OjoPúblico

THE SWISS ARMY KNIFE FOR JOURNALISTS

Digital investigative tools in the era of Big Data

By David Hidalgo & Fabiola Torres

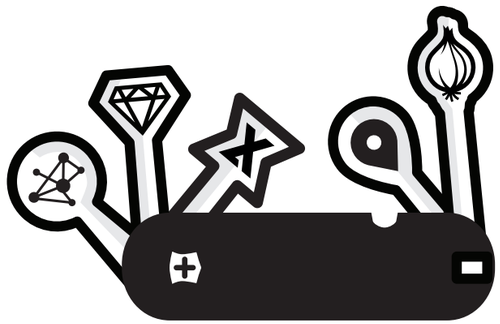
Translated by Florencia Melgar & Kirsty Styles

OjoPúblico

THE SWISS ARMY KNIFE FOR JOURNALISTS

Digital investigative tools in the era of Big Data

By David Hidalgo & Fabiola Torres



Translated by Florencia Melgar & Kirsty Styles

The Swiss Army Knife for Journalists

Research: Fabiola Torres and David Hidalgo
Assistant: Karina Valencia
Design: Kati Sanabria
Traslation: Florencia Melgar and Kirsty Styles

Spanish version (printed): 1a edición, Febrero 2016
English version (digital): August 2016

The original publication was possible with the support of the Peruvian Press Council, Hivos Foundation and The International Institute for Democracy and Electoral Assistance (IDEA).

Asociación de Periodismo de Investigación OjoPúblico
Jr. Pablo Bermúdez Nro. 150, Int. 11-A, Urb. Santa Beatriz
Lima - Lima - Lima.

Index

PROLOGUE

When Fabiola Torres and David Hidalgo first approached me with the idea of writing this prologue, I started wondering what a composition of my ideal prologue would be like:

-Visually appealing and entertaining so it won't waste the time I hope to spend on reading the actual book

-It would have extra information about the authors

-It will let me know if I am the type of reader they thought about when they wrote the book

-It will allow me to understand the importance of its content in the current context

So here's my best attempt. 

As you can see, I have decided to write it by hand, so it will be visually interesting. At the same time, a hand-written prologue in a data journalism book is kind of a synthesis of how young journalists feel today. Obviously, I first wrote this text in my computer, using a little bit of help from google translator but ended it by hand, adding extra-information. This kind of process has a lot to do with the mix of digital and analogic techniques journalists use nowadays.

We (current young journalists) belong to the generation of professionals that have received an academic formation destinated to perfectly fit into the traditional media universe with the little help of the Internet but the archaic process of the print paper. It's kind of important to clarify that I had never worked in traditional media. That had something to do with the lack of job opportunities while I started my career and my -kind of accidental- thrive to the entrepreneurial world at a young age.

I call our generation, "the in-betweens" because we had no train to think or produce journalistic pieces or journalistic products in the

current context. When I say “current context” I mean: social media, snak content, visual representation of the information, data processing and engineering, engaging, highly FOMO (fear of missing out) readers and an exponential information development.

What first came to my mind while I was devouring the pages of this amazing book, was: “What would have happened if I would have read this while I was studying journalism?” and the second thought was: “Can a student be prepared to face the current journalistic challenges without a good understanding data journalism?”. If you have this book in your hands you are about to find the answer to the second question.

David and Fabiola have created what I feel is more like a compass than a swiss army knife. In a few pages, you will easily find your destination to a world of incredibly vast information that could create an infinity of investigative opportunities.

I won’t duplicate the information you are about to read, instead I will tell you something that amaze me the most about the existence of this piece. While Fabiola and David wrote this book they were: publishing Ojo Público’s first investigations with an enviable quality; building and expanding their team; creating their diversify business model (a mix of grants and donations, and web and training services); and growing a large amount of followers and repercussion around the world.

They really fit into what I think journalism is today:

Journalism is a service for the audience, readers know better and are the real boss

Collective and interdisciplinary work is the only way to go

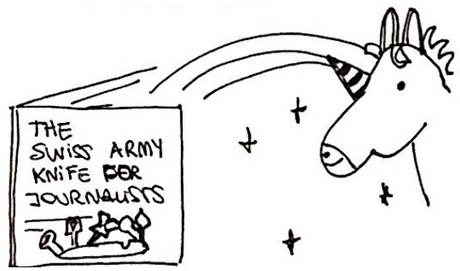
News it’s no longer to find only on the streets, it is also at our computers in a huge data base :)

Today journalism quality also depends on programing and designing

Sharing your knowledge benefits your colleagues and yourself

I think is important for you to know that Ojo Público’s founder team (which includes the journalists Nelly Luna Amancio and Óscar Castilla, and the programmers Antonio Cucho and Jason Mártinez Vera) is a very humble media startup that -in less than 2 years- has more journalistic accomplishments than an average traditional media founded decades ago. And THAT is a very good reason to follow their work and all the recommendations you are about to discover in this great book.

Mij Ebner
Regional Director and Co-founder
Sembramedia



THE NEW LITERACY FOR JOURNALISTS

How an encounter with a hacker accelerated the reinvention of journalism

1

The ecosystem of digital journalism

After becoming the most wanted man on the planet, the US security analyst Edward Snowden sat down with a couple of journalists and with a quiet voice recorded a phrase that could become a psalm to future civilisations. "Technology is the greatest equalizer of human history". 1 When he said this, he was hidden in the room of a Moscow hotel at the risk of being caught for revealing the biggest clandestine surveillance machinery that has ever existed. Instead of a cry for help, it was a message about the true sense of the digital revolution. "It helps us to adopt new faces, enter new communities, engage in new conversations and discover who we are and what we want to become." Snowden called for a fight against the powers that seek to use technology to tell whether people are good or bad. "It's not the governments who are the ones to decide. We are". His way of contributing to this fight was to become the most prolific journalistic source of all time.

Snowden is the epitome of this era of global informants. When his efforts came to light, the soldier Bradley Manning was already in prison for leaking further government secrets and Julian Assange had been granted asylum in the Embassy of Ecuador in London. Unlike Manning, who left a trail before the leak, and Assange, who made his activism a personality cult, Snowden only made his first move after a millimetric calculation. This allowed him to suggest things with a visionary voice, which was very well structured, on the impact his leak would have not only for American society, but for every person on the planet. The axis of this strategy was to ensure the intervention of Glenn Greenwald, a journalist known for his background as a human rights lawyer and his coverage of the massive espionage committed by the NSA, the US intelligence agency specialising in information gathering and analysis. The contact between the pair was a test of the challenges faced by journalists trying to

understand and record facts of public interest in today's digital society. The world's former most famous informant, Deep Throat, not only leaked secrets to the journalist, he also had to teach him about the technical resources required before starting work.

The first contact occurred in December 2012. Snowden sent Greenwald an email under a pseudonym that started by defending the need for secure personal communication. Speaking anonymously, the author explained that by using a normal email account, the journalist could be putting at risk those people who want to share sensitive information. This wasn't a surprise; stories about internet espionage made headlines in the first decade of the 21st century and have in recent years focused on the role that the giants of online communication have in facilitating government surveillance in countries with little democratic tradition, such as China or Syria. The mysterious source suggested that the journalist should install an encryption program, software that allows the encoding of passwords and messages, to make it impossible for a third-party, or government agency, to intercept the communications. He even offered to give him a hand if he found it difficult. "For some time I wanted to use encryption software," Greenwald writes in the book 'No Place to Hide', which covers the details of this investigation. "However, the program is complicated, especially for someone like me, little versed in programming and computers. It was one of those things you never find the time for".[1]

The program mentioned by Greenwald is called PGP and stands for Pretty Good Privacy. It is a popular tool among hackers and all kinds of people living at risk of being spied on. It works with a special key that one shares with the other person to establish a secure contact. "In essence, the program involves emails with a protective shield which is a password composed of hundreds, even thousands of random numbers and case-sensitive letters," says Greenwald in his

"Computers don't make a bad reporter into a good reporter. What they do is make a good reporter better."

Elliot Jaspin, *Pulitzer Prize-winning investigative reporter.*

1 GREENWALD, Glenn. "Sin lugar donde esconderse. Edward Snowden, la NSA y el estado de vigilancia de EE.UU.". Barcelona: Ediciones B, 2014.

SECURITY BOX

[Tools to avoid being spied on while using the internet]

Tor Project

https://www.torproject.org/

Free software for secure communications. Hides the IP address of the devices used. Navigates the web undetected and leaves no trace of visited websites or the geographic location of the user.

Mozilla Thunderbird

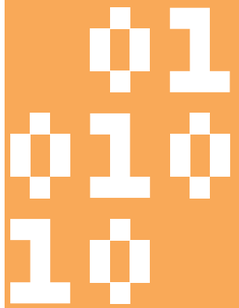
https://www.mozilla.org/en-US/thunderbird/

Free email program (with secure code) to receive, send and store emails. You can also manage multiple email accounts.

Enigmail

www.enigmail.net

Thunderbird add-on that allows you to send e-protected encrypted keys. The user keeps the key. To use Enigmail, you must also install GNU Privacy Guard (GnuPG).



book. Even the most advanced decipherment programs built by the most powerful intelligence agencies would take years to get through this protection. Although Greenwald knew its benefits while writing about cases like Wikileaks and Anonymous, he had not incorporated it into his tools, perhaps as yet unwilling to spend the time.

Days later, the anonymous sender wrote again with a series of instructions to install the program. He even offered to put him in contact with an expert to help him get started. Greenwald said he would do it, but didn't get around to it. He had a big workload and nothing guaranteed that he would get a great story. Weeks later, the stranger insisted and tried to make things easier by sending a tutorial video titled 'PGP for journalists'. Even then, Greenwald took no action. Nor did he in the following two months. By then, the informant had found another way to continue with his plans: he contacted the documentary maker Laura Poitras, simply because she did use encryption software. It was in fact Poitras who first knew the dimensions of the source and his secrets. "That's how close I was to lose the most significant and far reaching national security leak in US history," Greenwald would later recognise. He was lucky Snowden insisted on working with him.

Later, a series of security measures were put in place, which included new encrypted emails, more secure keys and the help of a computer security expert. Laura Poitras herself commissioned this ally technologist to teach Greenwald to use a system that is still not well known, called Tails (The Amnesic Incognito Live System), which can only used from a portable device.[2] The expert prepared a special version for the reporter on a blue USB stick and mailed it to Brazil. It is the accessory that appears connected to the computer while Greenwald interviewed Snowden in a room in Hong Kong for the documentary Citizenfour. This is how the story began. On his

2 LEE, Micah. "Ed Snowden taught me to smuggle secrets past incredible danger. Now I teach you". The Intercept. Link: https://theintercept.com/2014/10/28/smuggling-snowden-secrets/ [Visualized on November 22, 2015]

44

trillion gigabytes of data will be in the digital universe in 2020, according to projections by tech consultancy IDC.

current Twitter profile, Glenn Greenwald highlights his PGP public key and its corresponding fingerprint, a shorter number, only 40 digits, which facilitates the confirmation of the key. His meeting with Snowden represented a piece of history in which technology is no longer an accessory, but part of the habitat of human experience.

The most shocking leak of information this century led to the idea that one day journalists will have to imitate astronomers to capture some certainties in the expansive digital universe. Not all of us will be writing about Western intelligence service espionage or about certain governments' plans to capture internet – even though we should have it on the agenda - but it is just as valuable as the tools we use to understand the big data era. Even before the attacks of September 11, it was known that the major US intelligence agency was intercepting 1,700 billion communications every day. With Snowden's revelations, it became publicly known that in 2013, the same agency managed to capture a trillion pieces of metadata, used for identifying what people look at online, their hobbies and even their future behaviour. That same year, an American professor estimated that the volume of stored information worldwide was 1,200 exabytes, equivalent to covering the entire surface of the United States with books some 52 times.[3] There is almost no process that can't be quantified. But how do you understand an amount of data that, if stored on CDs, would be enough to build a tower from the Earth to the moon?

In essence, a change is required in the technical capabilities and operational thinking of the investigative journalist. The classic metaphor of the profession has to change, from having a toolbox to managing lab equipment. "Journalists don't need to learn to program, but they need to develop a mentality of massive data, so they unders-

3 SCHÖNBERGER, Viktor y Kenneth CUKIER. "Big data, la revolución de los datos masivos". Madrid: Turner, 2013.

tand that the data contains hidden stories to be told," says professor Víktor Schönberger, internet expert from the University of Oxford.[4]

The last big revolution in journalism was called 'New Journalism', and it was about scenes, dialogue and perspective: the experimental techniques that changed the way to tell good stories. This next new way of crafting stories means mastering concepts in the newsroom just like those found in a robotics workshop: encrypt, scrape, refine, visualise. This is a meta-language to define how we handle the data.

Neither the applications or the software define this new moment of journalism. But they both give us the possibility to find answers to questions that once seemed better placed in science fiction. "The data can reveal secrets to those who have humility, desire and tools to listen," wrote Viktor Schönberger with Kenneth Cukier in the book 'Big Data, the Revolution of Mass Data'. We are now more similar to astronomers in that we face such a big data universe, we have to improve our instruments overnight.

BIG (REALLY BIG) DATA

At the beginning of the century, the American economist Steven Levitt envisioned 'Freakonomics'. It was about finding surprising truths based on the way data was analysed. Levitt did this by pushing the logic to an extreme in his questions: "why do drug traffickers continue living with their mothers?" or "what is more dangerous, a gun or a swimming pool?". In one of his essays, he asks his readers about a couple who avoid sending their daughter to a neighbour's house because the father of the household keeps a gun at home, but they let her go to another friend's house where there is a pool in the yard. The question considers which was the right decision for the safety of the child. Levitt found that according to statistics, there is one child death per 11,000 pools in a country with six million. That's an average of 550 children drowned a year. Conversely, one child dies from a gunshot wound per 1.5 million arms. In a country with 200 million guns, the proportion is 175 children shot dead a year. Translation: a girl is statistically more likely to die in a pool than playing with the neighbour's father's gun. If humans often change their behaviour with scaled samples like this, what happens when the size of the information goes beyond our ability to store it? "The era of massive data calls into question the way we live and interact with the world," explain Schönberger and Cukier.

44%

of the data in the digital universe requires some protection, from security measures to cryptography, according to IDC.

"We need to humanize and personalize the large datasets in a way that does not affect the complexity or scale of the issues in question".

