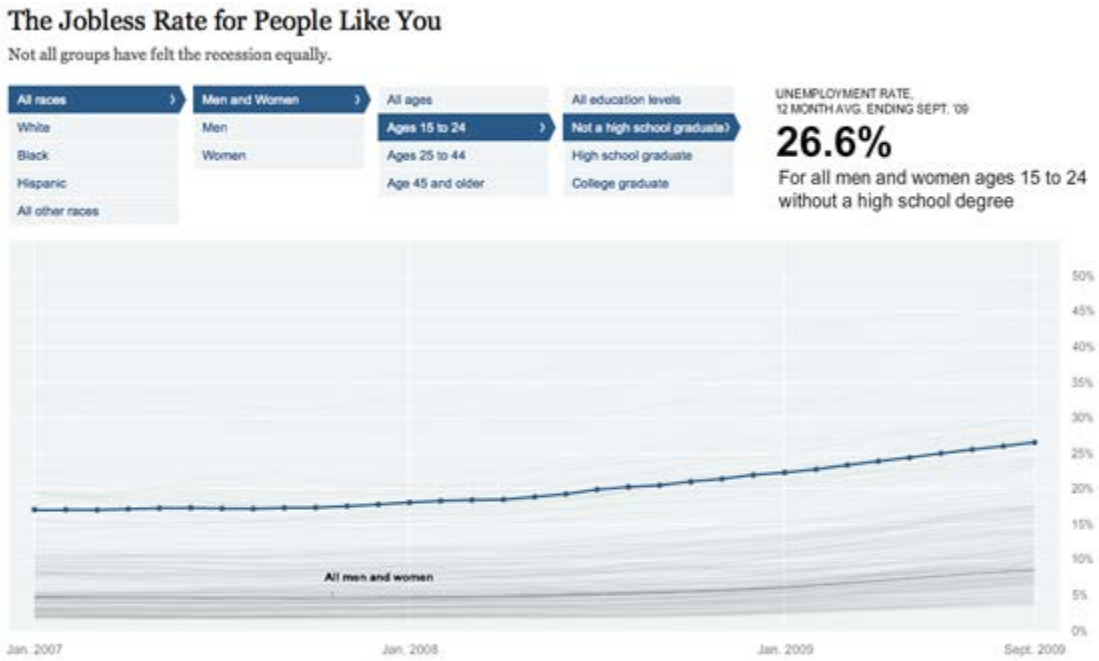


Figure 4. La tasa de desempleo para personas como usted (New York Times)

Incluso cuando ya ha dejado de mirarla, una buena visualización de datos se mete en su cabeza y deja un modelo mental duradero de un hecho, una tendencia o un proceso. ¿Cuánta gente vio la animación de tsunamis presentada por los investigadores en diciembre de 2004, que mostraba olas en cascada irradiando desde un terremoto indonesio a través del océano Índico, amenazando a millones de residentes costeros en el sur de Asia y África oriental?

Las visualizaciones de datos –y las asociaciones estéticas que engendran- pueden incluso convertirse en hitos culturales, tales como la representación de las profundas divisiones políticas en Estados Unidos luego de las elecciones del 2000 y 2004, cuando los estados republicanos “rojos” llenaban el centro del país y los estados demócratas “azules” formaban núcleos en el noreste y el lejano oeste. No importa que en los medios de EE.UU. antes del 2000 las principales cadenas de medios habían intercambiado el azul y el rojo muchas veces para representar a cada partido, optando algunas incluso por alternar cada cuatro años. De allí el recuerdo de algunos estadounidenses de la victoria épica en 49 estados “azules” para

los republicanos en 1984 liderada por Ronald Reagan.
Pero por cada gráfico que engendra un cliché visual, aparece otro que aporta un poderoso testimonio fáctico, tal como el mapa de 2006 de The New York Times que usó círculos de distintos tamaños para mostrar donde vivían cientos de miles de evacuados de New Orleans, desparramados por todo el continente por una mezcla de vínculos personales y programas de relocalización. ¿Estos evacuados "varados" podrían volver alguna vez a sus hogares?

Ahora que hemos hablado del poder de la visualización de datos, es justo preguntar cuándo debemos usarla y cuando *no*. Primero analizaremos algunos ejemplos en los que la visualización de datos podría ser útil para ayudar a narrar una historia a sus lectores.

Mostrar el cambio a lo largo del tiempo

Quizás el uso más común de la visualización de datos –personificado en el humilde gráfico de fiebre- es mostrar cómo han cambiado valores a lo largo del tiempo. El crecimiento de la población china desde 1960 o el salto en el desempleo desde la caída económica de 2008, son buenos ejemplos. Pero las visualizaciones de datos también pueden mostrar de modo muy poderoso el cambio a lo largo del tiempo a través de otras formas gráficas. El investigador portugués Pedro M. Cruz utilizó cuadros con forma de círculos animados para mostrar dramáticamente la declinación de los imperios europeos occidentales desde comienzos del siglo XIX. Medidos por su población total, Gran Bretaña, Francia, España y Portugal estallan como burbujas al lograr la independencia sus territorios extranjeros. Allí va México, Brasil, Australia, la India, y esperen... allí van muchas colonias africanas a comienzos de la década de 1960, con lo que casi desaparece Francia.

Un gráfico del Wall Street Journal muestra el número de meses que les llevó a varios empresarios llegar al número de US$ 50 millones en ganancias. Creado utilzando Tableau Public, una herramienta de gráficos y análisis de datos gratuita, la comparación semeja las estelas superpuestas que dejan múltiples aeronaves al despegar, algunas rápidas, otras lentas, algunas pesadas,.

Hablando de aviones, otro gráfico interesante que muestra el cambio en el tiempo presenta la participación en el mercado de las principales aerolíneas estadounidenses durante varias décadas de concentración en el sector.

Luego de que la administración Carter desregulara la aviación de pasajeros, una seguidilla de adquisiciones financiadas con deuda creó compañías de aeronavegación nacionales a partir de pequeñas aerolíneas regionales, como ilustra este gráfico de The New York Times.

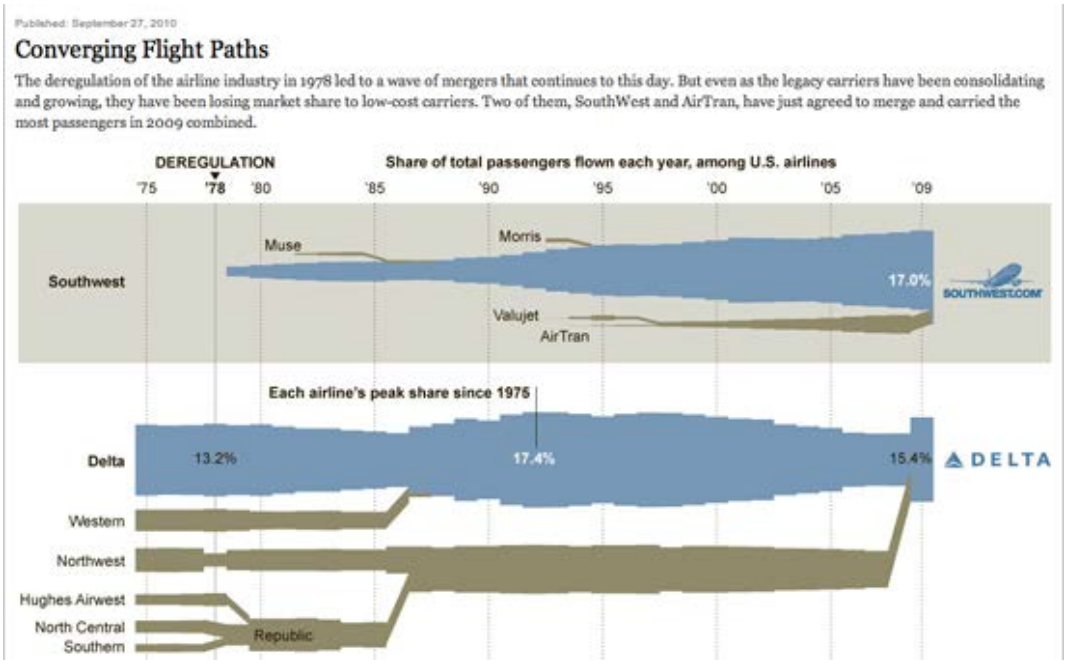


Figure 5. Rutas de vuelo convergentes (New York Times)

Dado que casi todos los lectores casuales ven el eje horizontal, de las "x" de un cuadro, como representa el tiempo, a veces es fácil creer que *todas* las visualizaciones deben mostrar el cambio en el tiempo.

Comparar valores

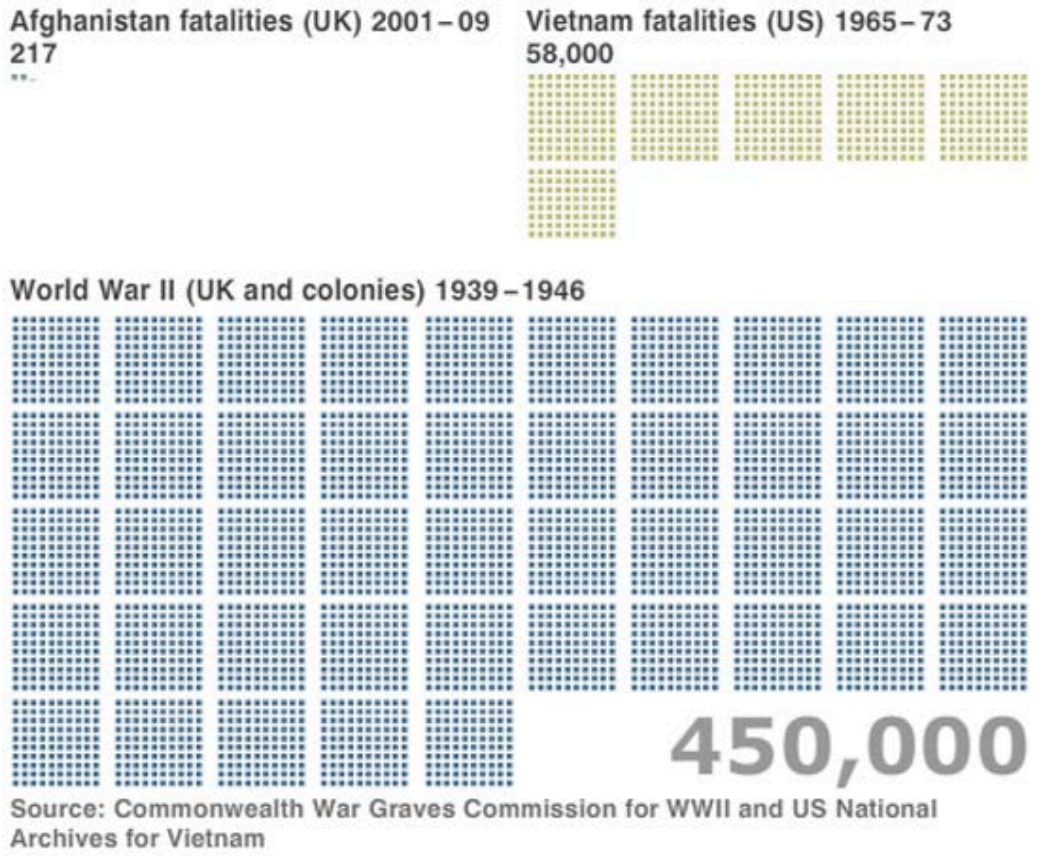


Figure 6. Contar el costo humano de la guerra (BBC)

La visualización de datos también es útil cuando se trata de ayudar a los lectores a comparar dos o más valores discretos, sea para poner en contexto la pérdida trágica de hombres y mujeres de las fuerzas armadas en los conflictos de Irak y Afganistán (comparándolos con los tantos miles de muertos en Vietnam y los millones que murieron en la segunda Guerra Mundial, como hizo la BBC en un slideshow de transparencias animadas que acompaña su base de datos de bajas); o cuando el National Geographic, utilizando un cuadro muy minimalista, mostró cuanto mayores son las probabilidades de morir de enfermedad coronaria (probabilidad de 1 en 5) o infarto (1 en 24) que en accidentes de aviación (1 en 5051) o por una picadura de abeja (1 en 56789), mostrando las probabilidades relativas de las distintas causas de muerte (todo dominado por un arco inmenso que representa las probabilidades generales de morirse: 1 en 1).

