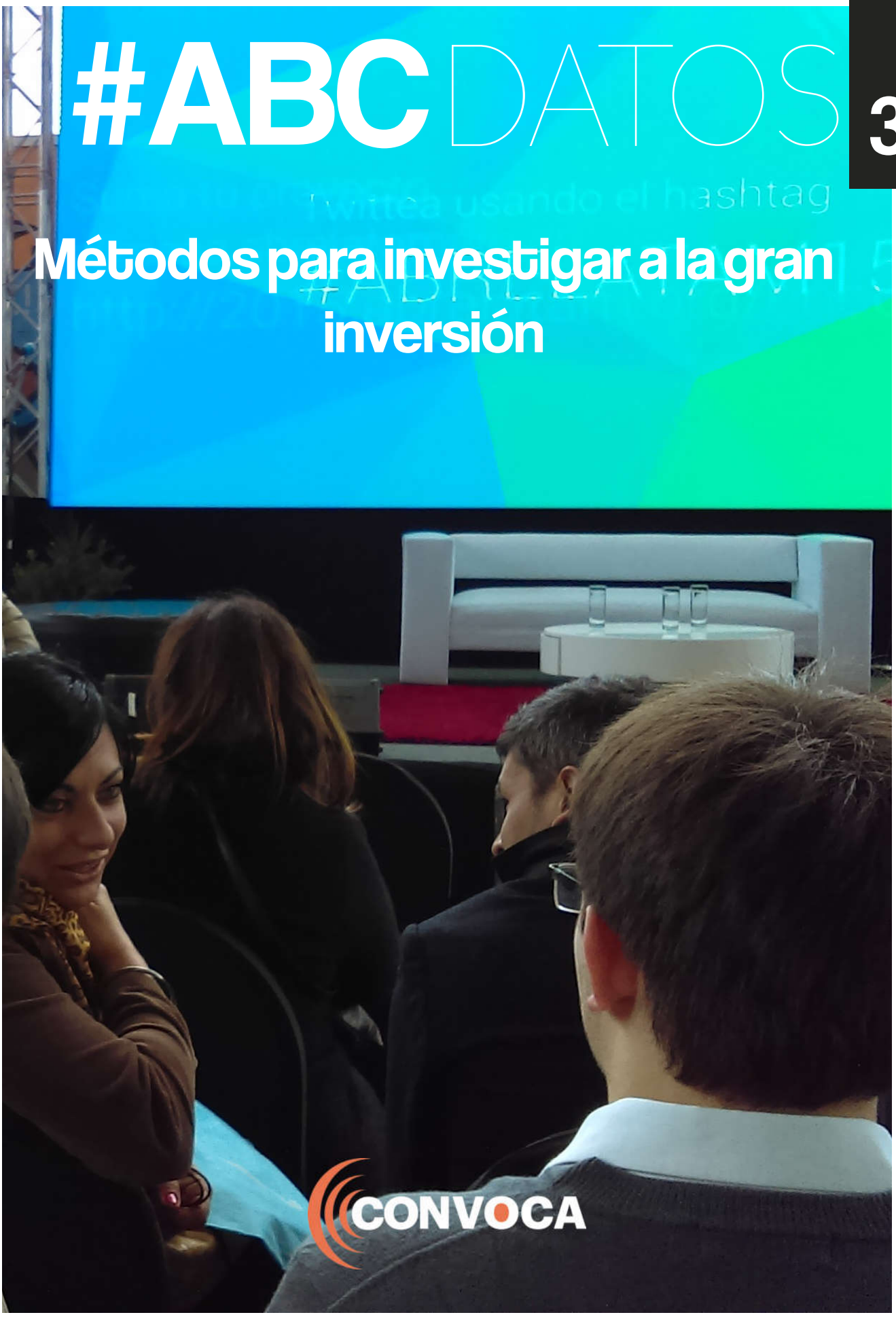
#ABC DATOS

3

Métodos para investigar a la gran inversión



Directora
Milagros Salazar

Editor
Carlos Bracamonte

Asesora editorial
Giannina Segnini.

Redactores
Aramís Castro
Gabriela Flores
Milton López

Diseño y diagramación
Stefany Aquise
Wendy Vega

Convoca es una plataforma de periodismo de investigación y análisis de datos que publica reportajes, realiza diversas actividades y elabora productos para promover las buenas practicas en el periodismo de investigación.

Índice

Presentación

Por: Milagros Salazar

Los datos pueden mentir más que las personas. Pero luego que ese mar de datos logra ser analizado, veri ficado y confrontado con la realidad, puede revelar un patrón de conducta, un sistema fallido y conexiones insospechadas que te permitirán trascender la filtra ción para conocer historias completas y reales. H a y muchas bases de datos disponibles en Internet, el desafío es saber como usarlas sin sacrificar la rigurosidad.

Después de dos ediciones anteriores sobre cómo in vestigar el poder y trabajar con datos cuando no es tán disponibles, este tercer número aparece bajo el nombre "Métodos para investigar a la gran inversión" que está dedicadoa conocer como un grupo de perio distas del Consorcio Internacional de Periodistas de Investigación (ICIJ por sus siglas en inglés), logró de - mostrar que 3.4 millones de personas fueron despla - zadas por proyectos financiados por el grupo de Banco Mundial y del Fondo Monetario Internacional (FM) en Lima.