Paul Bradshaw, *author of Online Journalism Blog.*

Levitt's favourite case was when the most famous search engine in the world saved America from a global epidemic. When bird flu arrived in the country, health systems collapsed due to the lack of timely information available to help plan strategies. The alert system was too slow to understand the spread of the disease. At that time, the science journal Nature published an article showing that Google engineers had found a way to predict the spread of the common flu. The method was to combine search trends for symptoms with historical information about the evolution of the disease. "Others had already tried to do this with internet search terms, but nobody had so much data, the capacity, or the know-how to process it, as Google does", the experts say. Just validating the keywords and phrases to start predicting flu outbreaks meant creating 450 million different mathematical models. The result was a group of 45 terms that showed the relationship between the search made by prospective patients and the evolution of flu. Unlike the traditional method, which could take weeks to retrieve the information, Google developed a software that could achieve that accuracy in real time.

What can we learn as journalists of today, when an algorithm can predict the moment when millions of people will be wiping their noses? "The big data refers to things that can be done on a large scale [...] to extract new insights or create new forms of value that transform markets, organisations, relationships between citizens and governments," said Schönberger and Cukier. Since the psychiatrist

4 GONZALO, Marilín. "Los datos masivos (o big data) son el nuevo oro". From eldiario.es (Spain), August 5, 2013. Link: http://www.eldiario.es/turing/Big-data_0_161334397.html [Visualized on November 22, 2015]

GUIDE TO INVESTIGATE WITH DATA

Paul Bradshaw, author of the blog Online Journalism, argues that the journalistic work with databases comprises five stages: gathering, debugging, analysis, verification and presentation of findings. We can take that sequence to propose the following exercise when starting a project.

The collection of data

You must know the file formats that contain the information and the tools you need to collect them. You can get masses of data using a script, a simple program that enables automated downloading of information. This process is known as scraping.

Are there any databases on the subject? How and why were they created? Are they on an official websites or should I make an FOI request?

If the database is on the web, it is downloadable or do I need to scrape it?

What is the best format (Excel, CSV, Json) to request a copy of these databases in? If the information is in PDF or JPG, how can I change its format to an Excel file?

If I build a new database: what variables should I include and what can I prove?

Debugging and context

There are several types of errors you might find: duplicated records, incomplete boxes, misspelled words, etc. You will need tools to identify and solve these problems. This is called 'cleaning the data'.

Is the database complete? How many lines of information does it have? Can I clean it using Excel or Open Refine? When do I need to do it by hand? When do I have to use data managers with more capacity, such as MongoDB?

Do I know and understand all of the terms, variables and acronyms that appear in the database? Are they the same ones used in similar databases? Is the criteria aligned with the meaning of the question I want to answer or do I need to see the same data in reverse?

Crossbreeding and analysis

In this phase, the value of the findings depends on the quality of the questions asked and the combination of two or more records to find revealing matches.

Do my databases have a concept or code in common that will allow me to cross the datasets: ID, Tax Code, full name?

Does the database crossing show trends, patterns or evolutionary processes in a given period? In what context?

Or, on the contrary, does it reveal atypical behaviours? In what context?

Verification

Investigative journalists must apply the traditional methodology: go to the necessary places, interview the people involved, and review new documents to detect weaknesses and strengths of the database.

Does the data represent the actual condition of the people? What has changed in the life of the alluded people: his health, economic stability, legal status or relations?

Does it affect the meaning of the findings? Does it confirm its relevance? Does it stress it or make it relative?

Is there an expert that can validate the methodology of the crossing? Is it possible to have correct findings that offer more than one interpretation?

Presentation

To present your findings in the most efficient way, you have to think about this from the beginning. There are libraries like d3js.org and software repositories like Github with examples that you can adapt to what you need.

What is more convenient: visualisation or an app? Which one will make the best contribution to the meaning of the story?

What should the user experience be like? What emotions should the chart or tool induce in the reader? Which elements of my app or visualisation are necessary for the user?

Is the tool responsive? Will it look good on mobile phones and tablets? Can it be shared? Can it be embedded?

OPEN REFINE, A SOFTWARE WITH STEROIDS

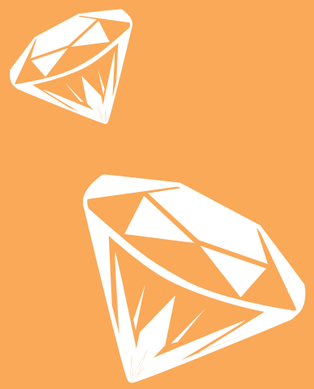
[How to detect data errors among huge data]

Anyone working with spreadsheets knows that **there are four common problems**: misspellings, names and words written in various ways, and invisible characters or spaces. They do not seem complicated for those who use a personal Excel files, but they are a nightmare when we manage databases with millions of lines. In these cases, **the most useful option is using Refine Open**, an open source tool that **allows you to debug and organise data in just a few steps**. We chose a log of gold exporters as a sample. It is possible that the name of a company has been filled out in various ways (OrogoldSA, OrogoldS.A. and OroGold). An initial analysis will count them as different companies. **Open Refine finds the matches and allows them to be edited in one single action** to give them uniformity.

In addition, **if there is an error when editing the database, it's possible to return to the previous version** or display the change history over time.

The program can be downloaded from http://openrefine.org/download.html. It is compatible with any browser and **it's available for Windows, Mac and Linux.** It also allows you to change the format of files, such as XLS, CSV, JSON, XML, TSV and Google spreadsheets.

It's a much needed resource in the toolbox of the investigative reporter. Some journalists define it as "Excel with steroids".



300

thousand records of companies registered in Panama were downloaded by the programmer Dan O'Huiginn, which he added to a searchable website.

Carl Jung deciphered how humans dream, no other operation to find stories in the abstract world was so powerful.

"Facts are sacred," says an old motto of British newspaper *The Guardian*. That's the principle that guides the work of its former digital editor Simon Rogers, the journalist who transformed statistics in graphics than remind Mondrian's paintings. "Most of the time we are the bridge between the data (and those who are desperate to explain) and people in the real world trying to understand what the story is really about," Rogers says in *Facts are Sacred*, a guide to how to convert data into visual concepts. In 2011, a team led by Rogers, who now heads Google's Digital Laboratory, explained England's civil administration using a bunch of colourful balloons. At a glance, readers could now understand a bureaucratic structure that was originally a list of more than 200,000 names, positions and salaries. The analysis allowed people to see that at least 90 bureaucrats were earning more than the British Prime Minister.

On another occasion, journalists David Leigh and Nick Davies, the research team at The Guardian, obtained a file from Wikileaks with detailed information about all military incidents recorded by the US army during the war in Afghanistan. It had been created by the soldiers to monitor their actions. The first step of their investigation was to obtain the encrypted information on email. The challenge was to review the data and find the story from a spreadsheet with 91,000 rows and 201 columns. The task was too overwhelming for the journalists and even the newspaper's systems experts found it difficult to handle. "It's like finding small gold nuggets amid a mountain of data," Leigh went on to say.[5]

By that time, the newspaper already had experience of handling large databases released by the government and had even created internal explorer tools that allowed journalists to search files like this. This time, they were again able to use this. The data was filte-

5 LEIGH, David y Luke HARDING. "Wikileaks y Assange". Barcelona: Ediciones Deusto, 2011.

10 DATABASES TO TRACK THE MONEY AND OTHER LEADS

This is a suggestive selection collected by Ojo Público from those attending the Global Investigative Journalism Conference 2015.

ICIJ Offshore Leaks Database

http://offshoreleaks.icij.org/search

It has data on more than 100,000 companies and funds held in tax havens. It is part of an archive of 2.5 million documents leaked to the International Consortium of Investigative Journalists (ICIJ).

OpenCorporates

https://opencorporates.com/

It contains information on 80 million companies and 90 million directors in more than 100 countries. You can search by company name, address and name of director.

Persons of interest

https://www.personadeinteres.org

It gives access to court records, property records and intelligence reports on people linked to organised crime, drug trafficking, corruption, etc. It has information on trials for drug trafficking in Peru.

Open Spending

https://openspending.org/

Website that follows the public spending of governments around the world and presents it in various forms of visualisation.

Registering property in Miami

http:// www.miamidade.gov

Miami Dade County's free data, which can locate properties in Miami, on behalf of the owner.

Search Systems

http://publicrecords.searchsystems.net

Specialised portal. It contains more than 55,000 databases by date of birth, date of death, marriage, licenses granted, stocks, mortgages, among other subdivisions.

Registration of companies Panama

http://ohuiginn.net/panama/

Independent site that reordered the public record information from the Panama leak to facilitate the research of more than 300,000 companies registered in the country. It allows searches by people's names.

FlightAware

https://es.flightaware.com/

A worldwide platform, including a mobile app, to track airline flights and their status. It allows you to search by the name of the airline, flight number, route and aircraft registration.

Marine traffic

http://www.marinetraffic.com/

It is a database updated in real-time that can track the location of any boat or ship, in addition to departures, arrivals and routes.

Investigative Dashboard

https://www.investigativedashboard.org/

Allows you to search shareholders, directors and financial reports of companies around the world. There are links to more than 450 online databases in 120 countries. Platform built by the Organized Crime and Corruption Reporting Partnership (OCCRP).

red according to a determined order, including date, time, description of the attack, number of victims and the coordinates where it occurred. The analysis established that the number of attacks that used homemade devices - the most unpredictable and lethal - had increased, and this increase had occurred precisely in the areas controlled by the UK and Canadian armies. Now the challenge was to find the best way to tell that story. It was then that a team of visualizers, led by Simon Rogers, joined the work. "The Wikileaks project was producing new data types, so they needed to be extracted with new types of journalism," Leigh wrote with Luke Harding, one of the reporters who participated in the research.[6] El resultado fue un mapa que mostraba por primera vez la evolución de seis años de atentados en ese país: entonces se confirmó que la racha sangrienta había dejado más muertos civiles que militares.[7]

"The story of Wikileaks is a combination of two things: the knowledge of traditional journalism and the power of technology, working together to tell an amazing story," wrote the pair who rebuilt the case.

Even with these samples of the potential of the data, until recently, many journalists, including researchers, tended to think that technology was an alien language. The very idea of investigating an Excel file with more than a thousand rows discouraged the traditional scourer of confidential documents. "You do not have to be a programmer," says Rogers in his book. "You can become luxury encoder if you want it that way, but the main task is to think of the data as a journalist rather than as an analyst". No digital tool will replace the exercise of asking what can provide relevant information or what would happen if you mix a database with another one, like the original economist of the weird thigs. What you cannot do on your own, you can do it with allies from this parallel universe.

88

cities in the world have 'hacks and hackers' communities that facilitate the collaboration between journalists and programmers.

"I never begin a project without a plan. I use a matrix developed by myself, tha allows us to focus on speficic questions".

Ginna Morelo, *Data Editor at El Tiempo (Colombia)*

CODE OF FORCE

It was only a matter of time before someone invented a space to integrate these two ways to view the information. The idea came from a chance meeting between a young correspondent and two veteran journalists. Burt Herman was a reporter for the Associated Press who had spent 12 years travelling around sensitive areas of the world, from Korea and the former Soviet Union, to the convulsed Iraq and Afghanistan. Between 2008 and 2009, Herman left the agency and chose a scholarship to explore innovations in journalism at Stanford University. Nearby, the digital maelstrom of Silicon Valley had begun to organise meetings of people interested in journalism and technology. Around the same time, Aron Pilhofer, editor of the New York Times, and Rich Gordon, a professor at Northwestern University, launched a call from Massachusetts to form a network to develop apps and digital tools to process information. Both initiatives agreed on a concept: to unite hacks, a term that refers to the ability of journalists to produce written stories, with hackers, the prolific writers of source code, otherwise known as the instructions that run the machines.[8]

Such a cross of languages would be enough for an episode of Star Wars: two alien races – one from the other one, at least - had arrived to an agreement to fulfil a mission. The only possible way to achieve this is to exchange knowledge: journalists learn the hackers' jargon and the principles governing cyberspace. In return, they are trained to use their skills to make sense of the information. The proof is in the experience of Burt Herman himself. While conducting his scholarship, he came into contact with Belgian programmer Xavier Damman and together they set out to create a tool that could exploit information from social networks.The result was Storify, an application that helps users bring together photos, videos, tweets and links to tell a story that can be inserted into any website. "The way to

6 Op. cit.
7 ROGERS, Simon. "Facts are sacred". Londres: Faber and Faber Limited, 2013.

8 HIDALGO, David. "Periodistas buscan hackers (de los buenos)". Link: http://hhlima.info/node/8 [Visualized on November 25, 2015]

make sense of the media is through human curation with the help of technology," says Herman.

This alliance is already generating changes in global journalism: the HacksHackers community has chapters across all continents. In each place, it has facilitated the creation of tools to process large amounts of information. In early 2014, for example, the chapter of Rosario, Argentina, collected data from the Ministry of Justice, police reports and newspaper articles, and built an interactive map where you can see the exact point where every murder occurred the previous year. "The purpose of the project was to create a platform that would demonstrate, through data visualisation, the increase of social violence in the city," writes Ezekiel Clerici, one of the organisers of that community. Some time before, in 2011, the Buenos Aires chapter created an application that allowed real-time monitoring of the presidential election results: you mark on a map the place you're interested in and the application gets the corresponding data. In all cases, the last axiom of the information age is confirmed: the problem is not the changes in the methods used by the journalist, but what we mean by doing journalism.

As is expected in this time of massive data, the potential is huge. In June 2014, several Latin America chapters joined in a regional hackathon to create tools that allow journalists to monitor the use of public funds. The activity was baptised with the expression that has always guided the best journalism: "the money trail". The Lima chapter gathered more than 50 members who locked themselves away for 12 hours in the auditorium of a school dedicated to teaching technology. During the day, teams of journalists and developers worked together to analyse information from various databases and make it news. In essence, it was about looking at the sexy side of an Excel table..

At the end of the afternoon, the community presented seven projects, ranging from the analysis of how funds are invested by the Environment Ministry to an estimate of the money allocated by the

"When I design a spread sheet for an article I think about the goals: what do I want to know and what are the possible patterns in the data".

Lise Olsen, *investigative reporter for the Houston Chronicle.*

ONLINE CATALOG

[News apps developed by Ojo Público]

Suprema Fortuna / Supreme Fortune

http://supremafortuna.ojo-publico.com/

News app that shows the evolution of the assets of the judges in Peru, their careers, and other information useful to draw a profile of the justice in the country.

Fondos de Papel / Paper money

http://fondosdepapel.ojo-publico.com/

Interactive report that offers an unprecedented xX-ray of the private funding of the Peruvian political parties and their presidential candidates during the 2006, 2011 and 2016 campaigns.

Congreso Airlines

http://ojo-publico.com/sites/apps/congreso-airlines/

Application that shows the international trips made by 113 congressmen in Peru and the expenses of the parliament, between 2011 and 2015.