La BBC, en colaboración con la agencia Berg Design, también desarrolló el sitio "Dimensions", que le permite superponer los contornos de los principales eventos mundiales –el derrame de petróleo de la plataforma marina Deepwater Horizon o las inundaciones paquistaníes, por ejemplo- a un Google Map de su propia comunidad.

Mostrar conexiones y flujos

La introducción del ferrocarril de alta velocidad en Francia en 1981 no achicó realmente el país, pero una representación visual ingeniosa muestra cuanto menos tiempo lleva alcanzar distintos destinos comparado con el ferrocarril convencional. Una grilla superpuesta al país aparece de forma cuadrada en la imagen de "antes", pero se ve aplastada hacia el centro, París, en la de "después", mostrando no solo que los destinos están más "cerca", sino que la mayor ganancia de tiempo se da en la primera parte del viaje, antes de que los trenes tengan que bajar la velocidad al llegar a vías no mejoradas.

Para comparar entre dos variables distintas, vea el cuadro de Ben Fry evaluando el desempeño de equipo de Baseball de las Grandes Ligas relativo a lo que ganan sus jugadores. Una línea dibujada en rojo (mal desempeño) o azul (buen desempeño) conecta los dos valores, dando de forma práctica una sensación de qué dueños de equipos lamentan lo mal que le ha ido con jugadores caros. Más aún, el recorrido de una línea de tiempo ofrece una imagen vívida de la competencia por el campeonato.

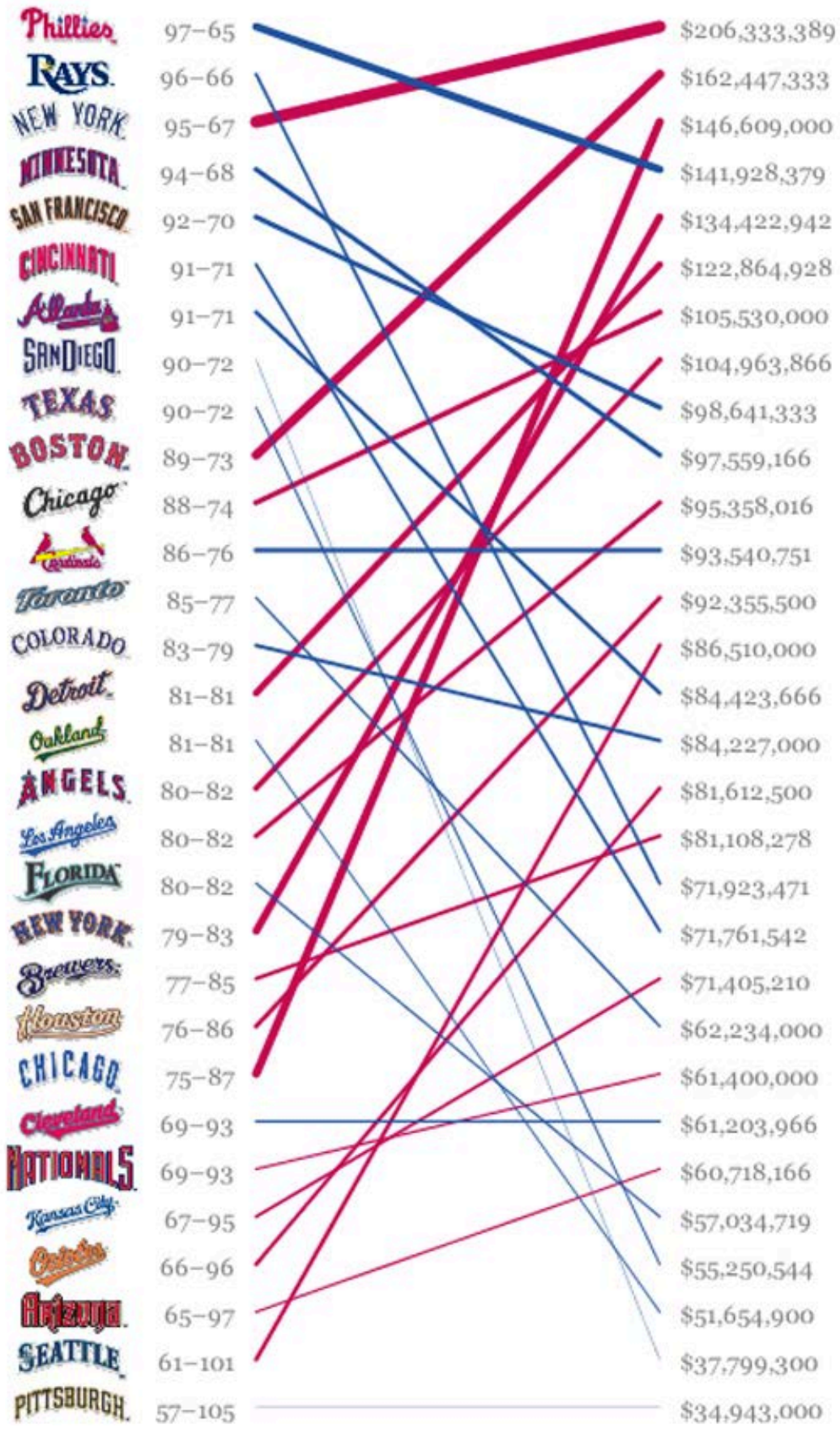


Figure 7. Salario vs. desempeño (Ben Fry)

Diseñar con datos

Similares a las conexiones gráficas en un sentido, los diagramas de flujo también codifican información en las líneas de conexión, generalmente de acuerdo al grosor y/o el color de las mismas. Por ejemplo, con la Eurozona en crisis y varios miembros incapacitados para pagar sus deudas, The New York Times buscó desentrañar la madeja de deudas que vincula a los miembros de la UE con sus socios comerciales al otro lado del Atlántico y en Asia. En un "estado" de la visualización, el ancho de las líneas refleja el monto del crédito que pasa de un país a otro, y tonos que van del amarillo al naranja indican lo "preocupante" de la deuda, es decir, la improbabilidad de su repago.

Sobre un tópico más feliz, la revista National Geographic produjo un gráfico que parece simple, mostrando las conexiones de tres ciudades de EE.UU. –New York, Chicago y Los Ángeles- con regiones productoras de vino importantes, y cómo los métodos de transporte con los que se trae el producto de cada una de las fuentes podrían resultar en una huella de carbono drásticamente diferente, haciendo que para los neoyorquinos, por ejemplo, comprar en Burdeos sea más "verde" que comprar vino de California.

"SourceMap", un proyecto iniciado en la escuela de estudios empresarios del MIT, usa diagramas de flujo para analizar rigurosamente el abastecimiento global de productos manufacturados, sus componentes y materias primas. Gracias a mucha investigación un usuario ahora puede buscar productos que van desde zapatos de marca Ecco hasta jugo de naranja, y saber qué rincón del globo es su origen y su correspondiente huella de carbono.

Mostrar jerarquías

En 1991 el investigador Ben Shneiderman inventó una nueva forma de visualización llamada "treemap" que consiste de múltiples cajas concéntricas. El área de cada caja indica la cantidad que representa, en sí misma y como adición de sus contenidos. Se trate de visualizar un presupuesto nacional dividido por entes oficiales y sub-entes, la bolsa de valores por sector y compañía, o un lenguaje de programación por clases y sub-clases, el "treemap" es una interfaz compacta e intuitiva para representar un ente y sus partes constituyentes. Otro formato efectivo es el dendrograma, que se ve como un cuadro de organización más típico, donde las subcategorías salen de un solo tronco central.

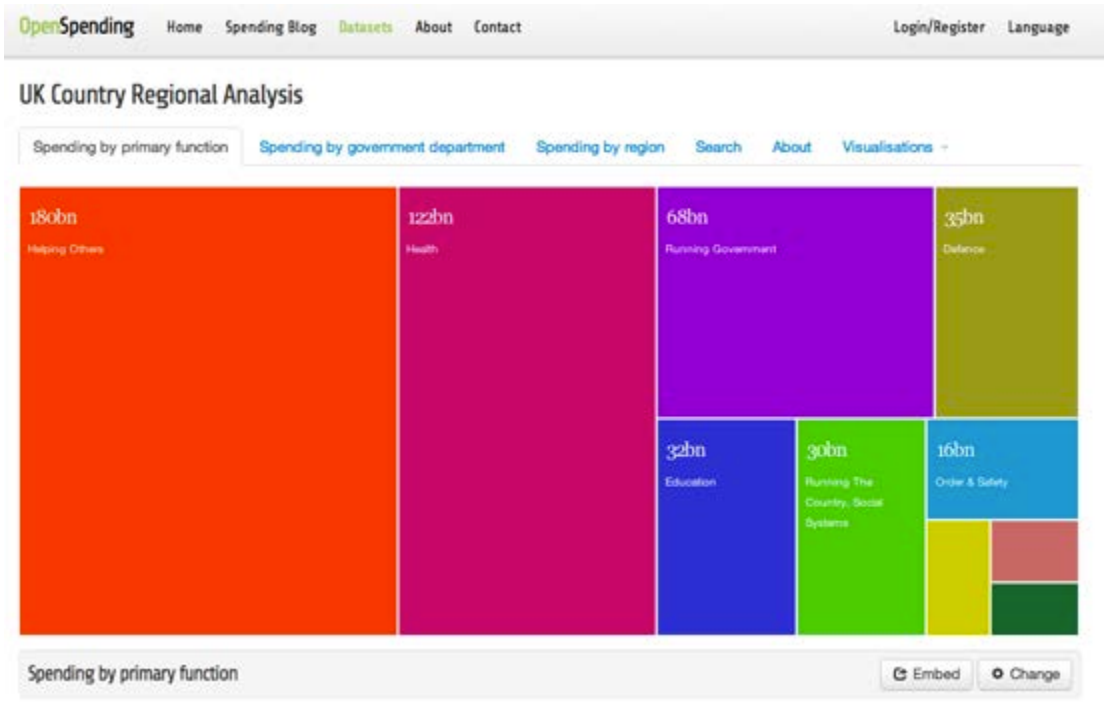


Figure 8. OpenSpending.org (Open Knowledge Foundation)

Explorar grandes bases de datos

A veces la visualización de datos es muy efectiva para tomar información familiar y mostrarla desde un ángulo totalmente nuevo, ¿pero qué sucede cuándo se tiene información nueva que la gente quiere navegar? La era de los datos trae consigno descubrimientos nuevos sorprendentes casi todos los días, desde el brillante análisis de fotos de Flickr de Eric Fischer hasta la difusión por la municipalidad de la ciudad de New York de miles de evaluaciones de docentes hasta ese momento confidenciales.