El investigador principal de la investigación del ICIJ, Sasha Chavkin, cuenta lo que significó traducir los tér minos técnicos que figuraban en más de 6 mil 600 do cumentos oficiales del Grupo del Banco Mundial para recién entonces trasladar los datos a una hoja de cál culo con el propósito de ser analizados y confrontados con un persistente trabajo de campo de periodistas de más de 50 países que reportearon en Honduras, Gua temala, Perú, Kosovo, Etiopía, Sudán del Sur y Ghana. Mientras que Aramís Castro de Convoca narra cómo se realizó junto a otros periodistas e ingenieros am bientales, la serie investigativa "Excesos sin castigo" sobre la fiscalizacióna ambiental a la gran minería y la industria petrolera a partir de más de cien pedidos de información al Estado. La investigación logró determi nar quiénes eran las empresas que infringían las leyes ambientales de manera reiterada en los últimos años en Perú y cómo el Estado les aplicaba multas fijas sin importar la gravedad de la infracción.

En el último año el panorama se agravó en Perú con la aprobación de una norma que premió a estas empre - sas con la suspensión y re - corte de las multas impues - tas, a pesar que varias de ellas prefirieron entrampar los procesos sancionado - res en los tribunales de jus - ticia mediante estudios de abogados antes que pagar sus multas al Estado. Para la investigación, también se construyó un registro de da - tos inéditos que demostró que se encarpetaron más de mil informes de supervisión ambiental en el sector hidrocarburos y electricidad du - rante tres gobiernos.

> "Este boletín cuenta cómo se realizaron investi - gaciones a partir de los datos para potenciar el perio - dismo en su bús - queda de la ver - dad y la narración de historias".

El #ABCDatos incluye un resumen de los desafíos que puso en evidencia el panel de Datos Abiertos y Rendi - ción de Cuentas en la III Conferencia Regional de Da - tos Abiertos de América Latina y el Caribe (Condatos) que se realizó en setiembre en Santiago de Chile. Y cierra con un artículo notable de la periodista costa - rricense Giannina Segnini, pionera en la investigación periodística a partir de los datos, en el que explica paso a paso cómo lograr ser exactos y rigurosos en medio de un volumen abrumador de información en Internet a partir de la revisión de bases de datos como la del Banco Mundial que cualquier periodista puede pensar que son confiables pero que nos pueden sorprender. Esta iniciativa de Convoca es un espacio abierto para el aprendizaje y la colaboración de periodistas, desa rrolladores, analistas de datos, estudiantes y profesio nales de diiversas disciplinas interesados en potenciar la investigación periodística y la vigilancia pública a partir del uso responsable de los datos. La meta es publicar una edición mensual y unir puentes para un trabajo colaborativo. Gracias a quienes hicieron posi ble esta edición y bienvenidos a todos los interesados en sumarse a esta iniciativa que crecerá cada día.

- Lima, 6 de octubre de 2015.

Un método que atravesó la gran muralla del Banco Mundial

Convoca conversó con Sasha Chavkin, reportero principal de la investigación que reveló cómo el organismo internacional estaba implicado en el despojo de tierras de millones de personas. La serie ganó recientemente el premio de la Online News Association en la categoría periodismo de investigación innovador.

Por: Milton López

El 16 de abril último, más de 20 medios de comunicación a nivel mundial publicaron, en simultáneo, una investigación que involucraba al Banco Mundial y a su ente prestamista, la Corporación Financiera Internacional (IFC, por sus siglas en inglés), en el financiamiento de proyectos que desplazaron a miles de personas de sus tierras. Y a otras miles que fueron afectadas en su medio de vida.

La investigación había comenzado un año antes, cuando el periodista estadounidense Sasha Chavkin y el editor general Michael Hudson, junto a más de 50 periodistas de los cinco continentes, accedieron a la base de datos del Banco Mundial y contrastaron los primeros hallazgos con documentos, exfuncionarios y técnicos del organismo mundial, creando así un estándar de información, "porque entender los datos lo más pronto posible es lo más importante", afirma Chavkin.

Este "método consistente", como lo califica Chavkin, revelaría un patrón: el financiamiento de múltiples "proyectos de desarrollo" que quitaban la tierra a habitantes de diferentes partes del mundo. Luego de un año de trabajo publicaron "Atrapados por el desarrollo", una serie de reportajes que descubría cómo se financiaba "proyectos para el progreso" al margen de las personas que habitaban desde mucho antes en estas tierras, y en algunos casos, incluso sin que existiera un plan para reubicarlos.

> "Accedieron a bases de datos y contrastaron los primeros hallazgos con documentos, exfuncionarios y técnicos del Banco."

Chavkin es miembro del Consorcio Internacional de Periodistas (ICIJ), que reúne a un grupo importante de los periodistas de investigación más destacados del mundo. Él contó a #ABCDatos cómo se realizó de principio a fin esta investigación que confrontó la base datos del Banco Mundial con otras fuentes.

Todo comenzó cuando Sasha accedió a los reportes del Defensor del Pueblo del Banco Mundial, que supervisa la actividad de los funcionarios del banco, investiga denuncias y da recomendaciones. Chavkin observó que existían denuncias de comunidades desalojadas por proyectos de hidroeléctricas y monocultivos, entre otras actividades económicas financiadas por el Banco. "Existían docenas de casos así", dice Chavkin, y a esto se sumó reportes de la prensa local y de organismos no gubernamentales. Existían antecedentes sobre esas historias que "si bien llamaron la atención, nadie las había investigado por completo".