Cuentas Juradas / Sworn Accounts

http://cuentasjuradas.ojo-publico.com/

It's a platform that shows - from 2003 to 2014 - the evolution of the wealth -declared by 38 mayors of Lima seeking re-election and 23 former mayors who wanted to return to power in the 2014 municipal elections.

state to the Catholic church. One of the most interesting tools was an application that allows the search of databases and specialised pages to identify links between public servants and organised crime, and the possible relationship with public funds. Another was an application that allowed the systematisation of the main state suppliers to see how much money they have invoiced to the country. Had it not been for that meeting, which functioned as a workshop for curious professionals, the information would still be buried in disjointed files. “This group brings together all these people: those who are working to help people make sense of their world,” says a statement in the original page on HacksHackers.

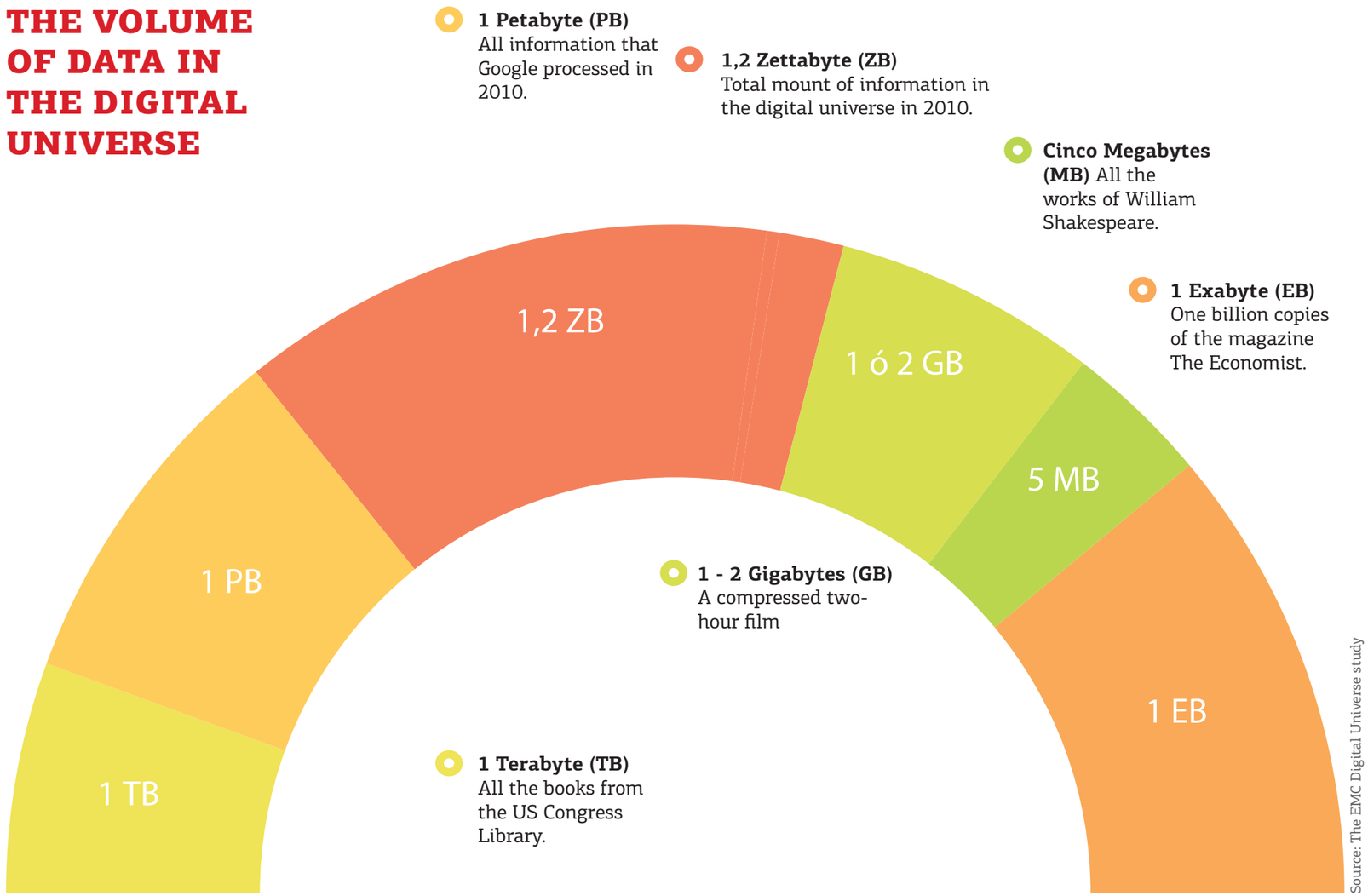
THE ALLIES

One afternoon in October 2010, the Costa Rican engineer Rigoberto Carvajal decided to quit his job in a software development corporation to enrol in journalism. Through a friend he had learned that the newspaper La Nacion was looking for a programmer for the company’s research unit. Carvajal applied for the job out of curiosity to see if a guy like him could be useful for uncovering secrets. After an initial interview, he was convinced it was the right place. “I studied programming because I like to solve problems and in journalism one can do it with a nobler purpose than increasing a business’ profit,” says Carvajal, who is now an expert in databases for the International Consortium of Investigative Journalists (ICIJ).[9]

The project he was asked to work on was a challenge for a Latin American newspaper: to gather all public databases of each country to be analysed, then find new knowledge and relevant stories to investigate. It was expected that the IT team would have the same boldness and passion for the truth as journalists do. In the job interview, Carvajal said he once managed to locate the whereabouts of a person only using their name, because he managed to contact the people who could give him remaining information he needed.

9 Personal interview with Rigoberto Carvajal.

THE VOLUME OF DATA IN THE DIGITAL UNIVERSE



Source: The EMC Digital Universe study

His first test showed that he'd kept this ability. When he was asked to find Shakira's properties in the US, the developer sought the real name of the singer and then tracked this down in property records. Over time, Rigoberto Carvajal - a fan of science fiction - was transformed into a sort of Spock, a character between human and Star Trek: a hybrid of computing and journalism.

"Back to what I did before, programming for commercial interests, would be returning to the dark side of the force," he says, referring to another of his favourite movies.

The same spirit encouraged Brit Dan O'Huiginn when he downloaded the entire public registry of Panama in 2010 for a research project on arms dealers. Until then, the official website only allowed searches based on the names of the companies. That presented a limitation for the investigations of reporters who were following the trail of suspicious characters. O'Huiginn extracted data from more than 300,000 companies, ordered the information and used it to create a site that allowed searches based on names of individuals.[10] The programmer usually says that work had nothing to do with illegal piracy. He just used his technical skills to automate the data collection. "I do not mind being called hacker in the literal sense," he said when the Panamanian media were surprised that an unknown subject, from a terminal on another continent, achieved such access. Its essence, he explained, is "a person who enjoys exploring the details of programmable systems and how to stretch their capabilities, as opposed to most users, who prefer to learn only the minimum necessary".[11]

The site created by Dan O'Huiginn, which receives 2,000 visits a day, allowed investigative journalists from many countries to verify whether officials suspected of illicit activity and corruption were secretly recorded in Panama properties. Using this tool in 2011, reporter Khadija Ismayilova showed that the daughters of the President of Azerbaijan, Ilham Aliyev, ran a telecommunications firm through offshore companies.

345

Latin American organisations surveyed by the Organization of American States (OAS) had suffered data theft attempts.

This corporation had more than 500,000 subscribers, covered 80% of the territory of the country and, at that time, was the only provider of 3G services. The site also identified the offshore companies of former Egyptian President Hosni Mubarak, and also provided evidence that allowed the identification of five people connected with the murder of the former governor of the province of Panama, Darío Fernández. They were all convicted.

Since 2010, the programmer O'Huiginn decided to fully devote himself to working with investigative journalists. He has been a fellow of the African Network of Centres for Investigative Reporting and collaborated with projects of the Organized Crime and Corruption Reporting Project (OCCRP). Today he lives in Germany and works on the Openoil project, the first open data map of oil concessions of 18 countries in the Middle East.

"If done well, people really have much of an appetite to see the data," says Scott Klein, a benchmark figure of investigative journalism using massive data.[12] "It's enough to see how many people understand –and like- incredibly sophisticated and impenetrable sports statistics." If an average reader is willing to read sports pages that look like stock market reports, why not get them to see the usefulness of a tool that examines the health system or the quality of water drunk in your area?

Klein is editor of ProPublica, one of the US' most innovative media outlets. In 2010, he was commissioned to set up a project that looked like something created in a Silicon Valley lab: the department of news applications. It was a team of reporters and technologists working together to do journalism using software. One of his most impressive projects was Dollars for Docs, which revealed payments

10 http://ohuiginn.net/panama/ [Visualized on November 22, 2015]

11 "El Registro Público solicitó que eliminara mi web". At Diario La Estrella (Panamá), October 6, 2013. Link: http://laestrella.com.pa/panama/nacional/registro-publico-solicito-eliminara/23502395. [Visualized on November 22, 2015]

12 Cited at HOWARD, Alexander B. "El arte y la ciencia del periodismo de datos". Tow Center for Digital Journalism.

DIGITAL RESOURCES TO HUNT STORIES

[The desktop utensils of the investigative reporter]

ScraperWiki

https://scraperwiki.com/

Online platform where you can download information from the web and group it neatly in a database (Excel, CSV, etc.). It offers the possibility for anyone to create their own script according to their interests.

DocumentCloud

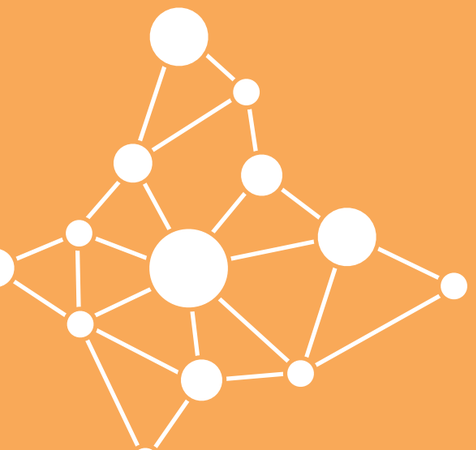
https://www.documentcloud.org/

Platform to manage documents. Extracts text from an image using Optical Character Recognition software. It can highlight data, annotate and organise it into links that are easy to access. It helps you search by subject, embed documents and place them in a public catalogue. Prior application is necessary to gain access.

Visual Investigative Scenarios (VIS)

https://vis.occrp.org/

A tool of special interest to investigative journalists. It allows you to establish relationships between people and organisations and attach documents to prove this relationship.



of $258 million to doctors who promoted the products of seven pharmaceutical companies among their patients. The team took advantage of a law that required laboratories to make public the money paid to doctors in commissions, lunches, conventions, and other ways. All documents had been published on the websites of the pharmaceutical companies, but were presented in complicated formats. ProPublica programmers developed a script – a small piece of software that automates processes - to collect all of the information. With this material, they developed an application that allows anyone in the United States to know if a doctor has received money from these labs, including how much and why.

Klein often refers to his team of five people as "programmers-journalists who think like reporters". This means they have the skills to handle digital tools, but they also have the instinct to detect a good story in a mountain of data. "Some have said that journalists with software development skills or vice versa are unicorns, rare. That's not true, you can develop in them the skills of journalism and engineering required today", he says.[13] The potential is greater than just making a visualisation that tells a story in a graph, or even an application that offers the potential to tell various stories, allowing the reader to make their own findings. It is the possibility of joining two rigorous and related methodologies that expands the frontier of knowledge. "Investigative journalism is the department of study and development of the profession," wrote Brant Houston, cofounder of the Global Investigative Journalism Network.[14]

Perhaps the image that best reflects this point is that conjured by Evan Smith, co-founder of the Texas Tribune. Smith says that journalists today must be like a Swiss army knife.[15] The idea of kee-

13 FALLAS, Hassell. "Simplificar es clave para crear aplicaciones de noticias". Blog La Data Cuenta. Link: http://hasselfallas.com/2014/09/11/simplificar-es-clave-para-crear-aplicaciones-de-noticias/ [Visualized on Noviembre 25, 2015].

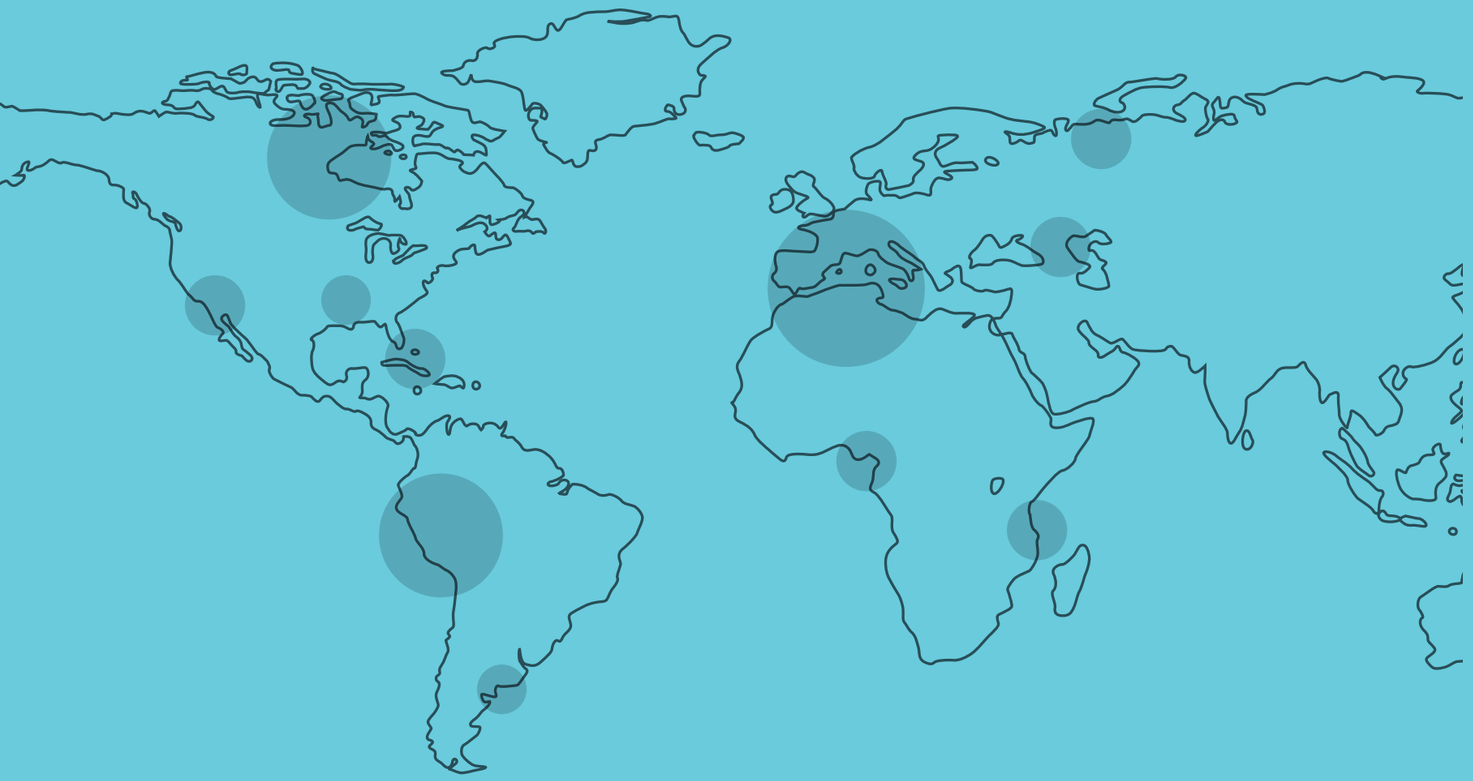
14 Cited at: KAPLAN, David. "Periodismo de Investigación Global: Estrategias para su Financiamiento". Center for International Media Assistance, 2013.