Estas bases de datos son más poderosas cuando los usuarios pueden meter mano y llegar hasta la información que les resulta más relevante.

A comienzos de 2010, se le dio acceso a The New York Times a los registros privados de Netflix de qué películas se alquilan más en cada área. Si bien Netflix se negó a difundir las cifras en crudo, el Times creó una base de datos interactiva atractiva que permite a los usuarios explorar las 100 películas más alquiladas en 12 zonas metropolitanas de EE.UU., subdivididas hasta el nivel de código postal. Un “mapa de calor” graduado por colores superpuesto a cada comunidad permitía a los usuarios ver rápidamente dónde un título en particular era más popular.

Hacia el fin del mismo año, el Times publicó los resultados del censo decenal de los Estados Unidos, apenas horas después de que fuera difundido. La interfaz, creada con Adobe Flash, ofrecía una cantidad de opciones de visualización y permitía a los usuarios llegar al nivel de cada bloque del censo en el país (de 8,2 millones) para ver la distribución de residentes por raza, ingreso y educación. Tal era la resolución de la base de datos que cuando se buceaba

en el conjunto de datos en las primeras horas después de su publicación uno podía llegar a preguntarse si era la primera persona del mundo en explorar determinado rincón de la base de datos.

Entre los usos igualmente aplaudibles de la visualización como presentación de una base de datos se incluyen la investigación por la BBC de muertes en las rutas y muchos de los intentos de indexar rápidamente grandes cúmulos de datos como la difusión por WikiLeaks de los registros de guerra de Irak y Afganistán.

Figure 9. Cada muerte en las rutas de Gran Bretaña 1999-2010 (BBC)

La regla de 65k

Al recibir la primera pila de datos de los registros de la guerra de Afganistán de WikiLeaks, el equipo que los procesaba comenzó a manifestar su entusiasmo por tener acceso a 65.000 registros militares.

Esto inmediatamente hizo sonar la alarma entre quienes tenían experiencia con Excel de Microsoft. Gracias a una limitación histórica del modo en que se accede a las filas, la herramienta de importación de Excel no procesa más de 65.536 registros. En este caso se descubrió que faltaban 25.000 filas.

La moraleja de esta historia (además de evitar usar Excel para tales tareas) es siempre desconfiar de cualquiera que alardee de tener 65.000 filas de datos.

— Alastair Dant, the Guardian

Imaginar resultados alternativos

En The New York Times, el “cuadro puercoespín” de Amanda Cox con proyecciones de déficit de EE.UU. trágicamente optimistas a lo largo de los años, muestra cómo a veces lo que sucedió es menos interesante que lo que no sucedió. La curva de Cox que muestra el alza del déficit fiscal luego de una década de guerra y exenciones impositivas muestra lo poco realistas que pueden resultar las expectativas del futuro.

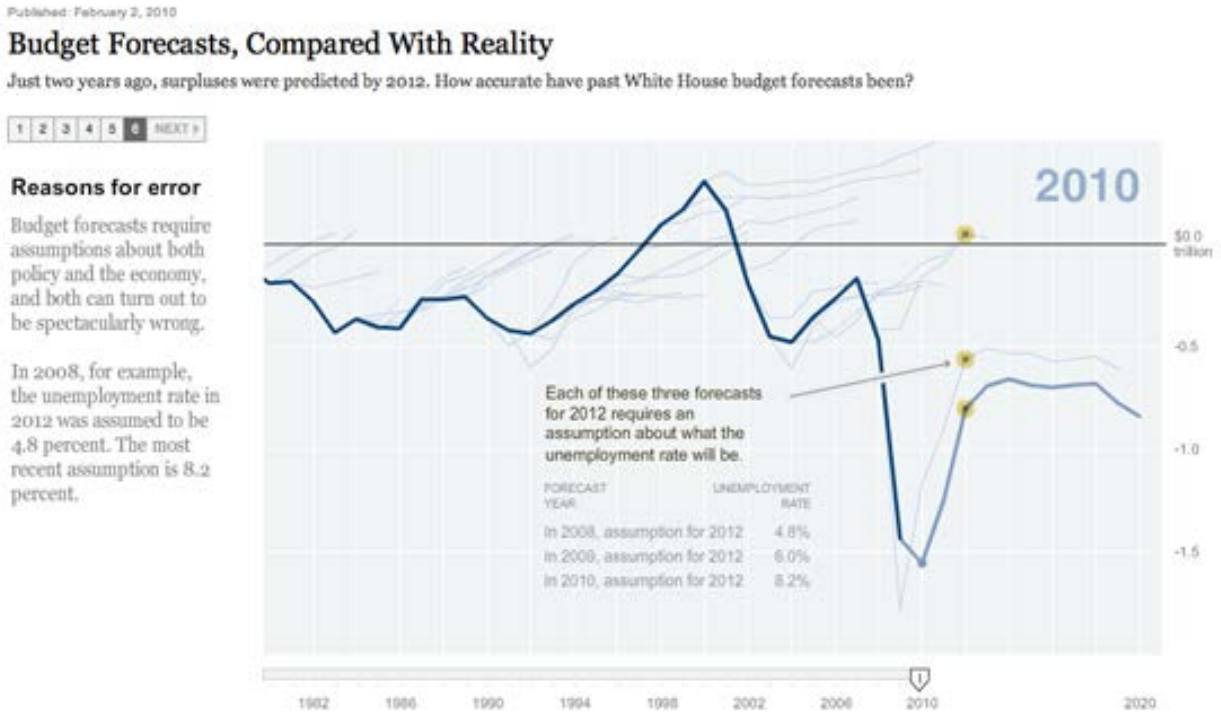


Figure 10. Pronósticos presupuestarios, comparados con la realidad (New York Times)

Bret Victor, un diseñador de interfaces de Apple de larga trayectoria (y originador de la teoría “_kill math_” o “matar la matemática” de visualización para comunicar información cuantitativa), ha hecho un prototipo de una especie de documento que se actualiza de conjunto cada vez que se modifica un dato. En su ejemplo, las ideas de conservación de la energía incluyen premisas modificables, por la que un paso simple como apagar las luces de los cuartos en los que no hay gente podría ahorrar a los estadounidenses la generación de 2 a 40 plantas de carbón. Cambiar el porcentaje que aparece en el medio de un párrafo de texto hace que el resto de la página se actualice en consonancia.

Para más ejemplos y sugerencias, aquí va el link con una lista de links de distintos usos de visualizaciones, mapas y gráficos interactivos compilada por Matthew Ericson de The New York Times.

Cuándo no usar visualización de datos

En definitiva, la visualización de datos efectiva depende de contar con información buena, limpia, precisa y significativa. Así como muchas citas, datos, y descripciones alimentan el buen periodismo narrativo, la visualización de datos es tan buena como los datos que la alimentan.

En qué casos su historia puede ser mejor narrada a través de texto o multimedia:: A veces los datos por sí solos no narran la historia del modo más convincente. Si bien un cuadro simple que ilustre una tendencia o una estadística puede ser útil, una narrativa que relate las consecuencias de una cuestión en el mundo real puede ser más inmediata y de mayor impacto para un lector.

Cuando tiene muy pocos datos

> Se ha dicho que “una cifra aislada no quiere decir nada”. Una frase común de los editores de noticias en respuesta a una estadística citada es: “¿comparado con qué?” ¿La tendencia sube o baja? ¿Qué es lo normal?

Cuando tiene escasa variación en su datos, sin una tendencia o conclusión clara:: A veces organiza sus datos en Excel o una aplicación similar y descubre que la información es ruidosa, tiene mucha fluctuación y muestra una tendencia relativamente chata. ¿Conviene elevar la base de cero a justo debajo del valor más bajo para dar un poco más de forma a la línea? ¡No! Parece que lo que tiene son datos ambiguos y necesita buscar y analizar un poco más.

Cuando un mapa no es un mapa

> A veces el elemento espacial no es significativo ni convincente, o distrae la atención de las tendencias numéricas pertinentes, como el cambio en el tiempo o mostrar las similitudes entre zonas no adyacentes.

Cuando bastaría con una tabla

> Si cuenta con relativamente pocos puntos de datos pero tiene información que podría ser útil para algunos de sus lectores, considere simplemente presentar los datos en forma tabular. Es limpio, de fácil lectura y no crea expectativas no realistas de una “historia”. De hecho, las tablas pueden ser una forma muy eficiente y elegante de presentar información básica.

— Geoff McGhee, Stanford University

Cuadros diferentes dicen cosas diferentes

En este mundo digital, con la promesa de experiencias 3D de inmersión, tendemos a olvidar que por tanto tiempo solo tuvimos tinta en papel. Ahora pensamos en este medio estático, plano, como un ciudadano de segunda, pero de hecho a lo largo de los siglos que hemos

estado escribiendo e imprimiendo, hemos logrado una increíble riqueza de conocimiento y prácticas para representar los datos en una página. Aunque los cuadros, las visualizaciones de datos y las infografías interactivas son la gran moda, nos llevan a dejar de lado muchas de las mejores prácticas que hemos aprendido. Solo estudiando la historia de cuadros y gráficos bien logrados es que podemos entender esos conocimientos acumulados y aprovecharlos con los nuevos medios.

Algunos de los cuadros y gráficos más famosos derivan de la necesidad de explicar mejor tablas de datos densas. William Playfair era un políglota escocés que vivió desde fines del siglo XVIII hasta comienzos del XIX. Por sí solo presentó al mundo muchos de los cuadros y gráficos que seguimos utilizando hoy. En su libro de 1786, *Commercial and Political Atlas* (Atlas Comercial y Político), Playfair introdujo el gráfico de barras para mostrar claramente las cantidades de importaciones y exportaciones de Escocia de un modo nuevo y visual.

Luego popularizó el cuadro de torta en su libro de 1801, *Statistical Breviary* (Breviario Estadístico). La necesidad de estas nuevas formas de cuadros y gráficos provino del comercio, pero con el paso del tiempo aparecieron otros que fueron utilizados para salvar vidas. En 1854 John Snow creó su ahora famoso “Cholera Map of London” (Mapa del Cólera de Londres), agregando una pequeña barra negra sobre cada dirección en la que se reportó un incidente. Con el tiempo, se pudo ver cualquier densidad evidente de la epidemia y actuar en consecuencia para contener el problema.