Así, desde julio de 2014, el equipo de reporteros bajo la coordinación de Chavkin descargó más de 6600 documentos oficiales del Banco Mundial sobre reasentamiento involuntario. "Sabíamos que el Banco Mundial tiene que hacer reportes, pero no cuenta la totalidad de casos", detalla Chavkin. Ese fue el principal desafío: trabajar con datos parcializados de este organismo y buscar fuentes dentro de la organización. Ante esta dificultad, el periodista contó en un hangout que organizó **Convoca** para América Latina el jueves 1 de octubre, que lograron romper la muralla del Banco como quien pela una cebolla: en la primera etapa se buscó a expertos de la sociedad civil, éstos a sus vez conectaron a los periodistas del ICIJ con ex funcionarios del Banco y ellos con técnicos que trabajaban dentro de la organización. En forma simultánea, la periodista Cécile Schillis-Gallego, se encargó de construir y analizar un registro completo de datos de los casos investigados.

Para lograr un trabajo colaborativo e intercambio de información entre periodistas de los países involucrados en el proyecto, ICIJ creó una plataforma de comunicación: Odyssey (como la nave espacial de la famosa serie Star Trek). Sasha la describe así: "es un foro que funciona como un espacio", además que los que integran el equipo "son fanáticos de Star Trek", dice.

Si bien los datos eran lo sustancioso de la búsqueda, no comunicaban nada sino existían historias que los humanizaran: testimonios de pobladores imprescindibles para Chavkin que "fueron los mismos que encontré al comienzo con los reportes del Defensor".
Así, como varios de los reporteros del caso, Sasha viajó a las zonas de donde provenían las denuncias por megaproyectos que desplazaban poblaciones sin

su consentimiento. Chavkin estuvo en Sudán del Sur, Etiopía y Honduras.

Este último país es el escenario de su reportaje 'Bañadas en sangre', que relata cómo la Corporación Dinant, con fondos de la IFC, invadió tierras de campesinos para la expansión de sus cultivos de palma aceitera. La incursión violenta de la trasnacional ha provocado varios asesinatos que han quedado impunes. Incluso Chavkin recibió intimidaciones del coronel encargado de "estabilizar" el conflicto: "si usted va a ese pueblo donde hay campesinos, no garantizamos su seguridad".

Al principio, el Banco Mundial negó las acusaciones, pero cuando en marzo de 2015, Sasha y el equipo les presentaron los primeros hallazgos y un cuestionario para que el organismo diera su versión, no pasaron más de cinco días para que el organismo diera una conferencia de prensa donde anunciaba que "era necesaria una reforma en el plan de reasentamiento".
Si bien los funcionarios del Banco Mundial conocían las irregularidades desde 2012, fue gracias a la publicación de los reportajes que decidieron reformar su sistema. La información procesada fue cuantiosa y densa, por eso fue necesario que "todos los periodistas tuvieran que compartir sus hallazgos, pese a que eso no es común en periodistas de investigación".

La colaboración fue lo principal; los plazos, inamovibles: la fecha de publicación final era el 16 de abril, sin prórroga. Ese día, la publicación se dio en paralelo en medios impresos, televisivos y digitales. Sasha resume en tres pasos generales la investigación: primero, se encontró un sistema roto; luego, se halló evidencia de un problema sistémico; por último, se analizaron las causas profundas. Hay que buscar casos individuales (como el de Honduras), que respaldado por datos oficiales y el contraste con fuentes, probarán un problema en el sistema. Regla básica de la prensa rigurosa. "Ese es el fin de todo periodista, el cambio estructural de un sistema fallido", explica Chavkin.

Explora más esta herramienta de los Desalojados y Abandonados AQUÍ

Cómo excavar en cientos de datos para investigar a las industrias extractivas

La investigación "Excesos sin castigo" de Convoca reveló la conducta ambiental de las industrias extractivas, que mueven la economía de Perú y a la vez son la principal fuente de conflictos sociales. Uno de los autores de la investigación relata la historia de estos reportajes que serán publicados en un libro electrónico el 7 de octubre.

Por: Aramís Castro

La descarga de una hoja de cálculo fue el primer paso hacia el extenso registro de las mineras y petroleras con más incumplimientos de las normas ambientales en los últimos años en Perú (de 2010 a 2014). La hoja de cálculo provenía de la página web del Organismo de Evaluación y Fiscalización Ambiental (OEFA) y, a primera vista, no parecía complicado: sólo había que analizar los datos y reportear; pero al hurgar más, la realidad era otra. La ardua ruta hacia el primer reportaje recién se iniciaba. La investigación completa nos tomaría seis meses.

El segundo paso fue acceder a los informes del OEFA (documentos públicos). Luego decidimos complementar los datos iniciales con otras hojas de cálculo (registros de supervisión, procesos llevados al Poder Judicial, concesiones de proyectos mineros, directorio de empresas mineras y de estudios de abogados, entre otros). Buscábamos cruzar información, y así obtener más datos para perfilar a las infractoras de las industrias extractivas. En ese proceso presentamos más de cien solicitudes de información pública que, una vez respondidas, fueron claves para crear nuevos registros de datos para determinar el nivel de reincidencia de las infracciones de las compañías y la responsabilidad de las autoridades peruanas para implementar una fiscalización eficaz. Por ejemplo, se logró establecer por punto de monitoreo de efluentes mineros (líquidos potencialmente contaminantes que salen de los campamentos mineros hacia ríos y suelos) los excesos en los que incurrían las empresas por cada mineral y su posible daño a los pobladores de las zonas aledañas.

Los resultados iniciales se gestaron gracias al uso de tablas dinámicas del Excel, que permite analizar en pocos minutos miles de filas y columnas llenas de información. Con ello se crearon los primeros rankings sobre las empresas más sancionadas por el OEFA que, en la mayoría de casos, eran las mismas incluso desde antes de 2010 en que Osinergmin, el antiguo órgano de fiscalización ambiental (Osinergmin), estuviera a cargo de la supervisión de las industrias extractivas. Se puso en evidencia que las multas no eran lo suficientemente disuasivas para que las empresas dejaran de incurrir en las mismas infracciones y decidieran mejorar sus procesos productivos para evitar el daño al medio ambiente y la salud pública.