15 "Journalists today have to be swiss army knives". Interview with Evan Smith at The future of news. Link: http://futureof.news/episodes/evan-smith-2/?utm_campaign=refdotfonesmithvideo2&utm_source=Twitter&utm_medium=esmith2_jrsch_hd_flw [Visualized on November 25, 2015].

ping these skills as separate areas is an anachronism comparable to those who at the time resisted leaving the typewriter for the computer. The editor describes a scene that he still faces, despite his company's influence on multimedia projects. "People come and say, 'I want to be an editor', 'what kind of reporter?', 'I just want to write; nothing else but write'. 'Don't you want to record a video with your phone, edit and post it?' 'No'. 'Don't you want to record an audio on your phone, edit it and post it?' 'No'. 'Don't you want to do anything in basic HTML?' 'No'. 'Don't you want to be in charge of social media?' 'No'. 'Then, this is what you will do: go to Home Depot (a US materials retailer), buy a lot of wood, build a time machine and return to Esquire 1964, because that was the last time anyone had that job'." The figure of the knife is not rhetoric: the reporter cannot use these tools every day, but they have to be there when needed.

80

journalists from 26 countries participated in the investigation of the 'Luxembourg Leaks', which revealed illegal agreements to benefit 340 corporations.

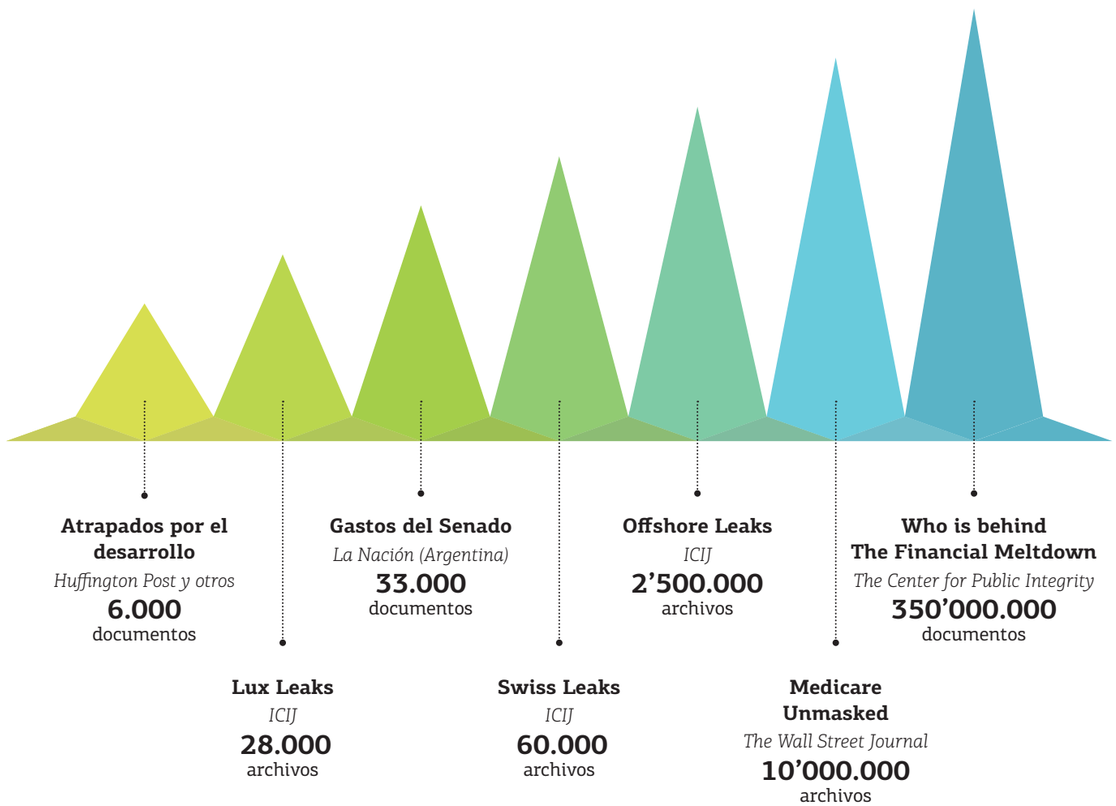


HOW TO TRACK CRIMES IN A DATABASE

Twenty research projects that changed the way journalism is done

2

The dimension of what's investigated



Since Wikileaks dynamited the global secrecy industry, investigative journalism has been immersed in data fever. It's now possible to track corruption across several continents, detect businesses and individuals seeking to evade taxes worldwide, and understand international movements of organised crime. In September 2011, Australian journalist Gerard Ryle, from the International Consortium of Investigative Journalists (ICIJ), received a hard-drive with 2.5 million files. Two computer engineers turned this tide into a reliable database. From there, an intensive reporting operation revealed more than 122,000 companies and 130,000 people operating in the shadows of the global financial system.

The discovery was praised but also raised alerts. "Investigative journalism should not be confused with what has been labelled as 'leaks journalism'", says David Kaplan. This observation is a missile to the core debate of journalism and technology working together. Are the documents obtained by hacking valid? How do we make them tell us what we really need to know? We must see spreadsheets as forensic records of reality: they offer details, but knowing the truth requires work. "The basic skills of investigative journalists," Kaplan says, "are similar to those of the most qualified prosecutors and police officers, field anthropologists and private investigators: the use of primary sources, verifying evidence, interviewing firsthand witnesses, and follow the traces of people, documents and money."

Many of the best examples of recent investigative journalism have started through better access to public information, or through building new databases using information gathered from different sources in order to answer a question that no one had asked before. These are some outstanding case studies.

INVESTIGATIONS WITH LEAKED DATA

Few processes show the peculiarities of this time as well as data leaks. Even with new forms of communicating with sources using specialised software, we face the challenge of understanding large amounts of information that could bury any enthusiasm if we didn't have the help of programmers. These are some examples.

Offshore Leaks

MEDIA OUTLET

International Consortium of Investigative Journalists (ICIJ)

It was a global effort led by the International Consortium of Investigative Journalists (ICIJ) with the collaboration of The Guardian, BBC, Le Monde, The Washington Post and 30 other media outlets. It featured the work of 112 reporters from 56 countries.

Date: April 2013

REVELATION

Politicians, aristocrats, bankers and criminals of various countries used tax havens to create companies or trusts in order to hide their assets or capital, and in many cases avoid paying taxes. Some of those listed are: President of Azerbaijan, Ilham Aliyev, and his family; Jean-Jacques Augier, treasurer of the electoral campaign of French President Francois Hollande; Spanish Baroness Carmen Thyssen-Bornemisza, who used offshore channels to buy works of art; and Maria Imelda Marcos, daughter of former Philippine dictator Ferdinand Marcos.

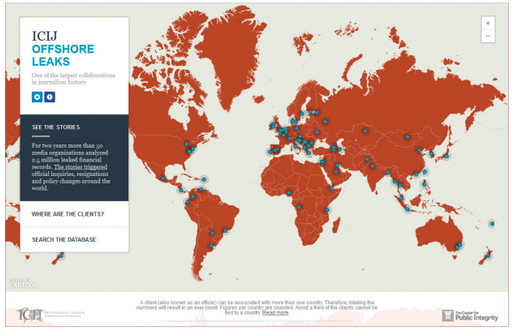
DATA ANALYSIS

The leak was 160-times larger than the number of diplomatic documents released by Wikileaks to that point. The ICIJ enlisted the help of computer experts, such as Sebastian Mondial, from Germany; Duncan Campbell and Matthew Flower, from England; Rigoberto Carvajal from Costa Rica; and Maltese Matthew Caruana. They cleaned and organised the data using Open Refine. The software dtSearch helped track the names in 260 gigabytes of data. And through the software Nuix, they found connections between keywords included in the attachments of emails, without even opening the documents. They also used the free software Talend Open Studio to integrate and organize the data in relational charts..

The programmers rebuilt the software systems of the companies that provided services to create these offshore businesses. This crucial task paved the way for the journalists to start their investigation. By offering structured files for them to navigate through, the journalists could find out who was behind the companies, who their partners were, along with any intermediaries and beneficiaries.

The hard-disk analysis detected more than 100,000 foreign companies or trusts located in places like the British Virgin Islands, Hong Kong and the Cayman Islands, among others. The documents revealed the involvement of 12,000 intermediaries, along with 130,000 people from 170 countries.

In June 2013, the ICIJ and the research unit of La Nacion newspaper of Costa Rica launched the interactive application Offshore Leaks Database, which allows you to search by name or country.



IMPACT

The case shook Europe and led to high-level resignations, such as the French economy minister, Jérôme Cahuzac, and the deputy speaker of the parliament of Mongolia, Bayartsogt Sangajav, because they were hiding bank accounts in Switzerland. Judicial investigations against officials and business people started in the Philippines, India, Greece and South Korea. Different social groups promoted campaigns against tax havens. In February 2015, the ICIJ was honoured with the George Polk Award, one of the main US journalism awards, in the Business Reporting category.

http://www.icij.org/offshore
http://offshoreleaks.icij.org/

† The Commission of the US Stock Exchange uses the Niox software to monitor emails that it confiscates from corporations when they suspect unlawful conduct.

Lux Leaks

MEDIA OUTLET

International Consortium of Investigative Journalists (ICIJ)

Another collaborative project led by ICIJ. It included 80 reporters from 26 countries.

Date: November 2014

REVELATION

More than 340 corporations, including Apple, JP Morgan, FedEx, Amazon and Pepsi have secret tax deals with Luxembourg to help them evade taxes. These deals, approved between 2002 and 2010, represent billions of dollars in lost tax revenue for states where these companies earn their profits. The agreements were signed when the current Prime Minister of Luxembourg, Jean-Claude Juncker, was finance minister of the duchy..

DATA ANALYSIS

The ICIJ accessed a 4.4 gigabyte file containing 28,000 pages of documents. For six months, 80 journalists got together using a secure communication platform called Enterprise, led by the ICIJ, to organise the content analysis. This tool allowed them to share everything, from transcripts of interviews and photos, to confidential material. "It was the closest thing to a global newsroom", said Marina Walker, assistant director of the Consortium. Each reporter reviewed very complex financial documents related to the companies in his or her country. The ICIJ also received advice from specialists in finance and taxation. At the same time, the ICIJ formed a team of reporters and computer engineers who developed a database for public exploration.

IMPACT

Since December 2014, the European Commission has been investigating whether the practices of Luxembourg constitute a tax system tailored to the needs of large corporations, to the detriment of community law.
The ICIJ received the George Polk Award in the Business Reporting category and the Data Journalism Award in 2015 for Best Investigation.

http://www.icij.org/project/luxembourg-leaks

Swiss Leaks

MEDIA OUTLET

ICIJ, The Guardian, CBS, Le Monde, Süddeutsche Zeitung and The Washington Post

This was the third ICIJ international collaboration project, done with The Guardian, CBS, Le Monde, Süddeutsche Zeitung and The Washington Post. It included 154 journalists from 47 countries.

Date: February 2015

REVELATION

Between 2005 and 2007, the Swiss subsidiary of HSBC bank established a system of tax evasion that helped more than 100,000 wealthy customers in 203 countries hide their money to avoid paying taxes. The list of beneficiaries included members of the royalty, politicians, celebrities, drug traffickers and businessmen, with accounts that collectively held more than $100 billion.

DATA ANALYSIS

It was based on records of secret bank accounts of HSBC customers, stolen by a former employee of its Swiss subsidiary, Hervé Falciani. The hard-drive contained 2.5 million records. In early 2014, the French daily Le Monde was given access to the data and passed it to the ICIJ to create an investigation strategy. The first step was to recreate the HSBC customer database based on the Excel files available. Then, they established relationships between names and countries, followed by the use of the software Talend to transfer the original database to the graphic database Neo4j that allows organizing the connections between the data. Linkurious was the tool that facilitated the visualization and analysis process. The reporters communicated via the platform Voyager -created with an open source software called Oxwall-, that allows to open thematic forums and to share files. The database led the team to create a diagram with 275,000 nodes and 400,000 relationships between them.

IMPACT

In February 2015, the Swiss prosecutor opened a criminal investigation against HSBC for aggravated money laundering. In the UK, the tax collection agency HMRC recovered $236 million from some of the 3,600 Britons identified as users of the HSBC branch in Geneva, but only started one prosecution. France initiated 103 legal proceedings against an equal number of people.
In June 2015, the ICIJ received the Data Journalism Award for Best Research of the Year.

http://www.icij.org/project/swiss-leaks

INVESTIGATIONS WITH PUBLIC DATA

90 countries in the world – including 14 in Latin America and the Caribbean - have Freedom Of Information laws, according to the Regional Alliance for Free Expression and Information. This presents an advantage for investigative journalism: it is now possible to track how corporate power influences state decisions and the impact this has on people's lives.

Evicted and Abandoned:

MEDIA OUTLET

The Huffington Post, The Guardian, The Ground Truth Project, and The Investigative Fund

Collaborative work that involved 50 journalists from 21 countries.

Date: April 2015

REVELATION

More than 3 million vulnerable people were displaced from the areas where they lived because of around 1,000 projects financed by the World Bank in 124 countries between 2004 and 2013.