Con la práctica los practicantes de estos nuevos cuadros y gráficos se volvieron más audaces y experimentaron más allá, llevando el recurso a los niveles que conocemos hoy. André-Michel Guerry fue el primero en publicar la idea de un mapa en el que regiones individuales se identificaban con distintos colores basados en alguna variable. En 1829 creó el primer coroplético dando distinto tono a las regiones de Francia representando niveles de criminalidad. Hoy vemos tales mapas utilizados para mostrar los resultados de encuestas políticas, quién votó por quién, distribución de la riqueza y muchas otras variables con distribución geográfica. Parece una idea tan simple pero aún hoy es difícil de dominar y comprender si no se la usa juiciosamente.

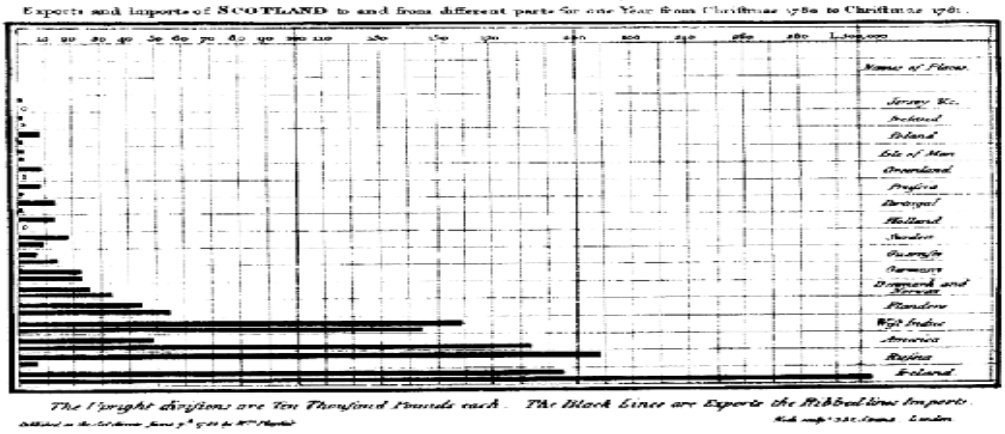


Figure 11. Uno de los primeros gráficos de barras (William Playfair)

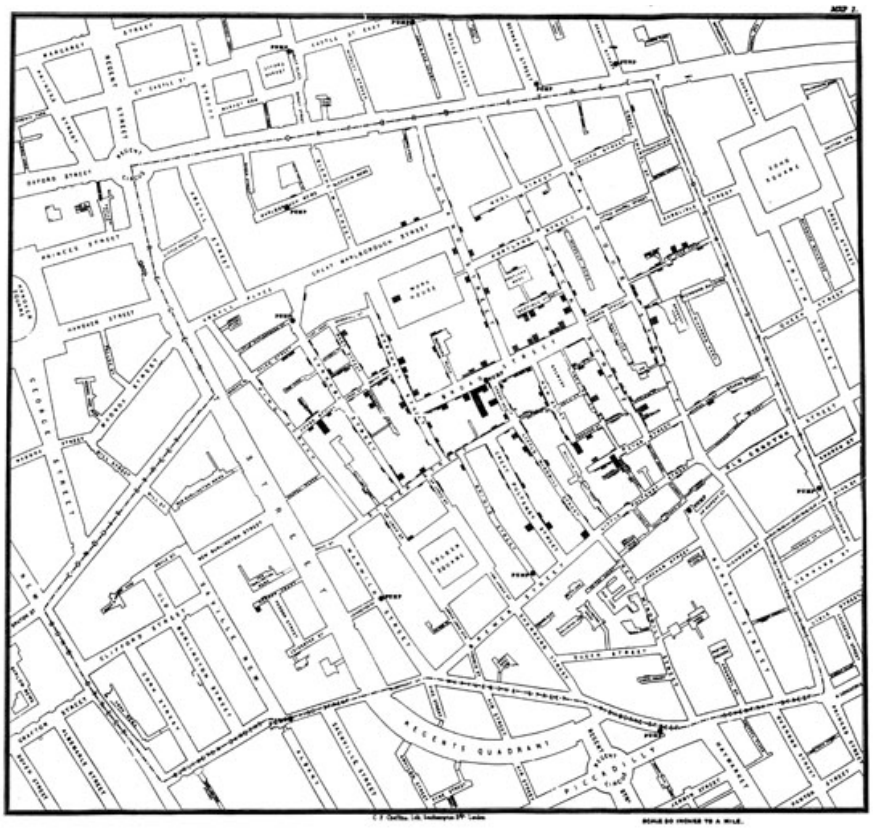


Figure 12. Mapa del cólera de Londres (John Snow)

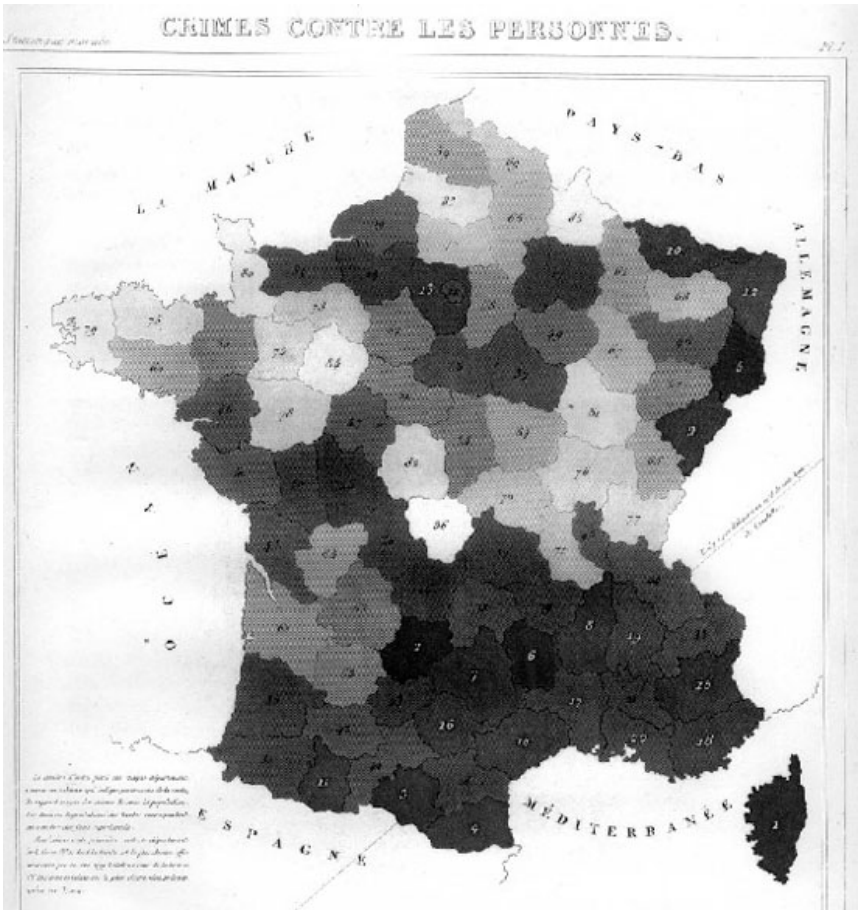


Figure 13. Mapa coroplético de Francia mostrando niveles de criminalidad (André-Michel Guerry)

Hay muchas herramientas que un buen periodista tiene que entender y tener en su herramental para construir visualizaciones. En vez de zambullirse directo en la parte más honda de la piscina, es importante tener una base en materia de cuadros y gráficos. Todo lo que cree tiene que originarse en una serie de cuadros y gráficos atómicos. Si puede dominar lo básico, entonces puede construir visualizaciones más complejas que se arman a partir de estas unidades básicas.

Dos de los tipos más básicos de gráficos son los de barras y de curvas. Si bien son muy similares en cuanto a los casos en los que se usan, también pueden diferir mucho en su significado. Tomemos por caso las ventas de una compañía para cada mes del año. Tendríamos las 12 barras que representan el monto de dinero que entra cada mes (Figure 14).

Analicemos por qué esto debe hacerse con barras en vez de un gráfico de curvas. Los gráficos de líneas son ideales para datos continuos. En el caso de las cifras de ventas, se trata de la suma de cada mes, no datos continuos. En base a las barras, sabemos que en enero, la compañía tuvo ingresos por $ 100 y en febrero $ 120. Si convertimos esto en un gráfico lineal, de todos modos representaría $ 100 y $ 120 el primero de cada mes, pero al día 15 del mes parece que hubiera tenido ingresos de $ 110. Lo que no es cierto. Las barras se usan para unidades discretas de medida, mientras que las líneas se usan cuando se trata de un valor continuo, como la temperatura.

Figure 14. Un cuadro de barras simple: útil para representar cantidad discreta de información

Podemos ver que a las 8:00 la temperatura era de 20°C y a las 9:00, 22°C. Si miramos la curva para adivinar la temperatura a las 8:30 diríamos 21°C, lo que es un estimado correcto dado que la temperatura es continua y cada punto no es la suma de otros valores; representa el valor exacto en el momento o un estimado entre dos mediciones exactas.

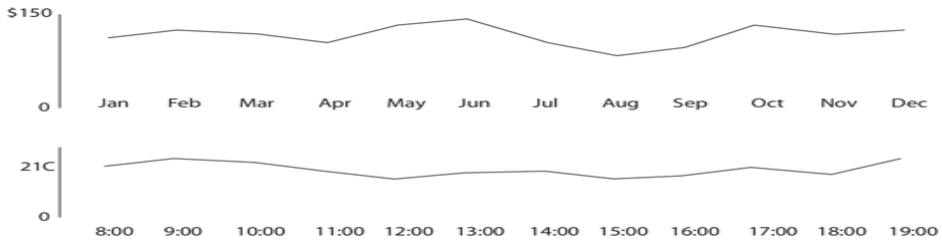


Figure 15. Gráficos de curva simples: útiles para representar información continua

Tanto el gráfico de barras como el de curvas tienen una variante de gráfico apilado (Figure 17). Esta es una excelente herramienta para narrar historias que puede funcionar de distintos modos. Pensemos, por ejemplo, en una compañía que tiene tres tiendas.

Para cada mes tenemos 3 barras, una por cada tienda, 36 en total para el año. Cuando las colocamos una junta a la otra (Figure 16) podemos ver rápidamente qué tienda ganó más en cada mes. Esta es una historia interesante y válida, pero hay otra oculta en los mismos datos. Si apilamos las barras, de modo que haya una sola por cada mes, ahora perdemos la posibilidad de ver fácilmente cuál tienda gana más, pero podemos ver en qué meses la compañía tiene mejores resultados de conjunto.