En el primer reportaje, 'El círculo minero de la infrac-

Foto: Julio Angulo - La República

ción', se presentó a las mineras más reincidentes, los potenciales daños a las comunidades que conviven alrededor de sus actividades extractivas y los nexos comerciales entre estas compañías infractoras. Ya con una historia publicada, el camino era más claro.

En la segunda historia, 'La tarifa plana de la gran minería', el hallazgo principal fue mostrar cómo la falta más recurrente de las empresas era sobrepasar los límites legales para descargas efluentes que salían de las minas para expandirse en suelos y ríos (LMP); y no solo eso: la construcción de una hoja de cálculo con datos sobre los niveles de contaminación por mineral vertido permitió mostrar que no importó la magnitud del exceso porque la multa fue siempre la misma: 50 Unidades Impositivas Tributarias (UIT), poco más de 130 mil dólares.

Entre las fuentes a las que acudimos, las entrevistas con ingenieros ambientales, de minas, toxicólogos y diversos especialistas del sector fueron muy importantes para entender las resoluciones y explicar los principales hallazgos a los lectores. Al menos 60 profesionales fueron consultados para la investigación. En paralelo al reporteo y las entrevistas, se realizó una "limpieza de los datos" con el objetivo de presentar la información completa bajo una herramienta útil que le permitiera a los ciudadanos saber quién contaminaba más en su región. Programadores, ingenieros, infografistas, diseñadores, periodistas y estudiantes participaron en la construcción del primer Mapa de Infracciones Ambientales a nivel nacional que en su versión inicial mostró datos de las mineras para luego incluir al sector hidrocarburos.

En resumen, trabajamos con más de dos mil documentos, entre resoluciones y apelaciones, y un mapa con las coordenadas de las faltas ambientales.

> "Se realizó una limpieza y análisis de datos con el objetivo de presentar la información (...) que le permitiera a los ciudadanos saber quién contaminaba más en su región".

Durante 2014, el actual gobierno cambió la normatividad ambiental. El equipo de Convoca decidió investigar esas reformas bautizadas por sus opositores como el 'Paquetazo ambiental' (Ley 30230). Con la nueva norma se eliminaban las infracciones por tres años para incentivar la inversión de las industrias extractivas. Convoca encontró, a partir de un cálculo basado en los datos oficiales del OEFA, que el Estado dejó de cobrar entre S/. 20 millones (ma´s de US$7 millones) y S/. 30 millones (más de US$11 millones) por la 'amnistía' a las mineras. En el análisis también se detectó que las más infractoras fueron las más beneficiadas por la ley.

Convoca construyó un registro inédito de más de mil supervisiones en el sector hidrocarburos y electricidad que fueron archivadas porque prescribieron. La construcción de la hoja de cálculo se realizó luego de una serie de pedidos de información. La información encarpetada permitió descubrir cuáles eran las empresas beneficiadas y las autoridades responsables. A lo largo de los reportajes de 'Excesos sin castigo', los datos adquirieron relevancia por el conocimiento que se obtienen de éstos: patrones de conducta, nexos y un sistema fallido. Para que eso sea posible, el trabajo con la computadora no reemplazó al reporteo tradicional.

Este miércoles 7 de octubre, **Convoca** presentará en una conferencia un libro digital gratuito con los reportajes de la serie 'Excesos sin castigo'.

Cómo se analizaron los datos del reportaje: "Los S/. 30 millones que no cobró el gobierno en multas mineras"

1 Acceso

Se accedió al Registro de Actos Administrativos publicado por OEFA y a las resoluciones de sanción de primera y segunda instancia, aprobadas entre julio de 2014 y marzo de 2015. En forma simultánea, se hicieron varios pedidos de información a OEFA para completar la verificación y el análisis.

2 Construcción del registro

Los datos fueron organizados en dos hojas de cálculo según las instancias de evaluación de OEFA: DFSAI (97 casos) y Tribunal (66 casos). En cada hoja de cálculo se puso la información fundamental del proceso: nombre de la empresa, número de expediente, motivo de sanción y se añadió el monto mínimo y máximo de la multa, representado en Unidades Impositivas Tributarias (UIT), y su conversión a soles y dólares. Se usó el tipo de cambio y el valor de la UIT, según el año en el que se emitió la resolución.

3 Organización y limpieza de los datos

Se cruzaron los datos entre las dos hojas de cálculo para evitar duplicidad en los nombres de las empresas y los procesos sancionadores. Al hacer una única lista de empresas, sin contar dos veces a aquellas que se beneficiaron en el DFSAI y en el Tribunal (seis compañías), en total son 49 mineras favorecidas.

4 Análisis

Se usó tablas dinámicas en Excel para conocer a las empresas beneficiadas y los montos.

DFSAI (primera instancia)

Solo se consideró las 57 sanciones firmes (ratificadas por el Tribunal o consentidas por la empresa). Aquí, la ley permitió congelar por 3 años el cobro del monto total de la multa.

97 — 57

Tribunal (segunda instancia)

Entre julio de 2014 y marzo de 2015, el Tribunal evaluó 66 casos, de los cuales solo se consideraron los 17 beneficiados. En esta instancia, la ley permitió reducir la multa en 50%.

66 — 17

El cálculo de las multas congeladas y reducidas arrojó como rango mínimo S/ 20 millones 222 mil ($7 millones de dólares) y como rango máximo, alcanzó los S/30.9 millones ($10.9 millones). Se consideró lo que registraron técnicos de OEFA en las resoluciones, según la escala de multas.