DATA ANALYSIS

In early 2014, the American journalist Sasha Chavkin noted that the reports of the World Bank's Ombudsman, which oversees the agency's activities, contained dozens of complaints from people and communities displaced by projects financed by the bank in developing countries. Chavkin downloaded more than 6,600 documents from the World Bank to build a picture of the projects, the beneficiaries of loans and the complaints. The information about the cases between 2004 and 2013 was incomplete, so to check the data, she sought sources within the organisation, including former officials and experts. The data identified a pattern: the World Bank and the International Finance Corporation - its investment arm in the private sector - did not respect its own policy to protect those who may be adversely affected by the projects it finances. It even gave loans to governments and companies accused of violating human rights. The reporters who travelled to countries such as South Sudan, Ethiopia, Guatemala and Peru communicated through the platform Odyssey.

http://www.icij.org/project/world-bank

IMPACT

The World Bank announced a plan to improve the monitoring of development projects in order to prevent the bad practices that cause displacement.
The Online News Association recognized this research at the Online Journalism Awards in the Innovative Investigative Journalism category.

Congress members back legislation that could benefit themselves, relatives

MEDIA OUTLET

The Washington Post (EE.UU.)

Date: October 2012

REVELATION

Seventy three US lawmakers approved laws that affected their investments or benefitted their relatives because they were not obliged to disclose potential conflicts of interest.

DATA ANALYSIS

The team put together a database in Excel with the financial information and other public records of the 535 members of the US Senate. They compared the personal investments of the members with reports of their activities monitored by LegiStorm, a non-profit watchdog group. The information was also crossed with reports the Office of Management and Budget at the White House. One legislator was found to have facilitated the approval of tax exemptions for horse owners and then bought seven race horses. Another one sponsored a bill that benefited the natural gas company in which his wife was a shareholder.

IMPACT

Congress started a debate about its ethical code of conduct and opened investigations into senators with clear conflicts of interest.

https://goo.gl/9WbZbT

Missed signs, fatal consequences

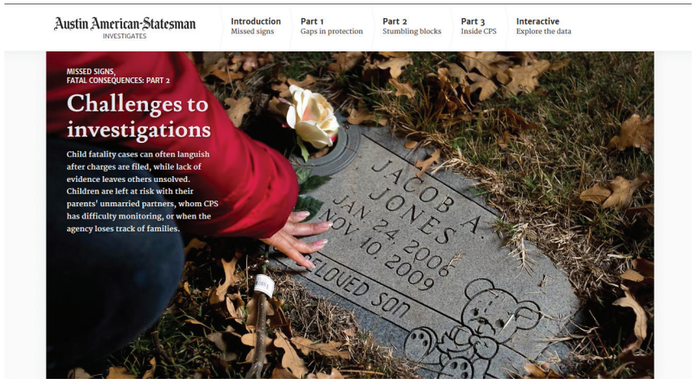
MEDIA OUTLET

Austin American-Statesman (EE.UU.)

Date: January 2015

REVELATION

Between 2010 and 2014, the supervising system of the Department of Family and Child Protective Services in Texas failed, which led to the death of 655 children by their relatives or caregivers. The officials did not take the necessary actions to protect them.



DATA ANALYSIS

The project began after obtaining 779 reports of deaths of children from violence in their homes. The documents, requested FOI were only available in PDF. The journalists had to transfer the contents to a format they could analyze, in which they created several fields to fill and order the information. They used Caspian, a very friendly online service to manage databases, which is based on Microsoft SQL Server system. With this tool the reporters could establish that the employees of the child protective services had visited the victims several times, but did not take the risk signals into account. The six-month work included the reconstruction of the events with interviews, collating documents and visits to the affected households.

IMPACT

The Department of Family and Child Protective Services of Texas reformulated the research system of child abuse and opened an investigation into 50 employees for various crimes.
The website Austin American-Statesman won the Online Journalism Awards 2015 prize, awarded by the University of Florida in the category Investigative Data Journalism.

http://projects.statesman.com/news/cps-missed-

The shoemaker's son always goes barefoot

En casa de herrero, cuchillo de palo

MEDIA OUTLET

La Nación (Costa Rica)

Date: April 2012

REVELATION

Half of the ministers of the president of Costa Rica, Laura Chinchilla, sub valued their properties to pay less taxes between 2009 and 2010.

DATA ANALYSIS

The team, made up of two reporters and two computer engineers, organized a database in Excel with information on the properties of the ministers contained in the public records, the values they had declared on the tax forms sent to municipalities where their properties were located and the value of the properties according to the Ministry of Finance. The first two databases of calculation were Google, a tool that organized the data to be displayed in a very attractive way.

IMPACT

The values of the properties of all public officials involved were updated. In 2013, the Institute for Press and Society (IPYS) gave the team the Latin American Investigative Journalism Award for good management of information technology to enhance the reporting in monitoring governance.

http://goo.gl/16ndl8

6 INSTRUMENTS FOR TO IMPROVE OUR STORYTELLING

[How to enrich a story with charts, infographics and sounding sentences]

1. Datawrapper:

https://datawrapper.de/

Allows to select data from a spreadsheet and to convert it into graphics and explanatory maps with color types and custom fonts. Includes the option of making pie charts, fevers or bars.

2. TimelineJS:

http://timeline.knightlab.com/

It is used to create interactive timelines, with photos, videos, and links. User doesn't need to open an account. Just insert dates, text and URLs in a Google spreadsheet and the tool will organize it for display in a very attractive way.

3. Infogr.am:

https://infogr.am/

It makes online info graphics. It offers templates that display data bars, circles, fevers. You can insert the information into its format or import files in Excel or .csv. It has a free version and a paid premium one. Graphics can be shared through Facebook, Twitter and Pinterest.

4. Tableau Public:

https://public.tableau.com/

It converts data from a spread sheet to interactive graphics (maps, tables, bars) and creates filters so that users can make consultations and have personalized results. There is no need to know programming to use this tool.

5. StoryMapJS:

http://storymap.knightlab.com/

It creates stories based on locations identified on a map. It allows inserting videos, tweets, texts or images that are displayed as a gallery associated with each selected site. The information is also inserted into a Google spread sheet.

6. Soundcite:

https://soundcite.knightlab.com/

It places sounds to a word or phrase in a text. Once the mp3 file is uploaded in SoundCloud, then you have to select it from Soundcite and the desired fragment is adjusted. There will be a code to insert in the selected phrase, which enriches the experience of reading with sound.

INVESTIGATIONS WITH BUILT DATABASES

In recent years, big cases started in investigation projects with data bases build by the journalists themselves. Here we put together seventeen experiences that can inspire the development of new topics.

Who's Behind the Financial Meltdown?

MEDIA OUTLET

The Center for Public Integrity (EE.UU.)

Date: May 2009

REVELATION

The US corporations Lehman Brothers, Merrill Lynch, JP Morgan & Co., Citigroup, Goldman Sachs & Co and Swiss bank Credit Suisse First Boston were part of the business model which generated the so-called housing bubble that broke the financial system. They were the owners of 21 out of the top 25 companies in the subprime mortgage industry that had granted high risk housing loans that triggered the economic crisis in 2008, and then benefited from the bailout.

DATA ANALYSIS

The CPI data editor, David Donald, began an analysis of 350 million mortgage applications approved from 1994 to 2007. The information had been previously collected from the public documents of the Home Mortgage Disclosure Act (the lending registration system based on the Mortgage Disclosure Act) and organized in spread sheets by the National Institute for Computer-Assisted Reporting (NICAR). The analysis identified that most high risk loans totalling more than a billion dollars were granted between 2005 and 2007. This allowed the journalists to refine the search and define who the main contractors were. A team of reporters wrote the profiles of the lenders and the report also included information on the contributions of the companies involved to members of the US Congress. To display the location of each subprime housing loan, they used heat maps made with the Palantir Government Software, which offers a powerful tool of visual analysis used in both academic projects and disaster management or intelligence programs.

http://goo.gl/cVX1ee

IMPACT

The series impulsed the US Congress to establish a commission for investigating the case. The team concluded that the disaster was avoidable and that the crisis was the result of failures in the regulations, corporate mismanagement and irresponsible risk of Wall Street.

The secret diaries of Parana

Diários Secretos

MEDIA OUTLET

Gazeta do Povo en colaboración con RPCTV Paraná (Brasil)

Date: Marzo 2010

REVELATION

Between 2006 and 2009, the Legislative Assembly of Parana (Brazil) hid a scheme to divert public funds that involved hiring ghost employees, overprized services, nepotism cases and other crimes. Embezzlement would have cost 400 million dollars.

DATA ANALYSIS

Journalists James Alberti, Katia Brembatti, Karlos Kohlbach and Gabriel Tabatcheik collected the legal gazettes of the Legislative Assembly of Paraná published between 1998 and 2009, several of which were not available in the archives of the Assembly itself. They hand digitized content of 724 legal newsletters to build a database in Excel that shows the staff hired and the managing of the budgets outlined in the gazettes. They were able to identify some twenty ghost employees, including dead people or children. The news team also found that relatives had been used by legislators and children of judges. The data analysis was completed with testimonies and documentary sources. Finally, they shared all the official gazettes in a public search engine, including the ones missing in the Assembly archives.

IMPACT

Since the publication until March 2015, fourteen employees and former employees of the Legislative Assembly of Paraná were sentenced to prison for embezzlement and other crimes.
"The secrets diaries of Paraná" won the Global Shining Light Award and the Latin American Investigative Journalism Award in 2011.

http://goo.gl/0qXwNF

The faces of homelessness

Los rostros del desamparo

MEDIA OUTLET

La Nación (Costa Rica)

Date: February 2011

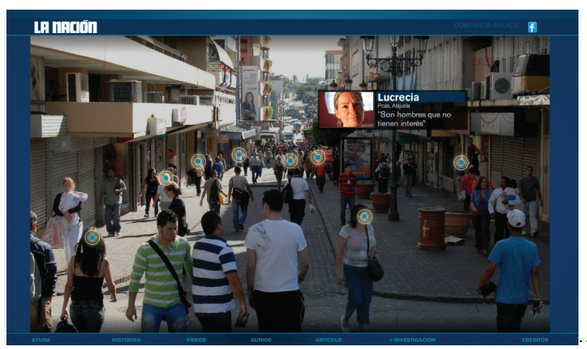
REVELATION

The official education subsidies programme 'Avancemos' is supporting the studies of teenagers abandoned by parents with earning high salaries.

http://goo.gl/BfMkQC

DATA ANALYSIS

This investigative series started with the hypothesis that there may be abuses when granting the scholarship benefit 'Avancemos', a grant programme to encourage more than 167,000 young people to continue their studies. The team, led by Giannina Segnini, could access the database of beneficiaries, so she completed it with the names of their parents, and crossed it with information that detailed their income and assets. A first check revealed that 75 beneficiaries had parents with salaries of between $2,000 and $9,000. The story took a turn, however, when reporters searched for the beneficiaries in person. It turned out they were children of people with a more than comfortable financial situation, but had been abandoned and were now living in poverty with a relative. The investigation took three months and resulted in a new, even more revealing story than the initial hypothesis: the state was subsidising the education of homeless youngsters whose parents were living in a good economic situation. This project included a team of three developers, three designers and four journalists. It was also accompanied with a special, multimedia report.



IMPACT

The research received a Special Mention at the II Regional Award for Journalism, Poverty and Human Rights in Central America in 2011.

Doctors without control: owners of public health in Chile

Médicos sin control: los dueños de la salud pública en Chile

MEDIA OUTLET

CIPER (Centro de Investigación Periodística)

Date: September 2010

REVELATION

A one-year track of five hospitals in Chile showed that doctors' attendance was not monitored, nor was their use of health services to benefit their personal interests, to the detriment of the patients.

DATA ANALYSIS

In late July 2009, CIPER requested the attendance records of doctors in five hospitals in Santiago. To obtain these records, they had to overcome several obstacles, because some hospitals refused to provide the information and others provided incomplete data. The team finally accessed more than 35,000 attendance records. After putting the information into a spreadsheet, they crossed it with the professionals' schedules in the clinics and in their private practices. At the same time, they visited the selected hospital to see if doctors were sticking to the schedules. They found that several doctors used public infrastructure to serve their private patients. It was also obvious that they did not comply with the working hours stipulated in their contracts.

IMPACT

Seven months after the publication of this research, the General Comptroller of Chile issued a report that confirmed the non-compliance with contracts of doctors in 13 hospitals.
The investigative series won an honourable mention in the Latin American Investigative Journalism Award in 2011.

http://ciperchile.cl/multimedia/medicos-son-control/

Dollars for docs

MEDIA OUTLET

ProPublica, in partnership with the Boston Globe, Consumer Reports, NPR, Chicago Tribune and Public Broadcasting Service (PBS).

Date: October 2010

REVELATION

Between 2009 and 2010, seven pharmaceutical companies made individual payments of more than $100,000 to 17,000 doctors so they would promote and prescribe their drugs in the US.

DATA ANALYSIS

The seven companies had published this information on their websites, but it was difficult to analyse because it was hosted in PDF and JPG formats. Two journalists and a developer downloaded the data and organised it into an Excel file that was broken down into categories, such as consulting, meals, travel and gifts. For the first time, the payments made to physicians by 36% of the US pharmaceuticals industry was visible. The team used the tool Google Refine to clean and standardise the names of the physicians benefiting from this money. Then they crossed this information with public databases that record licences to practice medicine and disciplinary records of physicians. ProPublica presented this research together with an app that allows anyone to search a doctor by name and discover if he or she received payments to promote certain products. Three years later, the database was updated and it showed that payments of $4 billion were made to 681,432 physicians, paid by 1,630 pharmaceutical companies or medical products manufacturers.

IMPACT

In 2012, the US Congress passed a law making it mandatory to publish gifts given and payments made to physicians in the United States. More than 125 news organizations, such as the Boston Globe and the Chicago Tribune, made investigations using this research tool.

https://projects.propublica.org/docdollars/

Terrorists for the FBI

MEDIA OUTLET

Mother Jones (EE.UU.)

Date: August 2011

REVELATION

FBI informants fabricated evidence to incriminate suspects accused of threatening US security in order to claim a reward of $100,000 offered as part of the war on terror.

DATA ANALYSIS

Journalist Trevor Aaronson examined more than 500 cases of people accused of terror charges and found that in nearly half of these cases, there was an FBI informant involved. Aaronson combined data extracted from court records from different states and FBI documents, with interviews with the agents and the lawyers of the accused. He spent months working with an assistant to build a database. Initially he used Excel and MySQL manager, which help him build relational databases. Then the team used the Drupal content management framework to build an online search tool.

IMPACT

The FBI investigated the officers accused of making up these cases, while some of the informants involved have been subject to legal proceedings.
The research won a Data Journalism Award 2012 in the Data-Driven Investigation category.

http://goo.gl/hKem0p

Methadone and the politics of pain

MEDIA OUTLET

The Seattle Times (EE.UU.)

Date: December 2011

REVELATION

The Medicaid programme, aimed at people on low incomes, gave the narcotic methadone to patients in order to cut drug purchase costs without taking into account the health damage caused. As a result, 2,173 people died between 2003 and 2011.