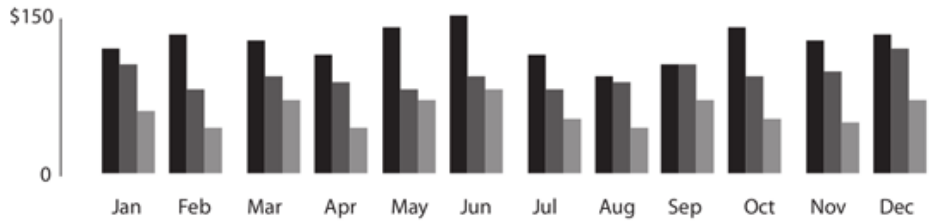


Figure 16. Un gráfico de barras agrupadas

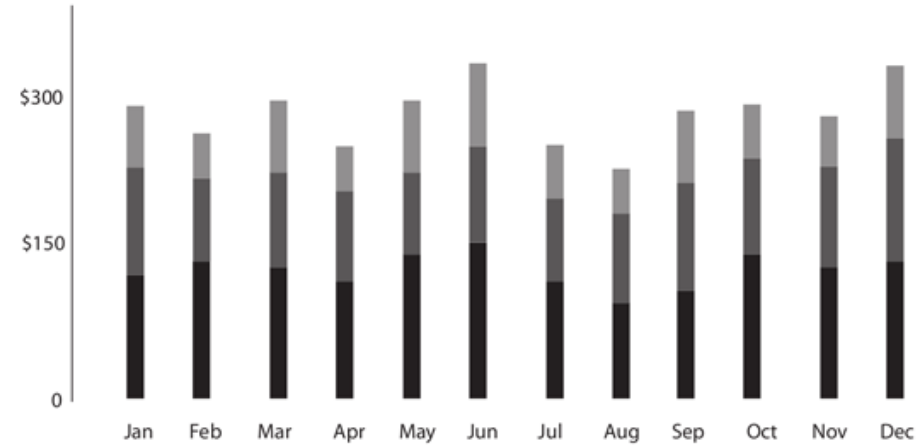


Figure 17. Un gráfico de barras apiladas

Ambas son representaciones válidas de la misma información, pero presentan dos historias diferentes usando los mismos datos. Como periodista, el aspecto más importante de trabajar con datos es que primero debe escoger qué historia quiere contar. ¿Se trata de cuál es el mejor mes en cuanto a ingresos totales o cuál tienda es la nave insignia? Este es solo un ejemplo simple, pero muestra cuál es el centro del periodismo de datos: Hacer la pregunta indicada antes de avanzar demasiado. La historia es la que guía la elección de la visualización.

Los gráficos de barras y curvas son en realidad lo básico del periodismo de datos. De allí se puede expandir a los histogramas, diagramas de área, "sparklines", gráficos de flujo y otros, que tienen propiedades similares y son adecuados para situación con ligeras diferencias,

incluyendo la cantidad de datos o fuentes de datos y la ubicación del gráfico en términos del texto.

En periodismo uno de los recursos gráficos más comúnmente utilizados son los mapas. En ellos hay tiempo, cantidades y geografía. Siempre queremos saber cuánto hay en un área comparada con otra área y cómo fluyen los datos de un área a otra. Los diagramas de flujo y los mapas coropléticos son herramientas muy útiles cuando se trata de visualizaciones para periodismo. Es clave saber cómo codificar un mapa con colores sin dar una representación equivocada o confundir a los lectores. Los mapas políticos por lo general tienen un código de color que indica todo a nada para determinadas regiones, aún si un candidato ganó en una parte del país por 1%. El color no tiene por qué reducirse a una opción binaria; se puede usar con cuidado gradientes de color basados en grupos. Entender los mapas es una parte importante del periodismo. Contestan fácilmente una de las cinco preguntas claves: ¿Dónde?

Una vez dominados los tipos básicos de cuadros y gráficos, se pueden comenzar a crear visualizaciones de datos más sofisticadas. Si no entiende lo básico, entonces está parado sobre terreno poco firme. De la misma manera que aprende a ser buen escritor –hacer frases cortas, tener presente el público y no complicar exageradamente las cosas para hacerse sonar inteligente, sino más bien transmitir el significado al lector- también debe aprender a mesurarse con los datos. Comenzar por algo pequeño es la manera más efectiva de narrar la historia, incrementando lentamente en la medida de lo necesario.

La escritura vigorosa es concisa. Una frase no debe contener palabras innecesarias, el párrafo no debe contener frases innecesarias, por el mismo motivo que un dibujo no debe tener líneas innecesarias y una máquina no debe tener partes innecesarias. Esto requiere no que el escritor haga que todas sus frases sean cortas o que evite dar detalles y que solo de un bosquejo de sus personajes, sino que toda palabra sea dicente.
Elements of Style (1918) — William Strunk Jr.
Está bien no usar todos los datos que tiene en su historia. No debiera tener que pedir permiso para ser conciso, esa debe ser la norma.

— Brian Suda, (optional.is)

Selección de herramientas "Hágalo Ud. mismo" para hacer sus propias visualizaciones de datos.

¿Qué herramientas de visualizaciones de datos se consiguen en la red en forma gratuita? Aquí en el Datablog y Datastore tratamos de hacer lo más posible usando las poderosas opciones gratuitas de internet.

Eso puede sonar un poco falso, dado que obviamente tenemos acceso a los increíbles equipos de gráficos e interactivos de The Guardian para las piezas en las que contamos con un poco más de tiempo, tales como este mapa de gasto público, creado utilizando Adobe Illustrator) o este interactivo de disturbios de Twitter.

Pero para nuestro trabajo cotidiano, a menudo usamos herramientas a las que cualquiera tiene acceso y creamos gráficos que cualquiera puede hacer.

¿Entonces, qué usamos?

Google Fusion Tables

Esta base de datos y herramienta de mapeado online se ha vuelto nuestra primera elección para producir mapas rápidos y detallados, especialmente aquellos que requieren zoom. Se tiene la alta resolución de Google Maps, pero puede abrir muchos datos, por ejemplo, 100 MB de CSV. La primera vez que uno lo intenta las Fusion Tables pueden parecer un poco complicadas, pero no se rinda. Lo utilizamos para producir mapas como el de Irak en la Figure 18 y también mapas de fronteras como la Figure 19 sobre los sin techo.

Figure 18. Los registros de guerra de WikiLeaks (The Guardian)

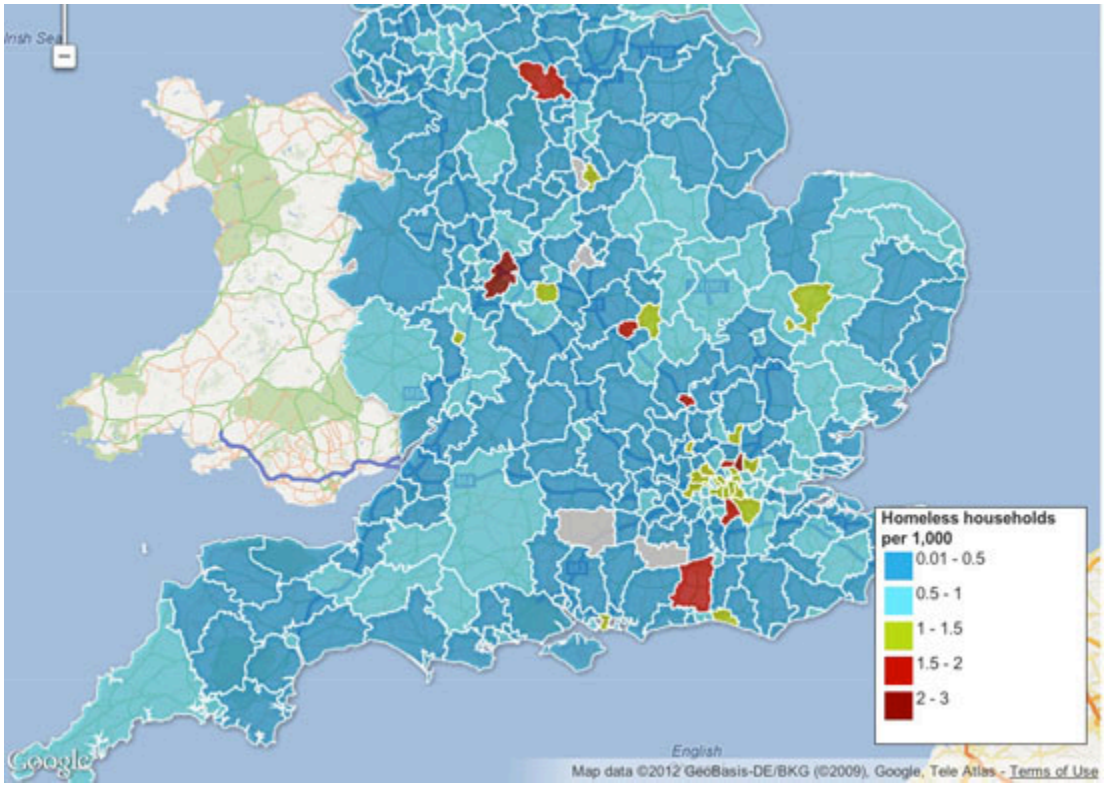


Figure 19. Mapa interactivo de personas sin hogar (The Guardian)

La principal ventaja es la flexibilidad –puede subir un archivo KML de fronteras regionales, por ejemplo- y luego fusionar eso con una tabla de datos. Además va a tener una nueva interfaz de usuario, lo que debe facilitar su uso.

No se necesita ser programador para hacerlo y esta herramienta de fusión de capas le permite unir distintos mapas o crear opciones de búsqueda o filtrado, que luego puede incorporar en un blog o sitio.

Este excelente tutorial de Kathryn Hurley de Google es un gran recurso para comenzar.

Note Use shpescape para convertir archivos .shp oficiales en Google *Fusion Tables*. También esté atento a que los mapas no sean demasiado complicados porque el programa no puede manejar más de un millón de puntos por celda.

Tableau Public

Si no necesita el espacio ilimitado de la edición profesional, Tableau Public es gratuito. Con este servicio visualizaciones bastante complejas de hasta 100.000 filas de modo simple y fácil. Lo utilizamos cuando tenemos que unir distintos tipos de cuadros, como en este mapa de tasas impositivas máximas en todo el mundo, que también tiene un cuadro de barras).

O incluso puede usarlo como explorador de datos, que es lo que hicimos en la Figure 20 con los datos de gastos en las elecciones federales de EE.UU, si bien nos quedamos cortos de espacio en la versión gratuita... algo a tener en cuenta). Tableau también necesita que los

datos estén formateados de modos bastante específicos para poder aprovecharlo al máximo. Pero si logra manejar eso tiene algo intuitivo que funciona bien. Por ejemplo, La Nación en la Argentina ha construido toda su operación de periodismo de datos en torno a Tableau.