5 Visualización

Se mostraron los resultados en Tableau Public, una herramienta interactiva para visualizar datos. La información de este análisis se presentó en dos gráficos: uno que expone a las empresas más beneficiadas del DFSAI y otro, a las compañías más favorecidas del Tribunal. De forma adicional, se mostró un análisis comparativo por año de lo que dejó de recibir OEFA por la vigencia del "Paquetazo ambiental" a partir de la construcción de un registro de procesos sancionadores de los últimos cuatro años*.

* Para establecer el año de la multa impuesta y su respectivo monto, se tomó en cuenta la fecha de resolución de la Dirección de Fiscalización, Sanción y Aplicación de Incentivos (DFSAI), que es la instancia que inicia el proceso sancionador.

La ruta de los datos abiertos en América Latina

¿Cómo trabajar con datos puede motivar la colaboración ciudadana? ¿Cuál es su rol en la transparencia del sector público y la información sobre las élites empresariales? Estas y otras preguntas se debatieron en la Tercera Conferencia Regional de Datos Abiertos de América Latina y el Caribe (Condatos), realizada en Santiago de Chile.

Por: Gabriela Flores

En el Perú aquellas empresas que quieren mo - verse en las grandes ligas financieras y cotizan sus acciones en la bolsa de valores tienen la obli - gación de entregar información periódica sobre sus activos, directorio, gerentes y estados finan - cieros. Esa información es de acceso público. Cualquier ciudadano curioso puede ingresar a la página web de la Superintendencia del Mercado de Valores (SMV) y acceder a los datos de gru - pos de poder que negocian con el Estado y que, no pocas veces, to man decisiones que afectan a las mayorías. Pero lo interesante no sólo está en saber quién es quién en la élite empresarial.

El dato por el dato no es suficiente. Lo interesante resulta de verificar y cruzar la información. Así lo hizo en México la organización Poder con iniciati - vas como Quién es Quién Wiki, una base de datos sobre empresas y empresarios, y RindeCuentas que, a partir de la base de datos, realiza indagaciones de casos emblemáticos. Hasta el momento, Poder ha logrado mostrar cómo los empresarios comparten el manejo de los principales grupos de poder y su cer - canía con el poder político.

> " El dato por el dato no es suficiente. Lo interesante resulta de verificar y cruzar la información " .

Para Eduard Martín – Borregón, coordinador del proyecto, el trabajo con datos debe servir para la rendición de cuentas de los grupos de poder.

El caso mexicano fue presentado la tarde del 9 de setiembre en una sala del Centro Cultural Ga - briela Mistral en Santiago de Chile, durante el pa - nel Datos Abiertos y Rendición de Cuentas que se desarrolló en el marco la III Conferencia Re - gional de Datos Abiertos de América Latina y el Caribe (Condatos). El debate intentó responder ¿quién, para qué y por qué deben abrir sus datos?

Foto: Gabriela Flores

Junto a la experiencia de Poder, Convoca pre - sentó la serie investigativa "Excesos sin Castigo" que reveló cómo las principales empresas de las industrias extractivas incumplen las normas am - bientales una y otra vez, evitan el pago de multas y se benefician con normatividad hecha a la me - dida. Los reportajes fueron el resultado de más de un centenar de pedidos de información, el acceso y análisis de más de dos mil documen - tos entre informes de supervisión ambiental que no eran públicos y resoluciones del Organismo de Evaluación y Fiscalización Ambiental (OEFA).

Apartirde este trabajo,**Convoca** construyó basesde datos, verificó y contrastó información que permitió los hallazgos periodísticos así como la publicación del Mapa de Infracciones Ambientales, herramien - ta que permite a los ciudadanos acceder directa - mente a las resoluciones analizadas por **Convoca.**

Estos casos muestran la utilidad y necesidad de los datos abiertos y la construcción de bases de

datos para la transparencia pública y privado. Los esfuerzos de la prensa independiente y de la sociedad civil son aún escasos; sin embargo, muchas iniciativas intentan replicarse en el resto de la región, aun cuando las leyes de transparencia pública por país tengan distintos alcances, y éstos no promuevan la rendición de cuentas del sector privado.

> "Transparencia Internacional presentó una iniciativa para que la información sobre los verdaderos dueños de las empresas no sea un secreto".

En la tercera conferencia Condatos se mostró que siempre hay maneras de obtener información que incomode a las élites empresariales. Al respecto, durante la reunión se presentó una iniciativa de Transparencia Internacional (TI) para que la información sobre los verdaderos dueños o "beneficial owners" de las empresas no sea un secreto. Fabiano Angélico, consultor independiente de TI, explicó que esta iniciativa busca combatir la corrupción y el desvío ilegal de fondos a través de testaferros y empresas de fachada.

Hay algo evidente: no solo el sector público debería trasparentar sus acciones. No obstante, lo que el poder privado no quiere publicar está siendo revelado por medios independientes y de la sociedad civil. Mientras no haya voluntad política firme, los esfuerzos por promover datos abiertos, elaborar bases de datos y despertar el interés ciudadano seguirán siendo imprescindibles en escenarios de alta desigualdad social como el latinoamericano.

Conferencia ConDatos en Santiago de Chile. Foto: Gabriela Flores - Convoca.pe

El ciclo virtuoso para verificar la calidad de los datos

En tiempos en que el periodista puede extraviarse en un universo infinito de datos, la destacada periodista de investigación, Giannina Segnini, muestra la ruta que puede seguir un reportero para verificar la calidad de los datos que obtiene. Segnini, profesora de la Universidad de Columbia, explica cada paso basada en más de dos décadas de trabajo.