DATA ANALYSIS

Following an alert raised by several doctors, two journalists looked for all the available records in Washington to track the number and circumstances of deaths linked to the drug. Through FOI requests, they found four databases: the records of death certificates, the forensic notes of each doctor, the clinical profiles of the patients and the costs of their methadone treatment in Washington hospitals. In parallel, they collected more data about the socioeconomic profiles of the deceased. These files were reviewed and annotated using DocumentCloud. This information was then reformatted in Excel and they then used Access software for handling the large volume of information. ArcGIS was then used to make maps. To present the findings they used Google Fusion Tables, Tableau, Final Cut Pro and Adobe.

IMPACT

Washington authorities sent an alert to more than 1,000 pharmacists and 17,000 health professionals about the risks of methadone. In January 2012, they sponsored a programme that instructed doctors to only use methandone as the last resort.
The investigation received the 2012 Pulitzer Prize in the Investigative Journalism category. That same year it was awarded the Data Journalism Award.

http://goo.gl/YoQcDs

Expenditures Senate

Gastos del Senado

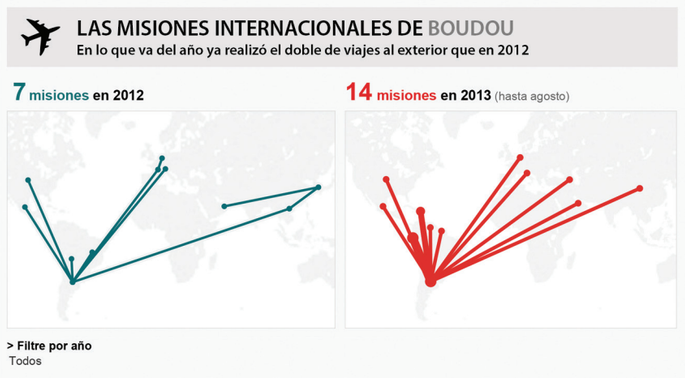
MEDIA OUTLET

La Nación (Argentina)

Date: February 2013

REVELATION

The team found recruitment, travel and other irregular expenses in the Argentinean Senate totalling more than $1 million between 2010 and 2012. The Vice President of the Senate, Amado Boudou, even bought luxury furniture for his office and made travel expenses claims for trips that overlapped.



DATA ANALYSIS

An anonymous informant sent an email to newspaper La Nacion with a photograph of the Vice President's office containing a luxurious table imported from Italy. The email prompted the journalistic team to download 33,000 documents from the Senate's official website. OmniPage 18 software was used to convert the PDF files into trackable ones. The information was then added to Excel, including furniture, travel and security personnel expenses, among others. Tableau Public was used to explore the data and make interactive graphics. To manage one set of the PDF files, the team developed the collaborative platform VozData, which allowed volunteers to classify the information into prescribed formats.

IMPACT

The Prosecutor launched an investigation of the travel expenses of the Senate's vice president. The issue was discussed on television, radio and other newspapers.
The work won the Data Journalism Award 2013 in the category Data-Driven Investigations Big Media, organized by the Global Editors Network (GEN).

http://www.lanacion.com.ar/gastos-en-el-senado-t49163

THE MIGHTY NEO4J

[Or how to find a global fraud ring using just circles and lines]

Mar Cabra, editor of the ICIJ's research and data unit, learned to use **spreadsheets as sophisticated software to keep track of fiscal and corporate corruption** across millions of files that apparently had no direct connection. For the Swiss Leaks series, she used a tool called **Neo4j**, which identifies **connections between large amounts of data** and display them in graphics. "The connections were crucial to identify who did business with who," Cabra said during the Global Investigative Journalism Conference held in October 2015 in Lillehammer, Norway. **Instead of tables, this tool uses nodes and edges, which makes reading data relationships more intuitive.**

This feature allowed the team of **OjoPúblico** to unravel the trail of the political campaign financing during the 2016 Peruvian elections, by crossing 16 databases and analysing 3 million records. "This system allowed us to understand, analyse and simultaneously cross-check databases", says Nelly Luna, the journalist who was in charge of the investigation. **What once would have taken several years, just took six months now using this tool**.

Homes for the Taking: Liens, losses and profiteers

MEDIA OUTLET

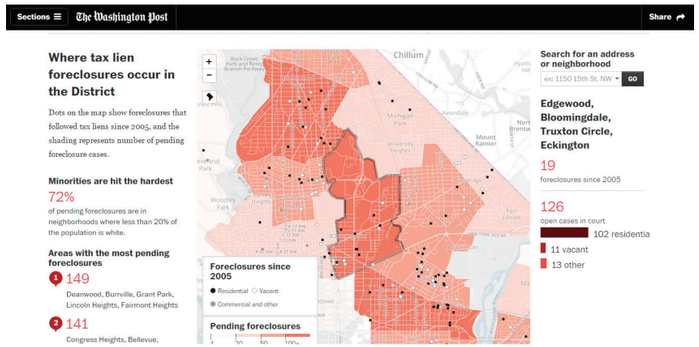
The Washington Post (EE.UU.)

Date: September 2013

REVELATION

About 200 people from Washington DC, mostly elderly, lost their homes in irregular mortgage foreclosures because of tax debts of less than $1,000. This investigation showed the abuses within the tax-withholding programme that allowed them to keep the properties.

http://goo.gl/ZGtrUP



DATA ANALYSIS

The research was based on the analysis of foreclosure documents between 2005 and 2013 available in the Office of Tax and Revenue, the Supreme Court and the Office of Tax and Revenue. The team put together a database of 200 elderly homeowners who lost their homes in auctions imposed because of the delay of payments under $1,000. In most cases, the properties were sold to real estate companies even though the original owners paid their overdue debts. When the reporters found the victims, they discovered more evidence of abuses. One of those affected was dying of cancer and owed $1,025 in taxes. Another was 95 years old and suffered from Alzheimer's disease and forgot to cancel $44. The documents were examined with DocumentCloud and the data was analysed using spreadsheets. The findings were worked up with the MapBox platform and the open-source library Leafletj, which allows you to make mobile-friendly maps.

IMPACT

A dozen senators asked the government to investigate the tax programs in order to protect vulnerable homeowners from losing their property due to small tax debts.
The investigation won the Research Data Journalism Award 2014 in the category Data-Driven Investigation.

Breathless and Burdened

MEDIA OUTLET

Center for Public Integrity and research unit and ABC News (USA)

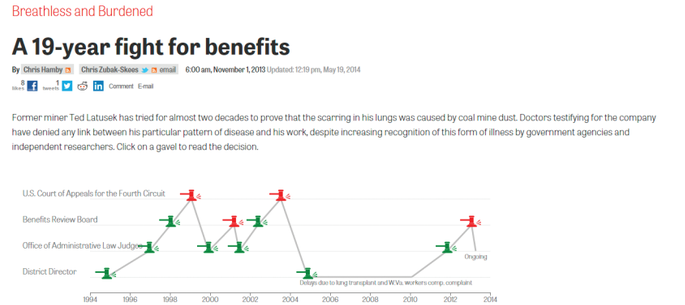
Date: October 2013

REVELATION

Doctors and lawyers working for the coal industry have helped divert aid meant for sick and dying miners with lung diseases.

DATA ANALYSIS

Reporter Chris Hamby built a database with information from x-ray examinations performed on 1,500 miners at Johns Hopkins Hospital between 2000 and 2013. The coal mining companies denied health treatment and social benefit payments because one doctor had signed a diagnosis that said they didn't have black lung disease. The miners took their cases to court. Meanwhile, the journalist used the legal records to build a second database about the legal strategies of the mining companies, reports of other doctors and the judges' verdicts. Hamby spent months reading and processing the information in his spreadsheets. By doing this, he was able to identify cases and patterns which showed the system was mounted to deny benefits to the miners.



IMPACT

Johns Hopkins Hospital suspended and then removed the physician responsible for the black lung programme. US Senators also then wrote legislation to reform the benefits system for coal miners.
The research won the Pulitzer Prize 2014 in the category of Iinvestigative Journalism category.

http://goo.gl/ZPZ9HE

DIGITAL MAPPING

[Resources to help you locate events and characters in exact places]

OpenStreetMap

http://www.openstreetmap.org/

A collaborative project to create free, editable maps. You can use geographic data captured using mobile GPS devices, on top of other sources.

My Maps

https://www.google.com/maps/d/u/0/

A tool to create maps through Google Maps. It's easy to use, share and embed the maps using the code generated. The only requirement is a Google account.

MapBox

http://mapbox.com/tour/

A website that lets you create custom maps easily. Uses free software.

InfoAmazonia

http://infoamazonia.org/

A website that hosts environmental maps of the nine countries of the Amazon.

Geocommons

http://geocommons.com/

Free programme to create maps with multiple layers. It allows you to use geographic location information added by other people and share your work.

Medicare Unmasked

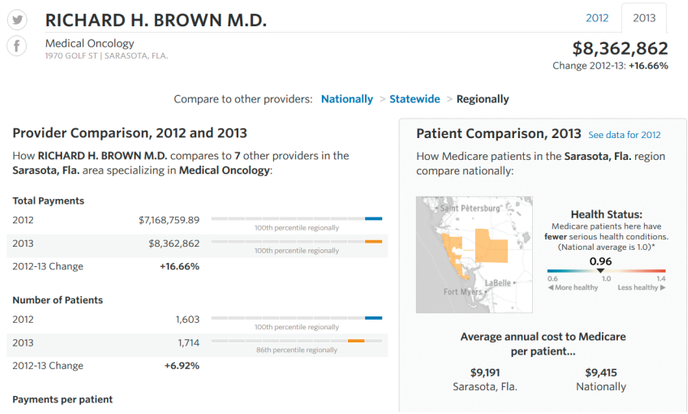
MEDIA OUTLET

The Wall Street Journal (EE.UU.)

Date: April 2014

REVELATION

Medicare, the health programme that helps people aged over 65 and young people with serious illnesses, made payments of $60 billion annually to more than 880,000 physicians, ambulance services and laboratories, several of which were cases of fraud, waste and abuse.



DATA ANALYSIS

In May 2014, after a five-year court battle, the WSJ managed to access nearly 10 million records of Medicare contracts and payments to medical providers, kept secret since 1979. They also got access to a second database purchased from the agency CMS (Centers for Medicare and Medicaid Services), which included records of payment claims of suppliers over a 6 year period. The crossing of information in a database was used to detect cases of fraud, excessive costs and abuse in a programme that spends more than $60 billion annually. Journalists and data experts used the programming language C# to convert records into relational tables and developed algorithms that could make connections between data. Then they imported the data from the information manager Microsoft SQL. With this, they produced interactive charts, rankings and a search platform of the payments given to each physician

IMPACT

Winner of the 2015 Pulitzer in the Investigative Journalism category.

http://graphics.wsj.com/medicare-billing/

ARESEP increases the price of petrol and diesel to make asphalt and gas cheaper

Aresep encarece gasolina y diésel para abaratar asfalto y gas

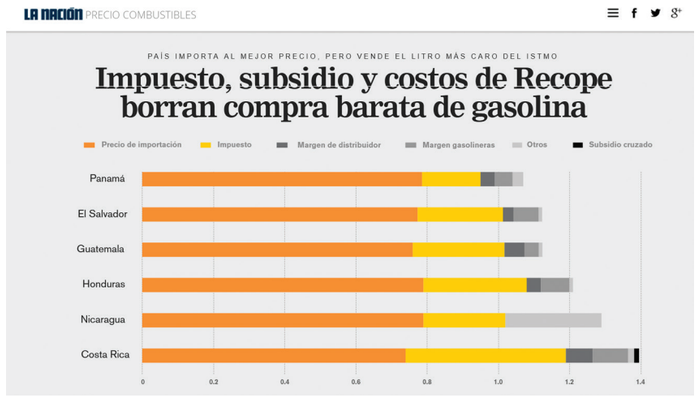
MEDIA OUTLET

La Nación (Costa Rica)

Date: December 2014

REVELATION

In 2008, the Regulatory Authority of Public Services of Costa Rica (ARESEP) secretly approved a cost overrun in the formula of diesel and gasoline prices in order to lower gas and asphalt costs. The subsidy benefited the cement companies to the detriment of thousands of consumers..



DATA ANALYSIS

The journalists Hassel Fallas and Mercedes Agüero manually created an Excel database with information from 59 resolutions on prices -both ordinary and extraordinary - issued by ARESEP between June 2009 and September 2014. The components of the pricing formula were disaggregated to find out how the operating costs of the Costa Rican Oil Refinery (Recope, in Spanish) were included in the fuel rate. With the help of specialists, it was found that ARESEP had established a cross-subsidy: it had assigned a higher price to a product than its real cost, in order to be able to reduce the price of another one. In this case, the cost of diesel and gasoline increased to lower asphalt and gas. The team had to build five versions of the database until they found the right one to prove this.

IMPACT

ARESEP modified the formula to calculate the price of fuel in order to eliminate the hidden costs in the methodology used since 2008..
The research was a finalist in the Data Journalism Awards 2015, GEN, in the Best Investigation of the Year category.

http://goo.gl/aNriTd

Courting Favor

MEDIA OUTLET

The New York Times (EE.UU.)

Date: October 2014

REVELATION

More than 20 corporations, including DirecTV, Pfizer, Coca Cola, Google and Citigroup, gave gifts and contributions to attorneys general in 12 states in order to influence their decisions.

DATA ANALYSIS

During a nine-month investigation, The New York Times reporter Eric Lipton saw how corporate lobbying had penetrated the attorneys general to twist their decisions in the interests of 21 companies.
Through legal petitions, Lipton managed to access about 8,000 emails from the public servants' email accounts and made a database in a spreadsheet with information extracted from their correspondence, in which the relations of officials and businesses were evident. Then, he documented that some prosecutors investigating corporations had received gifts and other contributions. That information was supplemented with photographs, as well as evidence gathered by the reporter attending academic conferences sponsored by such companies.
To complete his report, Lipman showed that the contributions paid by companies to Attorneys General Associations -both Democrat and Republican- quadrupled in four years.

IMPACT

The publication of this data generated investigations in four states and the Senate proposed a bill that prohibits attorneys general receiving gifts or financial contributions.
Courting Favor was awarded the 2015 Pulitzer Prize in the category of Investigative Journalism and the IRE Award in the category Print/Online Large.