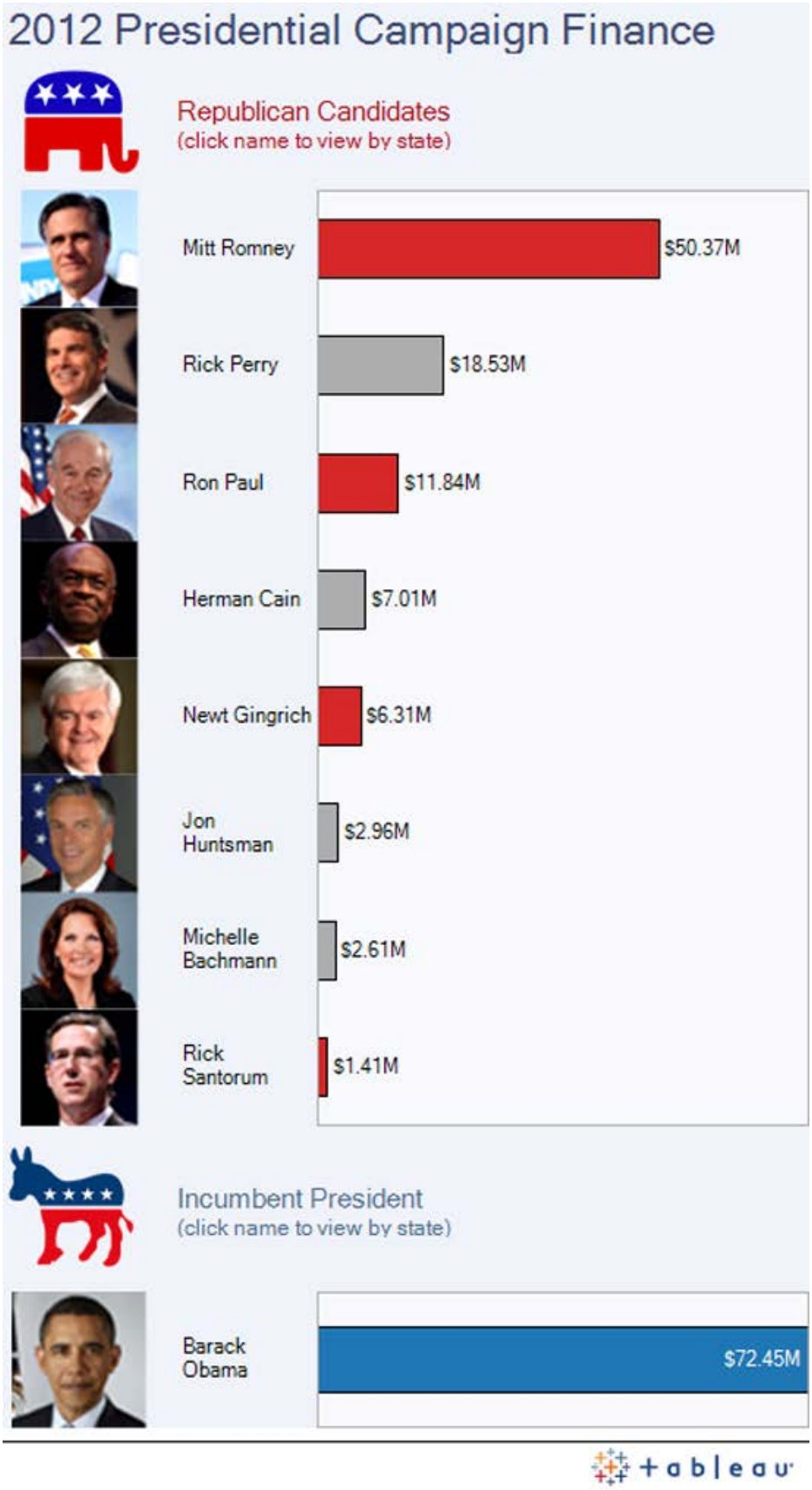


Figure 20. Finanzas de la campaña presidencial 2012 (The Guardian)

Tableau tiene algunos tutoriales online buenos con los cuales puede comenzar.

Note Tableau es para PC aunque se está preparando una versión para Mac. Use un "mirror" tal como "parallels" para hacerlo funcionar. (N. del T.: una aplicación de MAC para poder usar programas de Windows).

Gráficos con Google Spreadsheets

Puede acceder a esta herramienta en Google Spreadsheets

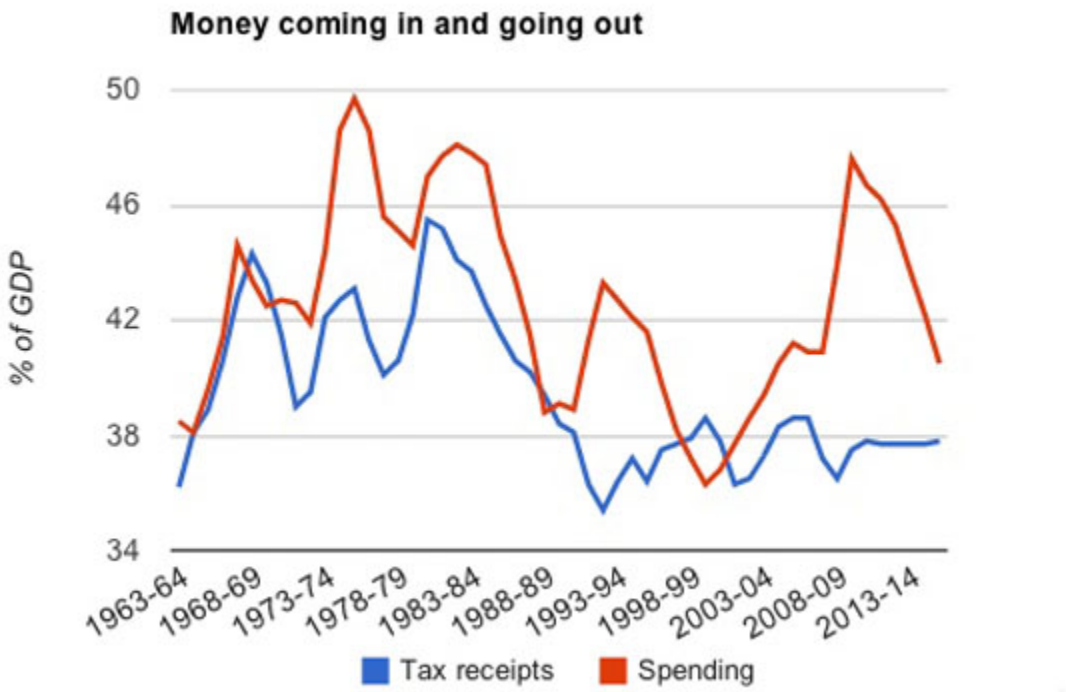


Figure 21. Gasto público e impuestos Reino Unido (The Guardian)

Luego de algo simple (como un gráfico de barras o curvas, o un gráfico de torta), encontrará que las Google Spreadsheets (que se crean con los documentos de su cuenta Google) pueden generar algunos gráficos bastante buenos, incluyendo las burbujas animadas usadas por el Gapminder de Hans Rosling. A diferencia de los gráficos API no necesita preocuparse por el código; es bastante similar a hacer un gráfico en Excel, en el sentido de que uno selecciona los datos y hace clic en el *widget* de gráficos. También vale la pena explorar las opciones de personalización; se puede cambiar el color, los encabezados y las escalas. Son bastante neutrales respecto del diseño, lo que es útil en gráficos pequeños. Los gráficos de curvas también tienen algunas opciones lindas, incluyendo opciones para anotaciones.

Note: Dedique algo de tiempo a las opciones de personalización de los gráficos; puede crear su propia paleta de colores.

Datamarket

Más conocido como proveedor de datos, Datamarket es en realidad una herramienta práctica para visualizar cifras. Puede subir sus propios datos o usar algunos de los muchos conjuntos de datos que ofrecen, pero las opciones son mejores si paga por una cuenta Pro.

Note Datamarket funciona de la mejor manera con datos de series temporales, pero no deje de ver su extensa variedad de datos.

Many Eyes

Si hay un sitio que está necesitado de un poco de atención y cuidado es Many Eyes de IBM. Cuando se presentó, creado por Fernanda B. Viégas y Martín Wattenberg, fue un ejercicio único en cuanto a permitir a la gente subir conjuntos de datos de modo simple y visualizarlos. Ahora, con sus creadores trabajando para Google, el sitio parece un poco desatendido, con sus paletas de colores apagados; hace tiempo que no ofrece nada nuevo en materia de visualizaciones.

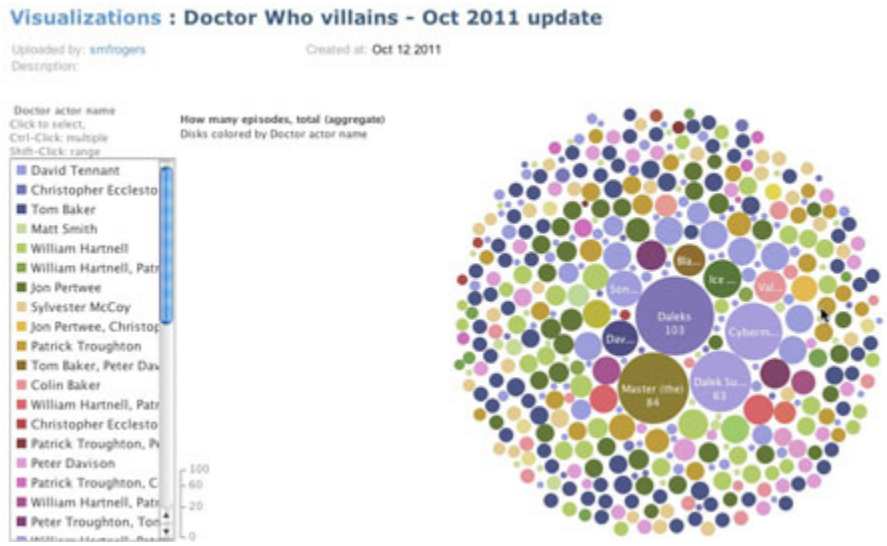


Figure 22. Villanos de Doctor Who; The Guardian

Note No se puede editar los datos una vez subidos, de modo que asegúrese de que estén bien antes de crear la visualización.

Color Brewer

No es estrictamente una herramienta de visualización, Color Brewer sirve en realidad para elegir colores de mapas. Puede escoger su color de base y obtener los códigos para toda la paleta.

Y algunos más

Si ninguno de estos le sirve, vale la pena ver lo que hay en DailyTekk que tiene aún más opciones. Las señaladas no son las únicas herramientas, solo aquellas que usamos con mayor frecuencia. Hay muchas más por allí, incluyendo:

- Chartsbin, una herramienta para crear mapamundis en los que se puede hacer clic.
- iCharts, que se especializa en pequeños "widgets" de gráficos
- Geoccomons que ofrece datos y datos de fronteras para crear mapas globales y locales.
- Y también está pikctochart.com que ofrece plantillas para esas visualizaciones de texto/cifras que son populares.

— Simon Rogers, the Guardian

Cómo presentamos los datos en el Verdens Gang

El periodismo busca llevar nueva información al lector lo más rápido posible. La manera más rápida de hacerlo puede ser mediante un video, una foto, un texto, un gráfico, una tabla o una combinación de éstos. Respecto de las visualizaciones, el objetivo debiera ser el mismo: información rápida. Las nuevas herramientas de datos permiten a los periodistas encontrar historias que de otro modo no podrían descubrir, y presentarlas de nuevas maneras. Estos son unos cuantos ejemplos que muestran cómo presentamos los datos en el diario más leído de Noruega, Verdens Gang (VG).

Cifras

Esta historia se basa en datos de la Dirección de Estadísticas de Noruega, datos de contribuyentes, y del monopolio nacional de lotería. En este gráfico interactivo el lector podría encontrar distintos tipos de información de cada condado y municipalidad noruega. La tabla muestra el porcentaje de los ingresos que se usa para jugar. Se creó usando Access, Excel, MySql y Flash.