Por Giannina Segnini*

Nunca antes los periodistas tuvieron tanto acceso a la información. Más de 3 exabytes de datos – equivalente a 750 millones de DVDs – son creados cada día, y ese número se duplica cada 40 meses. La producción global de datos es estimado hoy en día en yottabytes (un yottabite es equivalente a 250 trillones de DVDs de datos). Ya hay discusiones en marcha acerca de la nueva medición que se necesitará una vez que superemos el yottabyte.

El aumento en el volumen y la velocidad de la producción de datos puede ser abrumador para muchos periodistas, muchos de los cuales no están acostumbrados a usar grandes cantidades de datos para investigación o narración de historias. Pero la urgencia y el afán de hacer uso de los datos, y la tecnología disponible para procesarlos, no deberían distraernos de nuestra misión subyacente por la exactitud.

> "Los datos y la tecnología no deberían distraernos de nuestra misión por la exactitud."

Para capturar completamente el valor de los datos, nosotros debemos ser capaces de distinguir entre información cuestionable y de calidad, y ser capaces de encontrar historias reales en medio de todo el ruido.

Una lección importante que he aprendido de dos décadas usando datos para investigación es que los datos mienten – casi tanto como la gente, o incluso más. Los datos, después de todo, son creados y sustentados por la gente.

Los datos están destinados a ser una representación de la realidad de un momento específico de tiempo. Entonces, ¿cómo verificamos que un conjunto de datos corresponde a la realidad?

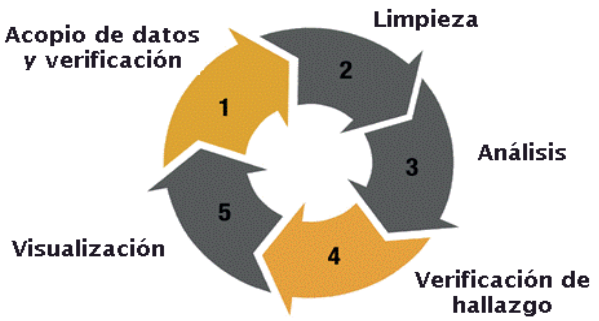
Dos tareas de verificación son claves durante una investigación basada en datos. Una evaluación inicial debe ocurrir inmediatamente después de obtener los datos; y los hallazgos deben ser verificados al final de a fase de investigación o análisis.

A. VERIFICACIÓN INICIAL

La primera regla es cuestionar todo y a todos. No hay tal cosa como una fuente totalmente fiable cuando se trata de usar datos para hacer periodismo meticuloso.

Por ejemplo, ¿confiarías totalmente en una base de datos publicada por el Banco Mundial? La mayoría de los periodistas a los que les hice esta pregunta dijeron que sí; ellos consideran al Banco Mundial como una fuente segura. Vamos a probar esa suposición con dos grupos de datos del Banco Mundial para mostrar cómo verificar datos, y para reforzar que incluso las llamadas fuentes confiables puedes proveernos datos equivocados. Seguiré el proceso señalado en el gráfico de abajo.

Fases de una investigación con datos



1.¿Están los datos completos?

Una primera práctica que recomiendo es explorar los valores extremos (altos o bajos) para cada variable en un grupo de datos, y luego contar que el número de registros (filas) se enumeren dentro de cada uno de los valores posibles. Por ejemplo, el Banco Mundial publica una base de datos con más de 10,000 evaluaciones independientes realizadas a más de 8,600 proyectos desarrollados alrededor del mundo por la organización desde 1964.

Sólo ordenando la columna del costo de préstamos en orden ascendente en una hoja de cálculo, pode-

Foto: Albert Marín - La Nación

mos rápidamente ver cómo múltiples registros tienen un cero en la columna de costo. Si creamos una tabla dinámica para contar cuántos proyectos tienen costo cero, en relación al total de registros, podemos ver cómo más de la mitad de ellos (53%) costaron cero (ver gráfico 1).

Gráfico 1

B19

	A	B
3	Count of Lending Project Cost	
4	Row Labels	Total
5	▼ 0	
6	0	5496
7	▶ 400000 to 8910000	4911
8	Grand Total	10407
9		

Esto significa que cualquiera que realiza un cálculo o análisis por país, región o año, que implica el costo de los proyectos, estaría equivocado si no pudieron dar cuenta de todas las entradas sin costo indicado. El conjunto de datos que se proporciona conducirá a una conclusión inexacta.

El Banco publica otra base de datos que supuestamente contiene los datos individuales para cada proyecto financiado (no solo evaluado) por la organización desde 1947 (ver gráfico 2).

Sólo con abrir el archivo api.csv en Excel (versión del 7 de diciembre, 2014), está claro que los datos están sucios y contienen muchas variables combinadas en una celda (como nombres de sectores o nombres de países). Pero incluso más notable es el hecho de que este archivo no contiene todos los proyectos financiados desde 1947.La base de datos de hecho sólo incluye 6,352 fuera de los más de 15,000 proyectos financiados por el Banco Mundial desde 1947. (Nota: el Banco eventualmente corrigió este error, para el 12 de febrero del 2015, el mismo archivo incluía 16, 215 registros.)