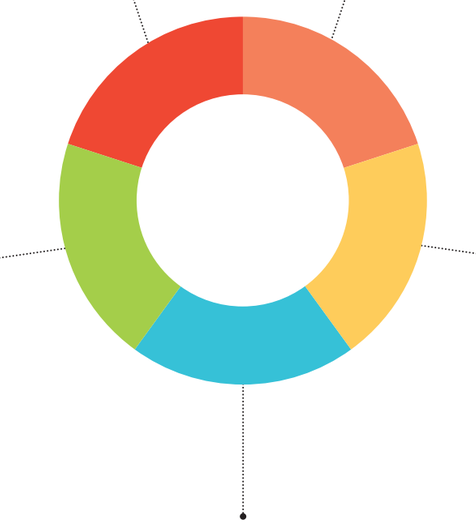
http://goo.gl/oKMMKg

THE ROAD TO A CULTURE OF INNOVATION

Digital laboratories of investigative journalism in Peru

3

The data community in Peru



OjoPúblico
Innovation lab focused on investigative journalism and new digital narratives.
http://ojo-publico.com/

HacksHackers Lima
Peruvian chapter of the international community that promotes partnerships between journalists and programmers.
http://hhlima.info/

Open Data Perú
Community that promotes openness and reuse of open data in Peru, and the creation of civic apps.
http://odpe.org/

HackSpace
Space for creativity and digital innovation aimed at academics, researchers and citizens interested in technology.
http://hackspace.pe/

Convoca.pe
Digital investigative journalism site that gathers reporters, data analysts and developers.
http:/convoca.pe/

Before journalism became friends with statistics, journalist Liz Mineo found a way to track corruption using numerical tables. In late 1997, while Peru was shaken by frequent allegations of government mismanagement, Mineo focused her attention on the public works planned to prevent the impact of El Niño, a cyclical process of climate change that often triggers disasters in different parts of the country. The reporter asked a basic question: how was the advertised public budget of S/.100 million (Peruvian Sol) used? The problem was that in those times, there was no government transparency. There wasn't an FOI Act either. In fact, the National Institute of Civil Defence (Indeci), the agency responsible for administering the money and making contracts, was controlled by the military and linked to the dreaded presidential adviser, Vladimiro Montesinos, who had created an almost police state. When she started making her initial inquiries, Mineo found that any information on public works was confidential and was not available to be released. So, she embarked on a pioneering, forensic analysis of the information: building a database to look for patterns, which ultimately helped her discover hidden information.

At the first stage, the reporter from El Comercio looked for internal sources within the government. She persevered until she found an official from Indeci who agreed to collaborate: he gave her an 80-page document with information about the public works and the contractors. It was the best he could get. "I was afraid he could be discouraged if I requested the information on a disk," Mineo remembers. The second step was to transfer all the information to an Excel table, which included 293 public works projects in 21 areas of Peru,

involving S/.100 million and 61 companies. “There were too many numbers to make calculations, by hand or with a calculator”, she recalls.[1]

The result was a host of irregularities: the first leak allowed her to see that many of the alleged prevention works had been carried out in departments that did not fall inside El Niño’s area. When she found out the names of the shareholders of the companies benefiting from the contracts, Mineo identified that 12 out of the 61 companies were owned by soldiers who were former comrades of the chief of Indeci, General Homero Nureña. In addition, one of the companies appeared to be owned by his private secretary and another one by a nephew. A third of the companies had been created just months before receiving the contracts and some even after they have obtained them.

The third stage was verification. Mineo, with the support of the newspaper’s correspondents, visited the places where the investments had been made and found unfinished works, while some of them were non-existent. She also found out that general Nureña had put money directly into Cajamarca, his hometown, although the area did not suffer the ravages of El Niño. General Nureña had even ordered the building of an elementary school that was named after his mother. If Mineo hadn’t done the journey in person, perhaps the findings wouldn’t have been so stark.

Mineo’s experience marked a change in the operational capabilities of journalists. In those times, the government was the only one with access to public information, so Liz Mineo had to combine traditional journalistic techniques with emerging resources, using a data-driven mentality in order to form a conclusive investigation. The case became a series of 15 articles in El Comercio. Later, General Nureña was sentenced to prison for embezzlement.

1 Personal interview with Liz Mineo.

8

Latin American news media have journalistic teams integrated with hackers to analyse massive data. In Peru, **OjoPúblico** and Convoca work together with programmers

For a long time, this pioneering experience remained an isolated case in the Peruvian media landscape. Replicating it was unlikely in a country with such little attachment to reality that some people even challenged the date of the president’s birthday.[2] At least until the end of the 20th century, the Peruvian public archives suffered from ‘amnesia’ and the government agencies managed their budgets more like necromancy (communicating with the dead!) than accounting. Investigative reporters were concentrated on tracking drug trafficking, terrorism or corruption, but they had to discover these networks through direct sources.

Only since 2001, with the publication of FOI legislation, did various state agencies begin to digitise their data and publish it online. Still, the informality of the mechanisms used to gather and update information means there is a lot of unreliable data. A dubious sample was detected in 2008 by journalists Gustavo Gorriti and Romina Mella, IDL Reporters (IDL-R) while they were looking into the records of crimes in Metropolitan Lima police stations.[3] When they requested information on the number and type of crimes in different jurisdictions, they realised almost all the results were the same. When they dug a little deeper, they found that the reports were filled with a simple copy and paste, under the premise that the situation was similar everywhere. So the reporters had to find another way to investigate the problem of criminality.

Until the first decade of the 21st century, the existence and use of databases in Peruvian journalism was based on individual efforts of investigative reporters trying to move forward in their investigations. In 2010, journalist Milagros Salazar Herrera, from IDL-Reporters, investigated the powerful fishing industry in Peru with digital tools that allowed her to compile, verify and analyse the spreadsheets of more than 47,000 reports of anchovy landing, be-

2 VALENZUELA, Cecilia. “Buscando la cuna de Fujimori”. From Caretas magazine [Lima], Link: http://www.caretas.com.pe/1475/fujimori/fujimori.htm. [Visualized on Noviembre 25, 2015].
3 Quote taken from Romina Mella’s speech at the Chicas Poderosas Perú Conference, November 20, 2015.

cause this was a species in permanent risk of over-exploitation. The documents correspond to the volumes reported by the fishing companies, plus the relevant reports recorded by state supervisors of the country, between 2009 and 2010. The detailed cross-checking of the data discovered a failing audit system that benefited the world's second-largest fishing industry. There was an undeclared volume that would have generated $100 million in taxes.

Salazar's report for IDL-R was later on expanded and included in a coordinated global coverage of the International Consortium of Investigative Journalists (ICIJ). In 2012, this work was also one of the winners of the Latin American Investigative Journalism Award, presented by the Institute for Press and Society (IPYS).

That same year, the Research Unit of the newspaper El Comercio published a series of reports about the businesses and family groups that benefited the most from buying in the millionaire Food Assistance Program (Pronaa) that later on was closed and replaced by the Qaliwarma program. A data base was manually built including the program's recruitments in the last ten years, based on the reports of the Electronic System of Procurement (SEACE). This way they could identify the companies that received more contracts, their owners and their background.

Shortly after, the same method was applied to investigate the state's purchases of drugs and it was found that a pharmaceutical monopoly was charging excessive prices to the public health system. The star source was a spreadsheet.

It was the beginning of a new phase for journalism, of exploring new tools to improve the analysis capacity and go from reporting on a particular event to a broader investigation, based on massive data. The process would have implications beyond the research results: from the focus of the investigation to the metalanguage of the profession itself.

"Investigative journalism is more alive than ever. Its capacity to support itself using technology has managed to increase the quality and impact of the stories."

David Kaplan, *Director of the Global Investigative Journalism Network*

JOURNALISM + TECHNOLOGY

One morning in June 2014, on the premises of a technology institute in Lima, more than 50 journalists and programmers gathered for the first time to create tools that enabled a leap in the methods for obtaining and processing information. The activity was a hackathon that lasted 12 continuous hours and it was baptised with the canonical expression that has guided the best journalism: "The money trail". It was an exercise in putting civic and journalistic eyes on the use of public funds. The meeting was held simultaneously with groups in 12 other Latin American cities, all of them belonging to the Hacks/Hackers community, which links journalists and programmers willing to reinvent the media.

The expectation was fuelled by recent media revelations, with technological support that were having high impact in the municipal and regional elections of that year. An important case was an unprecedented partnership between the NGO Transparencia, a hacker and two journalists from the popular news website Utero.pe. The result was called Verita, a software that crossed the information of the resumes of more than a hundred thousand candidates with several public databases, such as convictions for civil and criminal offenses. The findings were surprising: 1,395 candidates for mayors and regional governors had convictions; more than half because of not paying alimony to their own children.

The project confirmed a new trend for journalism in the era of big data: collaboration among experts from different fields of knowledge. What's special about this project was that its protagonists, journalists Marco Sifuentes and Ernesto Cabral, and the director of the NGO Transparency, Gerardo Távara never met face-to-face with the hacker that helped them. The architect of the massive download and the crossing of all the information did not live in Lima. He was a young Peruvian biologist studying a PhD in Europe and spent his free time programming. At the time, he preferred to be identified under the pseudonym Aniversario Perú (Anniversary Peru).

TWO DEVELOPERS SUPPORT JOURNALISM

Behind the pseudonym **Anniversary Peru** there is a 35-year-old biologist and father of two girls studying for a PhD in Europe. **In his spare time, he is a civic hacker**. He defines himself as: "A guy who knows some programming and wants to make information accessible to people." The architect of the software 'Verita', used to scan the resumes of the elections candidates, and 'Manolo', which extracts information from the records of official visits to state entities, believes that **good journalism is like science**. "In biology or physics you have to prove an idea and you do it by searching and analysing data to reach a conclusion. It is also what makes a good investigative reporter," he says.

Hacker **Antonio Cucho** has a similar experience. In 2014, he founded Open Data Peru, one of the main communities that promotes the release of public interest information and its conversion into information tools. Its mission is not easy in a not-very-transparent country, but in the last two years, Cucho has already managed to add **870 young professionals to his team, including programmers interested in creating a more open state**. The municipalities of Miraflores, San Isidro and Lima have received this message and they already offer reusable information on their online portals.

Shortly after this investigation, the Peruvian media landscape was invigorated with new independent media outlets that looked a bit more like startups than traditional media organisations. They were small outfits with teams of highly qualified journalists using dynamic research methods. The essential characteristic that made them different, even from similar organisations in other parts of the world, was that from the beginning they planned to produce investigations of impact, with the support of digital resources.

The first one of these media outlets was **OjoPúblico**, created by reporters Oscar Castilla, David Hidalgo, Nelly Moon and Fabiola Torres, who already had much experience working in some of the most important media organisations in the country. They worked in partnership with the programmer Antonio Cucho, an open data activist and founder of the Open Data community in Peru.

OjoPúblico shook public opinion with the launch of 'Sworn Auditors' ('Cuentas Juradas' in Spanish) in 2014, the first journalistic app that revealed the evolution of the patrimony of people working in local authorities who sought re-election in the same electoral process. Through a massive analysis of the affidavits submitted by these candidates and the resumes given to the electoral body, the portal published a series of investigations on the inconsistencies, gaps and other irregular aspects related to property and income. The work was done in partnership with the NGO Suma Ciudadana, which is in charge of an essential part of traditional FOI requests. It also had the support of the members of the HackSpace of the National University of Engineering, whose members downloaded and processed most of the necessary information for crossing the data.

"If I didn't partner with the journalists of **OjoPúblico**, a giant file of 10 years of affidavits of the mayors of Lima would have ended up as a pile of useless paper," said Javier Casas, president of Suma Ciudadana, about 'Cuentas Juradas'. His organisation had collected dozens of affidavits from mayors requested from the Comptroller since 2012, but it took almost two years until he found a team of jour-

nalists willing to process and interrogate the documents in order to turn them into compelling stories.

Soon after the digital site Convoca was launched, led by journalist Milagros Salazar Herrera, who organised a team of five young reporters and two developers. Their first job was building a complete record of more than 1,200 disciplinary cases opened by the Environmental Assessment and Control Agency (OEFA in Spanish) between 2010 and 2014 against companies in the mining, oil, electricity and fishing sector.

This database built by the journalists themselves, combined with six months of journalistic work, generated the series 'Unpunished excesses: the environmental trail of extractive industries'. The analysis revealed that the mining and oil companies that had received most of the fines by the supervisory body of the Ministry of Environment were also the worst repeat offenders. They had also established a legal framework to appeal to the judiciary, which had meant they had frozen more than $30 million in sanctions. It was another revelation, thanks to multidisciplinary work, where journalists detected the potential of a story, then organized the strategy and resources ad hoc to find the evidence.

The result is a curious phenomenon in Peruvian journalism: while the major traditional media outlets were reducing human resources because of a crisis of revenue, new digital media outlets were becoming a source of quality content, impact and innovation. The reports produced by **OjoPúblico**, Convoca and Utero.pe started online and were later published by national newspapers that were interested in the quality of the investigations, and in the traction on social media.

While working with databases and establishing the required workflow is an ongoing challenge in the local newspaper industry, some, such as El Comercio and La Republica, have jumped on the trend with the use of free digital tools, like Tableau to display sports and election results, or Thinglink for entertainment content. The concern about using technology to do better journalism is already looming. Fortunately, the process is irreversible.

"We believe the future of media will be discovered through lots and lots of experiments."

Corey Ford, *Executive Director at Matter*

Case Study: INTENSIVE CARE

[A health news app, or the news that never dies]

The meeting between journalism and technology offers a range of possibilities to present findings to a new audiences. It's accepted that, while visualisations allow us to understand a story from a chart, apps allow the audience to understand several stories within the same piece and give the user the opportunity to find alternative ways to consume the content. It's a very different process from the traditional one-way direction of information; the difference between listening to a speech and having a conversation. The applications offer readers free access to specific data of interest, a personal navigation experience and, consequently, the possibility of understanding an issue in the way that suits them best.

A clear example is 'Intensive Care' ('Cuidados Intensivos'), a news app created by **OjoPúblico** to investigate the private healthcare sector in Peru. At first, the tool reveals the penetration of large national financial groups within the business of clinics and medical centres, a sector that had grown in previous years with little state supervision. But the greatest value for the reader-citizen-patient is that the application is the first full record of clinics and doctors that have received administrative penalties and criminal prosecutions for medical malpractice, as well as poor care practice.

This application is the result of building databases to understand the different dimensions of the sector. The first step was making 52 FOI requests. Then, they reviewed documental archives and massive downloads of data from the website of 44 clinics in the country. This volume enabled them to build a search platform showing tabs of 61,372 registered doctors, 9,920 health facilities and 21 fund management companies (between health care providers, insurers and the clinics that offer their own health care programmes). So, the tool allows any user to enquire if a doctor or clinic is authorised to provide services, what their specialty is, their effectiveness and whether the facility has received administrative sanctions or has been involved in malpractice lawsuits.