Redes

Analizamos las redes sociales para estudiar las relaciones entre 157 hijos e hijas de las personas más ricas de Noruega. Nuestro análisis mostró que los herederos de la gente más rica de Noruega también heredaron las redes de sus padres. En total había más de 26.000 conexiones, y los gráficos se terminaron a mano usando Photoshop. Usamos Access, Excel, Notepad, y la herramienta de análisis de redes sociales Ucinet.

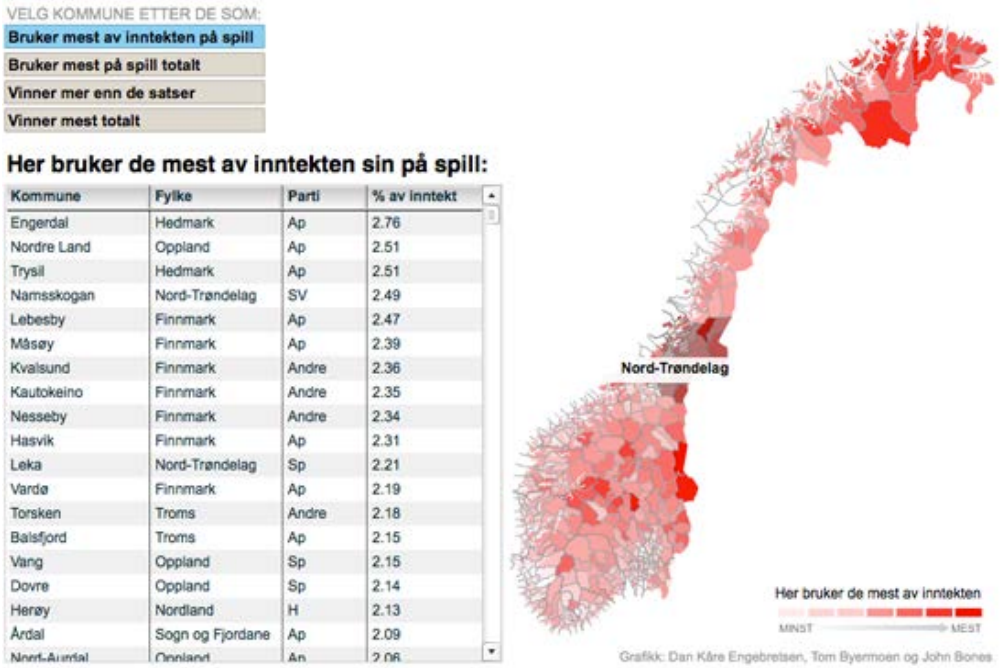


Figure 23. Mapeado de datos de contribuyentes y de la lotería (Verdens Gang)

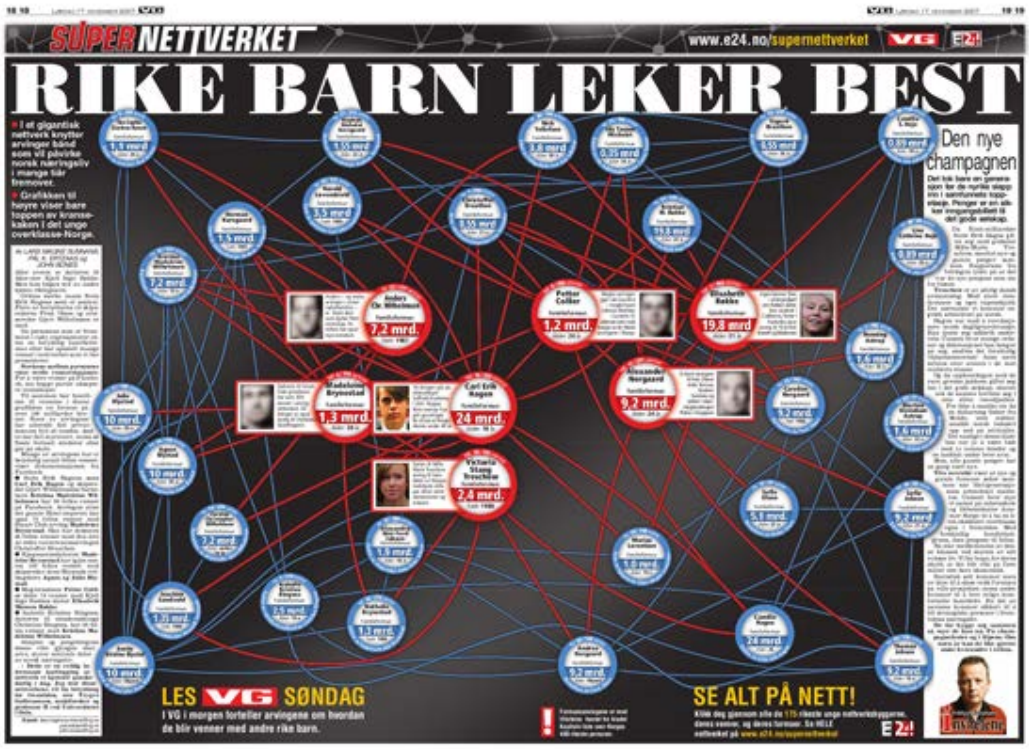


Figure 24. Los pájaros del mismo plumaje se unen (Verdens Gang)

Mapas

En este mapa de calor animado combinado con un gráfico de barras simple se puede ver la incidencia de crímenes en un mapa del centro de Oslo, hora por hora, a lo largo de los fines de semana por varios meses. En el mismo mapa de calor animado, se puede ver la cantidad de agentes de policía trabajando al mismo tiempo. En los momentos en que se dan los

crímenes, la cantidad de agentes de policía está en su punto más bajo.

Figure 25. Mapa de calor animado (Verdens Gang)

"Text Mining" (Minado de texto)

Para esta visualización, hicimos minería de texto de los discursos de siete líderes de partidos noruegos durante sus congresos. Todos los discursos fueron analizados y los análisis aportaron los argumentos de algunas historias. Cada historia se vinculó con el gráfico y los lectores pudieron explorar y estudiar el lenguaje utilizado por los políticos. Creamos esta visualización utilizando Excel, Access, Flash e Illustrator. Si ésto se hubiera hecho en 2012, hubiéramos creado el gráfico interactivo con JavaScript.

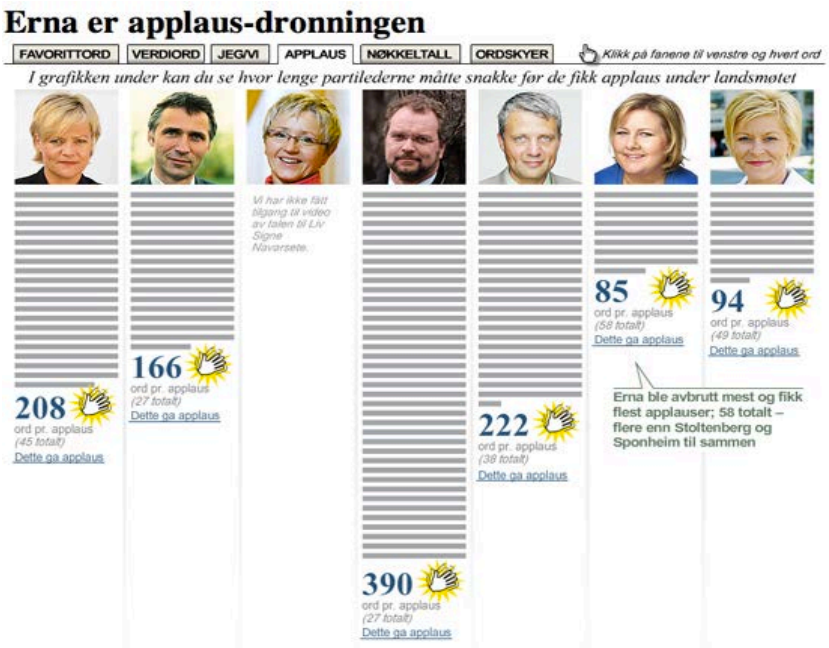


Figure 26. Minado de texto de discursos de líderes partidarios (Verdens Gang)

Notas finales

¿Cuándo necesitamos visualizar una historia? La mayoría de las veces no es necesario, pero a veces queremos hacerlo para ayudar a nuestros lectores. Las historias que contienen una gran cantidad de datos a menudo necesitan de una visualización. Pero tenemos que ser bastante críticos al elegir qué tipo de datos vamos a presentar. Conocemos todo tipo de cosas cuando informamos sobre algo, ¿pero qué necesita saber realmente el lector sobre la historia? Quizás baste una tabla, o un gráfico simple que muestra un proceso que va del año A al año C. Cuando se trabaja con periodismo de datos, el objetivo no es necesariamente presentar grandes cantidades de datos. Se trata de periodismo.

Ha habido una clara tendencia en los últimos dos o tres años a crear gráficos y tablas interactivas que permiten al lector investigar distintos temas. Una buena visualización es como una buena imagen. Se entiende de qué se trata con solo mirar uno o dos instantes. Cuanto más se mira la visualización, más se ve. La visualización es mala cuando el lector no sabe por dónde empezar o donde termina, y cuando la visualización está sobrecargada de detalles. En este caso, quizás una pieza de texto sería mejor.

— *John Bones, Verdens Gang*

Los datos públicos se vuelven sociales

Los datos son valiosos. El acceso a los datos tiene el potencial de clarificar cuestiones de un modo que genere resultados. Pero el mal manejo de los datos puede ubicar los hechos en una estructura opaca que no comunica nada. Si no promueven la discusión o aportan una comprensión en contexto, los datos pueden ser de limitado valor para el público.

Nigeria volvió a la democracia en 1999 luego de largos años de gobierno militar. Analizar los hechos detrás de los datos se consideraba una afrenta a la autoridad y como un intento de cuestionar la manchada reputación de la junta. La Ley de Secreto Oficial obligaba a los empleados públicos a no difundir información oficial. Aún pasados trece años del regreso a la democracia, el acceso a los datos públicos puede ser una tarea difícil. Los datos sobre el gasto público comunican poco a la mayoría del público que no conoce demasiado la contabilidad financiera y la aritmética compleja.

Al imponerse el uso de dispositivos móviles y con un creciente número de nigerianos online, junto con BudgIT vimos una gran oportunidad de usar tecnologías de visualización de datos para explicar y hacer que la gente se interesara por el gasto público. Para hacer esto, tuvimos que dirigirnos a usuarios de todo tipo de plataformas y llegar a los ciudadanos vía organizaciones no gubernamentales. Este proyecto apunta a convertir los datos públicos en objeto social y crear una red extensa que exija cambios.