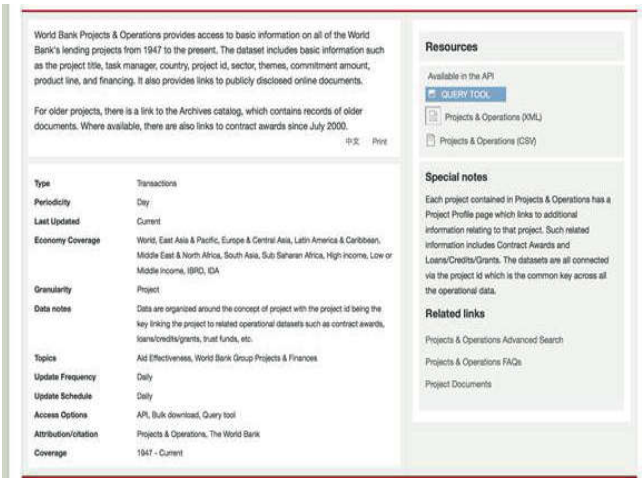
Después de poco tiempo de examinar los datos, vemos que el Banco Mundial no incluye el costo de todos los proyectos en su base de datos, publica datos sucios, y falló al incluir todos sus proyectos en al menos una versión de los datos. Debido a todo eso, ¿qué esperarías ahora sobre los datos publicados por instituciones aparentemente menos confiables?

Otro ejemplo reciente de inconsistencia de base de datos que descubrí de camino al taller que estaba dando en Puerto Rico para el que usamos las bases de datos de los contratos públicos de la Comptroller's Office. Algunos de los 72 contratos públicos, fuera de todos los contratos del último año, tienen valores negativos. ($-10,000,000) en sus campos de costo. Open Refine es una excelente herramienta para explorar rápidamente y evaluar la calidad de las bases de datos.

> "Recomiendo explorar los valores extremos (altos o bajos) para cada variable en un grupo de datos"

En la primera imagen debajo, puedes ver cómo Open Refine puede ser usado para ejecutar una "faceta" numérica en el campo "cuantía" (cantidad). Una faceta numérica agrupa números en acumuladores de distancia numéricos. Esto te permite seleccionar cualquier rango que abarque un número consecutivo de contenedor (ver gráfico 3).

Gráfico 2



La segunda imagen muestra que puedes generar un histograma con los rangos de valores incluidos en la base de datos. Los registros pueden entonces ser filtrados por valores moviendo las flechas dentro del gráfico. Lo mismo se puede hacer para fechas y valores de texto.

2. ¿Existen registros duplicados?

Un error común cuando se trabaja con datos es no identificar la existencia de registros duplicados.

Cuando sea que el procesamiento de datos desglosados o información sobre las personas, compañías, eventos o transacciones, el primer paso es buscar una variable de identificación única para cada elemento. En el caso de las base de datos de la evaluación de proyectos del Banco Mundial, cada proyecto es identificado a través de un código único o "Proyecto ID". Otras bases de datos de entidades pueden incluir un único número de identificación o, en el caso de contratos públicos, un número de contrato.

Si contamos cuántos registros hay en la base de datos para cada proyecto, vemos que algunos de ellos son duplicados hasta 3 veces. Por lo tanto, cualquier cálculo por país, región o fecha utilizando los datos, sin eliminar duplicados, sería un error (Ver gráfico 4).

En este caso, los registros se duplican porque se realizaron varios tipos de evaluación para cada uno. Para eliminar duplicados, tenemos que escoger cuál de todas las evaluaciones hechas es la más segura. (En este caso, los registros conocidos como "Informes de Evaluación de Rendimiento" [PARs] parecen ser los más confiables porque ellos ofrecen una imagen mucho más fuerte de la evaluación. Estos son desarrollados por la Independent Evaluation Group [IEG], el cual de forma independiente y aleatoria muestrea 25% de los proyectos del Banco Mundial por año. IEG envía a sus expertos al campo para evaluar los resultados de esos proyectos y crear evaluaciones independientes.

3. ¿Son los datos exactos?

Una de las mejores maneras de evaluar la credibilidad de un grupo de datos es escoger un registro de muestra y compararlo con la realidad.

Si clasificamos la base de datos del Banco Mundial (que supuestamente contiene todos los proyectos desarrollados por la institución) en orden descendente por costo, encontramos que un proyecto en India fue el más costoso, que aparece con una cantidad total de US$29,833,300,000.

Si buscamos el número de proyecto en Google (P144447), podemos acceder a la documentación de aprobación original para ambos, el proyecto y su crédito, que cuenta efectivamente con un costo de US$29,833 millones. Esto significa que la figura es exacta.

Siempre es recomendable repetir este ejercicio de validación en una muestra significativa de los registros.

4. Evaluando la integridad de los costos

Desde el momento en que se introduce por primera vez en una computadora hasta el momento en que se accede a ellos, los datos pasan por varias entradas, almacenamiento, transmisión y procesos de registro. En cualquier etapa pueden ser manipulados por personas y sistemas de información.

Es por lo tanto muy común que las relaciones entre tablas o campos se pierden o combinan, o que algunas variables fallan al ser actualizadas. Es por esto que es esencial realizar pruebas de integridad.

Por ejemplo, no sería raro encontrar proyectos listados como "activo" en la base de datos del Banco Mundial muchos años después de la fecha de aprobación, incluso es probable que muchos de ellos ya no estén activos.

Para comprobar, cree una tabla dinámica y agrupé los proyectos por año de aprobación. Luego filtré los datos para mostrar solo aquellos marcados como "activo" en la columna de "estado". Ahora veremos que 17 proyectos aprobados en 1986, 1987 y 1989 siguen listados como activos en la base de datos. Casi todos ellos están en África . En este caso, es necesario aclarar directamente con el Banco Mundial si estos proyectos siguen activos luego de casi 30 años.