The journalistic work behind Intensive Care

The investigation into the private health system in Peru and the construction of the Intensive Care app were carried out by five journalists and a developer. The team of reporters processed the data in Excel and used the program Refine Open to clean and cross the information from the various databases.

The scope of the data collection work and the processing to obtain relevant conclusions allow us to establish some clear lessons for any journalist who wants to accept a challenge of this nature:

1. The dimension of the data is transformed

The starting point of Intensive Care was the design of a structure of key databases to understand the sector, as well as identifying the state institutions that kept the necessary information. In this first phase, the team set out to develop four databases: the first, of the corporate groups with investments in the health sector; the second, of all private medical establishments registered in the country (from optometrists or GP clinics to specialised clinics); the third, of the insurance companies and health fund managers; and the fourth one, of registered doctors in Peru.

The main challenge was to check when that the data was updated. The first time **OjoPúblico**'s reporters asked the National Health Authority (Susalud) about the official registration of private medical establishments, the answer was that the list was published on the entity's website. Here, they found a list of 2,500 private health services. A month later - while processing the first list - they were told that Susalud had created a new form to classify information of private establishments. The new standard had over 9,000 registrants and included new terms. The initial work was incomplete and now

also outdated. So the team had to scrape all the online forms again to turn them into an Excel file.

A similar problem occurred with the standard of 60,000 registered doctors. At one point, the team took random samples to check the data, and found that the Medical College did not regularly update the accredited specialties of their members. In several cases, the data was incomplete and so the problem had to be fixed by hand, with specific searches into the specialists' history.

2. Official information is contradictory

The team conducted a second process of information gathering to create more sensitive index cards of the health facilities, doctors, insurance companies and health funds managers. They requested all sanctions imposed on private health companies from 1992 to mid-2015, issued by the National Institute of Competition and Protection of Intellectual Property (Indecopi in Spanish). During that period, Indecopi was the only public agency responsible for supervising and sanctioning malpractice in the private sector, to the detriment of patients. But it had only filed resolutions since 2011.

Indecopi's first response was that the team could download the PDF documents stored on their website, but they insisted they wanted the hard copies. Only then, did they find out that the agency had more information but that had not been processed. By comparing the soft and hard copy versions, they found 30 resolutions against clinics that did not appear on the website.

The Intensive Care application processed more resolutions than Indecopi to produce its ranking of clinics that had received administrative sanctions. With the full list, the journalists were able to verify that most clinics did not pay the fines and even challenged the sanctions in court.

3. Technical terms hide revelations

During the investigation, the team had to learn the technical terminology used by the state agencies to categorise health facilities. Without a clear understanding of this management jargon, they would have missed important data. A clear example was when an indicator was detected which the National Health Authority (Susalud in Spanish) calls: level of operational risk. The corresponding figure was a percentage without further explanation.

In the official jargon, this concept refers to the result of the evolution of private service and measures their degree of compliance with the standards of patient care (conditions and equipping of emergency services, intensive care units, pharmacies). Susalud inspectors registered a percentage for each service provider and this figure actually corresponded to the percentage of compliance. So, when reports said that a clinic had "operational risk level: 6%", what it actually meant was that the establishment did not meet 94% of the care standards. The impact of the data collected had changed radically.

To understand the terminology, the journalists consulted a number of health experts who helped them explain it in ordinary language for users. It was then placed in a comprehensive way in all of the records of the health facilities assessed.

Project's digital tools

The programming work was done using Python, chosen because of its efficiency and performance. The structure was made with the Django framework, which has a powerful content management system, as well as being a modular system, so you can potentially scale the application. For the management system database, PostgreSQL was chosen for its ability to store large amounts of data. And the system search was conducted in Elasticsearch, which offers a powerful engine.

4. If the database does not exist, there are ways to build it

One of the biggest challenges was to find a solution to the lack of information. Peru lacks an official record of sanctioned medical malpractice. The information was requested from the Ministry of Health, the Medical Association, the Association of Private Clinics and the judiciary. None of these institutions had any files available on the subject. The team decided to build an initial database based on formal complaints reported by the media.

To do this, the team plunged into the files of three of the country's largest newspapers: El Comercio, La Republica and Ojo. They reviewed the period from 1991 to mid-2015 and then dumped the information into an Excel table with the following fields: name of the victim, clinic or hospital where the malpractice occurred, medical or

health professional, reported brief description of the facts, and year of occurrence. With this background, they compared the names of the doctors and establishments in the records of the Public Ministry and the judiciary. Only the formalized cases were included.

5. If information costs, treat yourself to free it

The files of the companies that provide health services have information that comes from the National Superintendency of Public Registries (Sunarp) and the judiciary. One small detail is that both state agencies charge a fee for each search. To view each item, Sunarp charges S/.4 ($0.3), and one company can have several registry certificates.

To find out the status of a demand for information, S/.1 has to be paid to the judiciary. The team decided to bear the cost to access the records of about 50 companies on which the investigation was concentrated and the data was released within the Intensive Care project.

WORKSHOPS FOR REPORTERS

[Partner organizations to combine journalism and technology]

Investigative Reporters & Editors, IRE

http://www.ire.org/

Organizes conferences and training courses for journalists. It's based in the School of Journalism at the University of Missouri. NICAR runs the programme, which promotes the use of databases for investigative journalism.

Global Investigative Journalism Network

http://www.globalinvestigativejournalism.org/

The Global Investigative Journalism Network, created in 2003 in Copenhagen, organizes the Global Investigative Journalism Conference every two years. The next event will be in Johannesburg, South Africa, in 2017.

Knight Centre for Journalism in the Americas

http://www.knightcenter.utexas.edu/

The Knight Centre is permanently training journalists from Latin America and the Caribbean on the latest digital tools through free seminars online. Among its lecturers are leading database journalists.

Internacional Center for Journalists, ICFJ

http://www.ijnet.org/

This organization offers training in investigative journalism and digital tools. It has a program to develop innovative journalism projects.

To have or not to have data

[Transparency Law vs. Protection of Personal Data Law]

Thirteen years after it came into force, the Law of Transparency and Access to Public Information in Peru has become an essential tool for investigative journalists. However, there have also been several setbacks in this area. The main stumbling block is the criteria that officials apply to personal information. In June 2015, the lawyer Javier Casas asked the Comptroller General of the Republic for a copy of the affidavit of income, assets and income of the President, Ollanta Humala, arguing that this document is public and can therefore be requested, according to law. After four months of continuous demands, Casas received a letter that rejected his request, with a dubious explanation: affidavits are the private information of public servants regulated by the Law on Personal Data Protection since 2011.

Casas, the president of the NGO Suma Citizen and a specialist in access to information law, considered that the Comptroller had overlooked the constitutional provision that established that a state is obliged to make public the affidavits of officials to their citizens. But the case showed something else: the friction between the legislation that promotes the opening of public data and other state agencies that have started to use their discretion to limit access.

Months earlier, Casas had interviewed the head of the National Authority for Protection of Personal Data (ANPDP), Jose Quiroga León, who said the following: “The affidavits of civil servants do not require consent of the holders to be delivered, because they are not within the norm that regulates the protection of personal data.” The argument of the Comptroller was then a particular interpretation in order to deny the release of such information.

27

FOI requests made by journalists or civil society organizations have been rejected or not answered fully by the Peruvian Government between 2003 and 2015.

The paradox of this situation is that the state spends S/.11 million a year to maintain a Ministry of Public Administration in order to promote the openness of information, but at the same time, several of its ministries and agencies refuse requests by citizens or journalists.

A review of the reports of the Liber Center (Centro Liber), the Institute for Press and Society (IPYS) and the Ombudsman identified 17 public entities that between 2003 and 2015 had refused to respond to FOI requests, or they only did so partially, or after the deadline. Most of these FOI requests had journalistic purposes.

Can we read the emails of a Secretary?

In mid-2014, the hacker group Anonymous and Lulz Security leaked the largest ever number of emails of a senior Peruvian official. Its members compromised the account of the then Prime Minister, René Cornejo, and made public 6,482 messages. The press dubbed the event as ‘Cornejoleaks’.

The volume of the leak had great political and media impact, because it revealed the alleged secret lobbying of cabinet members in favour of various corporations, and also opened a legal debate on the public nature of official communications from the authorities.

Among the emails hacked, there was a huddle of emails between ministers of Energy and Mines, Eleodoro Mayorga, and Environment, Manuel Pulgar Vidal, on a decision that had been approved to directly benefit companies in the oil sector. When the journalists began reporting on this, the Liber Center, a not-for-profit organization that promotes state transparency, requested from Mayorga’s office a copy of the “emails received by the minister to the official email account or any another email that has been created in the

ministry, and their respective response, when the communications were to discuss issues related to the New National Hydrocarbons Regulations or similar". The request was made based on the Law of Transparency and Access to Public Information, but it was rejected. The ministry argued that the order violated the ministers' secrecy of communications.

Then the president of the Center Liber, former anti-corruption prosecutor July Arbizu, decided to file a *habeas data* arguing that the content of the conversations had by the Minister of Energy and Mines was in the public interest and did not violate in any way privacy or secrecy of communications.

The case was in the hands of Judge Hugo Velasquez Zavaleta, from the Fifth Constitutional Court of Lima, who upheld the claim of habeas data almost a year later, when Mayorga had already left office. "Public information may be requested and the state administration is obliged to deliver. It can be recorded in any form of expression, whether graphic, sound, visual, electromagnetic or work in any other material support," says the ruling in June 2015. The decision considered that "with the evolution of technology, communication is no longer done only through the support of paper, but through other means such as e-mails".

Judge Velasquez's conclusion was based on two principles: advertising and maximum disclosure. The first is governed by Article 3 of the Law of Transparency and Access to Public Information, which states that "all information held by the State is presumed to be public, unless expressly provided provisions". The second principle was developed by the Inter-American Court of Human Rights and enshrined in the jurisprudence of the Constitutional Court, stating that "advertising on the performance of public authorities is the general rule, and secrecy is the exception, as it is mentioned in the constitutional coverage exception".

The Ministry of Energy and Mines appealed the ruling despite his successor, Minister Rosa Ortiz Rios, saying in a TV interview that she would provide the public information found in the email of her predecessor.

"The reuse of data made public by the State allows citizens to exercise the right of access to information."

Miguel Morachimo, *director of the NGO Hiperderecho.*

The Liber Centre also filed habeas data resources requesting the emails of the Minister of Agriculture, Milton Von Hesse, and Economy, Luis Castilla, when they addressed issues directly related to their management of their portfolios. Closing this manual edition, both processes were pending court ruling.

The debate about these cases made it clear that FOI laws are powerful tools to allow journalists to access information on sensitive issues that concern both the management of specific officials and on public policy. It is not an automatic guarantee to getting the evidence, but should be one of the essential tools in the mind of an investigative journalist.

The existential dilemma

When is private data in public interest?

In July 2011, the Presidency of the Council of Ministers (PCM) brought into legislation the Law on Protection of Personal Data in order to ensure proper treatment of private information of any citizen. In theory, it is a positive ideal, it prevents the disclosure of sensitive data (such as health and personal property), which is very abundant in the digital era. It does not regulate or restrict the use of public data, but admits there are grey areas that have challenged the use of public access databases and data containing the names of people.

A clear example is the case of the portal Datos Peru. In October 2014, the National Authority for the Protection of Personal Data (APDP in Spanish) fined the site for $228,000 for replicating the legal appointments, and administrative sanctions made against officials and state employees, originally published in the bulletin attached to the official gazette El Peruano. This information is public in itself, although very few people read it.

Two people had requested that the site administrators eliminate information about their cases, but they refused to do so because the data was from a public document. According to APDP, Data Peru violated the Law on Personal Data Protection because it didn't have the consent of the individuals to publish such information, even though the same content also appeared on the websites of El Peruano and the Ministry of Justice.

This was the first penalty given by the APDP and it created controversy, to begin with because of the apparent contradiction to call data private when the same information is disclosed on a portal of the state or on another site. To this issue, there was also a greater concern: the potential impact on journalistic practice and citizen oversight. Can an official decree secrecy about what the state itself is obliged to inform citizens of? And in a more fundamental sense: where does the public interest begin and end?

22

indicators are included in The Action Plan Open Government 2012-2014, but these have not yet been fully met.

"An autonomous authority is required to follow, and monitor the delivery of public information, because officials are limited by the orders of their superiors."

Roberto Pereira, *Centro Liber.*

In times when the world is moving towards a culture of open data, investigative journalists still have several obstacles to overcome. The first step is to know about the tools available to get the information you need and then process it in innovative ways. Until recently, it was closer to craft and intuition, but the powerful union of journalism and technology has enriched the methods and standards of the profession. More than a set of new instruments, we have a strategic resource. It is, as it's said, both a diverse and practical tool, like a Swiss Army knife.

THE FINE PRINT: REQUESTING INFORMATION EFFECTIVELY

The Law of Transparency and Access to Public Information establishes that public bodies have a deadline of **seven working days to issue a response** after receiving an FOI request. In case of a delay, they have a five-day extension to deliver, after first notifying the applicant.

According to the Citizens' Manual to Access Public Information, prepared by the Peruvian Press Council, if a public body breaks the law, as in several of the cases presented, **citizens have the right to file a *habeas data***. This legal resource not only applies when the public institution refuses an FOI, but also when the delivered information is ambiguous, or when there is no response within 10 business days after the FOI request was received. **The reporter may also file a habeas data when the entity denies access to the requested information** after an appeal addressed to the responsible official, to be reviewed by their superior. In these cases, without being represented by a lawyer, a journalist has 60 working days to submit a habeas data before the civil judge or the judge of the area where they live, or where the public institution that denied the information is located.

Although the law does not determine an official application form, this model helps FOI requests

FOI REQUEST FORM

Name of the institution	
I. Official responsible for delivering information:	
II. Applicant details:	
Full name or company name:	Identity and number:
Address:	
Email:	Phone number:
III. Requested information:	
IV. Unit to which the information is requested:	
V. Method of delivery of information:	
Format:	Date:
Signature:	

Observations: _____

_____

_____

ABOUT THE AUTHORS

David Hidalgo

News Director
of OjoPúblico

He is the author of the book 'Shadows of a ransom', about the latest armed action of the terrorist group MRTA. In 2006, he won the National Prize for Journalism and Human Rights. He was Fellow of the Edward R. Murrow Program for Journalists, organized by the US Department of State. He integrated the team that won the Data Journalism Award 2015.

Fabiola Torres L.

Editor of Data Analysis
of OjoPúblico

Investigative reporter specialized in health, corporate power and public management research. Member of Investigative Reporters and Editors (IRE). She was a Fellow of the Kiplinger Foundation of the University of Ohio, and the Global Investigative Journalism Network (Gijn). She integrated a team that won the Data Journalism Award

OjoPúblico