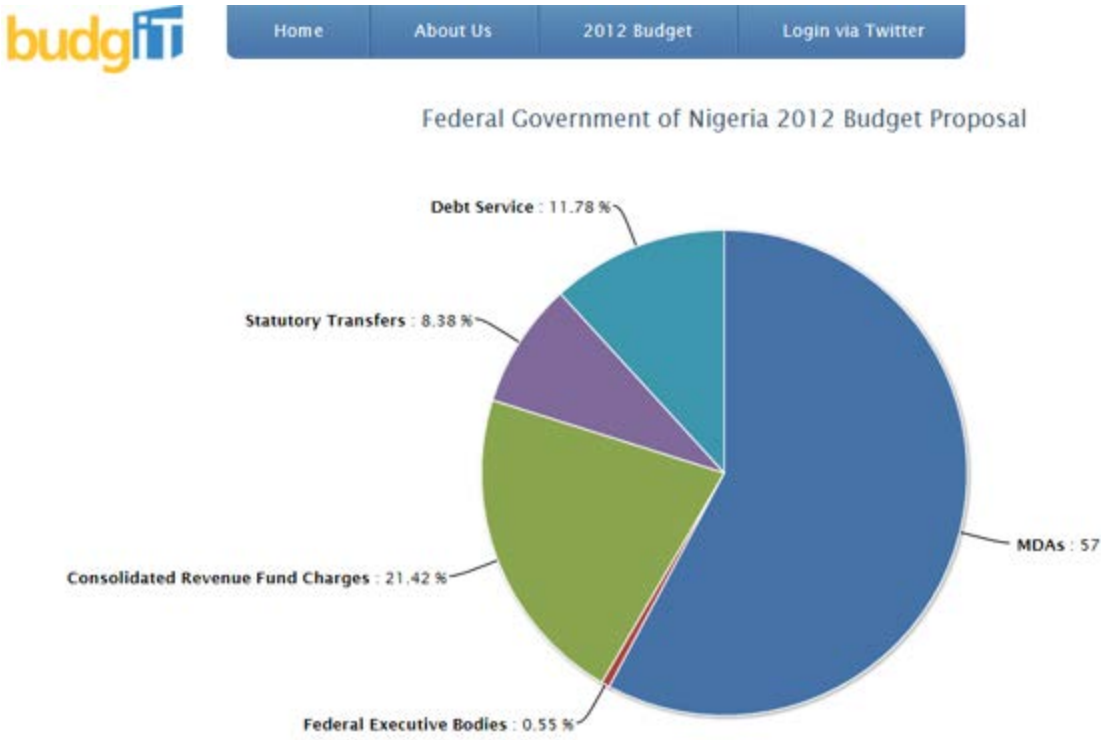


Figure 27. La aplicación de BudgIT (BudgIT Nigeria)

Para entablar exitosamente la relación con los usuarios, tenemos que entender lo que quieren. ¿Qué le importa al ciudadano nigeriano? ¿Dónde sienten que hay falta de información? ¿Cómo podemos hacer que los datos sean relevantes para sus vidas? El blanco inmediato de BudgIT es el nigeriano alfabetizado promedio conectado a foros online y medios sociales. Para competir por la limitada atención de los usuarios inmersos en una amplia variedad e intereses (juegos, lectura, socialización) tenemos que presentar los datos de modo breve y conciso. Luego de difundir una imagen de los datos como un tuit o una infografía, existe la oportunidad de una relación más sostenida con una experiencia más interactiva para dar a los usuarios una visión más amplia.

Al visualizar datos es importante comprender el nivel de manejo de datos que tienen nuestros usuarios. Por hermosos y sofisticados que puedan ser, los diagramas complejos y las aplicaciones interactivas pueden no comunicar de un modo significativo a nuestros usuarios en base a sus anteriores experiencias con la interpretación de datos. Una buena visualización habla al usuario en un lenguaje que puede entender, y presentará una historia con la que puede relacionarse fácilmente.

Hemos consultado a más de 10.000 nigerianos respecto del presupuesto, y los dividimos en tres categorías de acuerdo a su perfil para asegurar el valor óptimo. Explicamos brevemente las categorías a continuación:

Usuarios ocasionales

Son usuarios que quieren información de modo simple y rápido. Les interesa tener una idea de los datos, no un análisis detallado. Podemos dirigirnos a ellos vía tweet o gráficos interactivos.

Usuarios activos

Usuarios que estimulan el debate y usan los datos para incrementar su conocimiento de un área determinada o cuestionan los supuestos de los datos. A estos usuarios, queremos proveerles mecanismos de retroalimentación y la posibilidad de compartir su visión con sus pares vía las redes sociales.

Acaparadores de datos

Estos usuarios quieren datos en crudo para hacer visualizaciones o análisis. Simplemente les damos los datos para sus propósitos.

Con BudgIT nuestra relación con los usuarios se basa en lo siguiente:

Estimular debates en torno a tendencias actuales

BudgIT sigue debates online y offline y busca proveer datos sobre estos tópicos. Por ejemplo, con las huelgas del combustible en enero de 2012, hubo constante agitación entre los manifestantes respecto de la necesidad de que volviera a haber subsidios al combustible y reducir los gastos públicos extravagantes e innecesarios. BudgIT siguió el debate vía los medios sociales y en 36 horas con mucho esfuerzo creó una aplicación que permite a los ciudadanos reorganizar el presupuesto nigeriano.

Buenos mecanismos de retroalimentación

Nos relacionamos con los usuarios a través de canales de debate y medios sociales. Muchos usuarios quieren conocer las historias detrás de los datos y muchos nos piden nuestra opinión. Nos aseguramos de que nuestras respuestas solo expliquen los hechos detrás de los datos y no se vean afectadas por nuestros puntos de vista personales o políticos. Tenemos que mantener abiertos canales de retroalimentación, responder activamente a comentarios y relacionarnos con los usuarios de modo creativo para asegurar que se mantenga la comunidad creada en torno a los datos.

Hacerlo local

> En el caso de un conjunto de datos que apunta a un grupo en particular, BudgIT busca localizar su contenido y promover un canal de debate que se relacione con las necesidades e intereses de grupos particulares de usuarios. En particular, nos interesa relacionarnos con usuarios en torno a cuestiones que les preocupan vía SMS.

Luego de poner los datos sobre el gasto público en yourbudgit.com, buscamos tomar contacto con los ciudadanos a través de varias ONG. También pensamos desarrollar un marco de participación en el que ciudadanos e instituciones oficiales puedan realizar asambleas públicas para definir ítems claves del presupuesto que deben ser priorizados.

El proyecto ha sido cubierto por medios locales y extranjeros, desde CP-Africa hasta la BBC. Hemos emprendido un estudio de los presupuestos entre 2002 y 2011 para el sector de seguridad para un periodista de la AP, Yinka Ibukun. La mayoría de las organizaciones de medios son "acaparadores de datos" y nos han pedido datos para usar en sus informes. Estamos planeando nuevas colaboraciones con periodistas y organizaciones noticiosas en los meses venideros.

— Oluseun Onigbinde, BudgIT Nigeria

Interactuar con la audiencia en torno a sus datos

Casi tan importante como publicar los datos es lograr una reacción de su audiencia. Usted es humano; va a cometer errores, se le van a pasar algunas cosas y va a tener una idea equivocada de tanto en tanto. Su audiencia es uno de los activos más útiles que tiene. La gente puede verificar y señalar rápidamente cosas que usted quizás no consideró.

Pero relacionarse con esa audiencia tiene sus complicaciones. Está tratando con un grupo de gente acostumbrada, debido a años de uso de internet, a saltar de sitio en sitio, dejando nada más que un comentario sarcástico. Crear cierto nivel de confianza entre usted y sus usuarios es crucial; tienen que saber lo que van a recibir, cómo pueden reaccionar y comentar, y que lo que ellos aporten va a ser tomado en cuenta.

Pero primero tiene que pensar en la audiencia que tiene o quiere tener. Que se informará y será informada por el tipo de datos con los que usted trabaja. Si es específica de un sector particular, entonces va a querer explorar comunicaciones particulares con ese sector. ¿Hay organizaciones con las que puede tomar contacto y que podrían estar dispuestas a publicitar los recursos que tiene y el trabajo que ha hecho a una audiencia mayor? ¿Hay un sitio o foro comunitario con el que puede tomar contacto? ¿Hay publicaciones especializadas que podrían querer informar de algunas de las historias que usted está encontrando en los datos?

Los medios sociales son otra herramienta importante, aunque también depende del tipo de datos con los que trabaje. Si, por ejemplo, analiza estadísticas globales de embarques, es improbable que encuentre en Facebook o Twitter un grupo especialmente interesado en su trabajo. En cambio si está viendo índices de corrupción de todo el mundo, o estadísticas locales de criminalidad, eso es probable que interese a una audiencia mayor.

En Twitter el mejor enfoque tiende a ser contactar figuras de alto perfil, explicando brevemente por qué es importante su trabajo e incluir un *link*. Con suerte ellos lo retuitearán a sus lectores. Es una gran manera de maximizar la exposición a su trabajo con mínimo esfuerzo, pero no sea cargoso.

Una vez que tenga gente en su página, tiene que pensar en cómo va a interactuar su audiencia con su trabajo. Sin duda podrán leer la historia que escribió y mirar los gráficos o mapas, pero dar a sus usuarios una vía para responder es inmensamente valioso. Lo más importante es que probablemente eso le de a usted una mayor visión del tema sobre el que escribe, que puede incorporar a trabajos futuros sobre el tema.

Primero, no hace falta decir que tiene que publicar los datos en crudo junto a los artículos. Presente los datos en texto separado por comas en su sitio, o ubíquelo en el sitio de un servicio como Google Docs. De ese modo sólo habrá una versión de los datos y puede actualizarlos si encuentra errores que deba corregir más tarde. Mejor aún, haga ambas cosas. Facilite lo más posible a la gente obtener sus materiales en crudo.

Entonces comience a pensar si hay otras maneras en las que pueda conseguir que la audiencia interactúe. Siga las métricas para ver qué partes de sus conjuntos de datos reciben la mayor atención: es probable que las áreas más visitadas digan algo que usted no vio. Por ejemplo quizás no le prestó mucha atención a las estadísticas de pobreza en Islandia, pero si esas celdas están recibiendo mucha atención, entonces quizás haya allí algo que vale la pena analizar.

También piense más allá de los comentarios. ¿Puede vincular comentarios con celdas particulares en una hoja de cálculo? ¿O una región particular en un gráfico? Si bien la mayoría de los sistemas de edición no permiten este tipo de incrustación, vale la pena estudiar la posibilidad si está creando algo más particular. No se debe subestimar los beneficios que puede obtener de ello.

Asegúrese de que otros usuarios puedan ver los comentarios también, tienen casi tanto valor como los datos originales en muchos casos y si usted se guarda la información, entonces está privando al público de ese valor.

Finalmente otra gente podría querer producir sus propios gráficos informativos e historias basados en los mismos datos de origen; piense cuál es la mejor manera de vincularlos y

hacer un perfil de su trabajo. Podría usar un *hashtag* específico para ese conjunto de datos o, si se basa mucho en imágenes, podría compartirlo en un grupo de Flickr.

También podría ser útil contar con una ruta para compartir información de modo más confidencial; en algunos casos puede no ser seguro para la gente compartir públicamente sus aportes a un conjunto de datos o quizás simplemente no se sienta cómoda haciéndolo. Esa gente puede preferir enviar información a través de una dirección de correo electrónico o incluso una opción de comentarios anónimos.

Lo más importante que puede hacer con sus datos es compartirlos lo más ampliamente y del modo más abierto que sea posible. Permitir a sus lectores verificar su trabajo, encontrar errores y señalar cosas que se le pueden haber escapado harán que su periodismo y la experiencia para su lector sean infinitamente mejor.

— *Duncan Geere, Wired.co.uk*