Podemos, desde luego, realizar otras pruebas para evaluar la consecuencia del Banco Mundial. Por ejemplo, sería una buena idea examinar si todos los

Gráfico 3

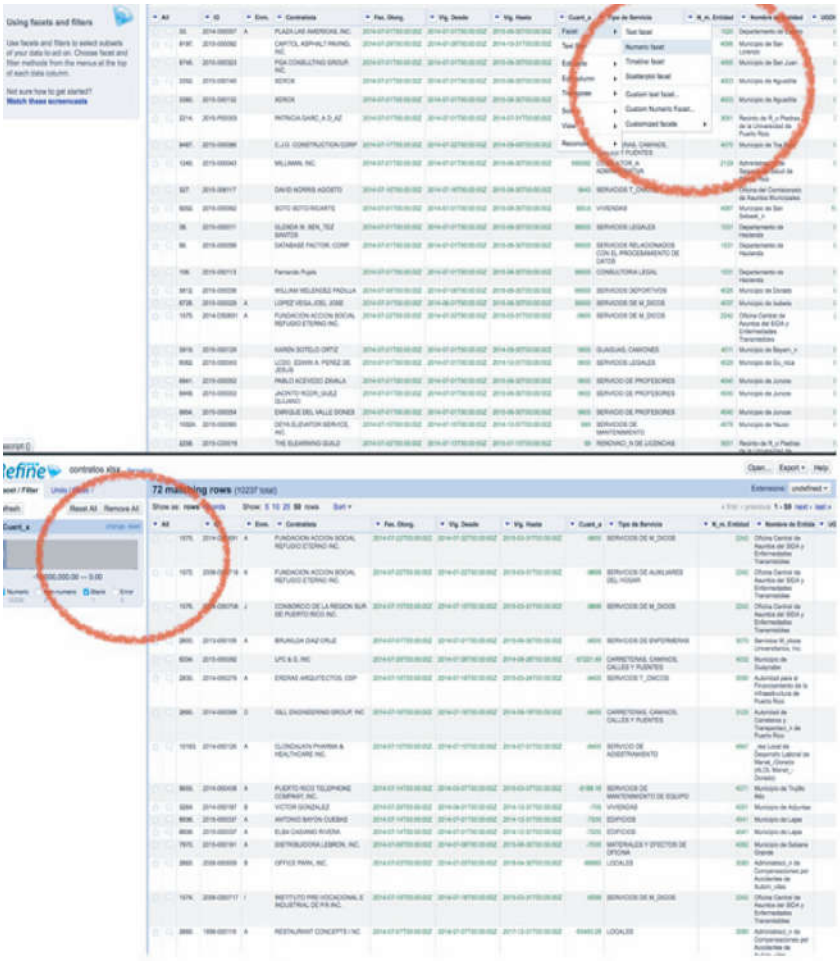


Gráfico 4

	A	B
1	Count of Project ID	
2	PROJECT ID	Total
3	P010162	3
4	P010410	3
5	P070371	3
6	P078994	3
7	P000075	2
8	P000082	2
9	P000090	2
10	P000099	2
11	P000101	2
12	P000105	2
13	P000106	2
14	P000120	2
15	P000133	2
16	P000138	2
17	P000139	2
18	P000142	2
19	P000145	2
20	P000147	2
21	P000173	2
22	P000178	2
23	P000183	2
24	P000186	2
25	P000188	2
26	P000258	2
27	P000261	2
28	P000263	2
29	P000267	2
30	P000270	2
31	P000274	2
32	P000276	2
33	P000282	2
34	P000285	2
35	P000290	2
36	P000293	2
37	P000304	2

beneficiarios de los préstamos (identificados como "prestatarios" en la base de datos) corresponden a las organizaciones y/o a los gobiernos actuales de los países listados en el campo "Nombre de país", o si los países se clasifican dentro de las regiones correctas ("nombre de región").

5. Descifrando códigos y siglas

Una de las mejores formas de espantar a un periodista es mostrándole información compleja plagada de códigos y terminología especial. Esto es un truco preferido por los burócratas y organizaciones que ofrecen poca transparencia. Ellos esperan que no sepamos cómo dar sentido de lo que nos dan. Pero los códigos y las siglas pueden además ser usados para reducir caracteres y apalancar la capacidad de almacenamiento. Casi todos los sistemas de base de datos, ya sea público o privado, utilizan códigos o siglas para clasificar información.

De hecho, muchas de las personas, entidades y cosas de este mundo tienen mucho o varios códigos asignados. Las personas tienen un número de identificación, número de seguro social, número de cliente de banco, número de contribuyente, número de viajero frecuente, número de estudiante, número de empleado, etc. Una silla de metal, por ejemplo, es clasificada bajo el código 940179 en el mundo del comercio internacional.

Traducción: Mayra Valera y Melanie Betetta de Convoca.

Este texto se publicó originalmente en inglés como parte de una guía de verificación de contenidos digitales editado por Craig Silverman, editor de Regret the error de The Poynter Institute. Ver AQUÍ.

*Giannina Segnini es profesora de la Universidad de Columbia y fue editora de la Unidad de Investigación y de Inteligencia de Datos del diario *La Nación* de Costa Rica. Es miembro del Consorcio Internacional de Periodistas de Investigación (ICIJ) y ha ganado diversos premios internacionales a lo largo de más de dos décadas de experiencia en el periodismo.

WWW.CONVOCA.PE
/CONVOCA
@CONVOCAPE

#ABCDATOS

Si deseas contarnos tu experiencia de investigación y trabajo con datos, escríbenos a este correo info@convoca.pe.
Suscríbete a este link y únete a nuestras redes sociales

Octubre de 2015