Backstage to the Panama Papers: big data analytics and collaborative journalism

blogs.lse.ac.uk/polis/2017/04/21/backstage-to-the-panama-papers-big-data-analytics-and-collaborative-journalism/

2017-4-21

The Panama Papers has received the Pulitzer Prize in the Explanatory Reporting category. The International Consortium of Investigative Journalism (ICIJ) – the Washington D.C.-based organization that coordinated the project – interpreted this as a prize to collaborative journalism itself. The use of technology was its fundamental ally. Looking closer at how the Panama Papers came alive, through a network of more than 100 journalists and reporting partners around the world working in secret for over a year, opens a window into how collaborative and networked journalism work today. And what vital a role technology plays.
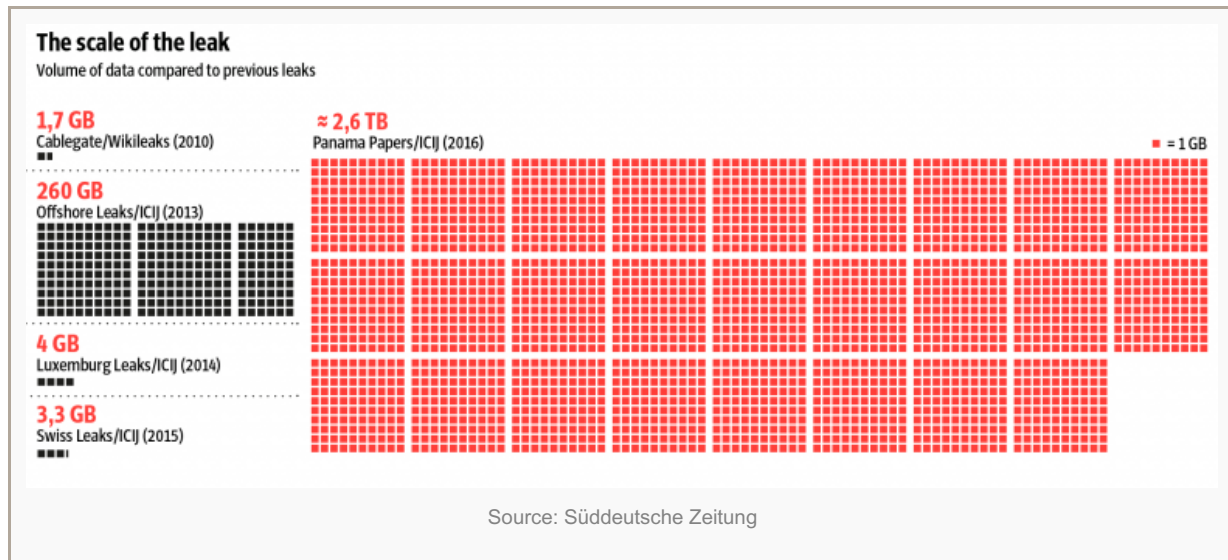
This report by Clara Aguirre Hernando.

So, how *big* was the big data in the Panama Papers?

The size of the leaked data was unforeseeable from the start. Bastian Obermayer and Frederik Obermaier were the two journalists from the German investigative paper Suddeutsche Zeitung initially contacted with the leak. Little could they have known, when the anonymous source 'John Doe' approached them with data from the Panamanian law firm Mossack Fonseca, that they would see the biggest leak in history. This source, whose identity was never revealed, gradually provided them with all sorts of documents corresponding to shell companies and off-shore accounts opened for some of the most powerful people in the world.

The data from the Panama Papers would grow to 2.6 terabytes, 11.5 billion documents. As reference, the diplomatic

cables published by WikiLeaks consisted of 1.7 gigabytes. This means that the Panama Papers were roughly 1,500 times bigger. Actually, reaching 261 gigabytes already made the Panama Papers the largest leak in the history of journalism.

The scale of the leak
Volume of data compared to previous leaks

1,7 GB
Cablegate/Wikileaks (2010)

≈ 2,6 TB
Panama Papers/ICIJ (2016)

■ = 1 GB

260 GB
Offshore Leaks/ICIJ (2013)

4 GB
Luxemburg Leaks/ICIJ (2014)

3,3 GB
Swiss Leaks/ICIJ (2015)

Source: Süddeutsche Zeitung

The "superpowers" of data technology

The ever-growing amount of data required bringing more people into the project. Bringing ICIJ on board opened up the opportunity of inviting journalists from news outlets from all around the world. The participation of journalists specializing in different national contexts made it possible to crack more cases and find more stories. Even more importantly, as the Obermai/yers described in their book (*The Panama Papers: breaking the story of how the rich and powerful hide their money*, Oneworld, 2016), given the potential of the investigations to touch so many powerful interests, relying on a wide network of people to protect both themselves *and* the data, was of the essence. The 'safety in numbers' principle.

The use of technology was the crucial piece. First, ICIJ specialists set up a secure forum –a platform similar to a chat room, based on the open source programme Oxwall, improved by ICIJ– to which every journalist in the project had access to. They all periodically posted new leads found in the data, in real time.

As a first step in making sense of the data, ICIJ introduced a software programme used mainly in forensic investigation called Nuix Investigator. This programme converts PDFs and image files into 'readable' information – in other words 'searchable' information. Rather than having to sift through each file individually, journalists could now extract and aggregate data from the whole database.

As the data continued to grow, a multi-level security database for secret documents accessible worldwide was set up. Later on, a new idea came into the project: the use of 'graph databases'. As opposed to traditional databases which show data in a spread sheet-like structure, graph databases allow users to visualize data as a network of nodes and connections. This brought to light the hidden links between accounts and their owners even when they were buried under layers of shell companies. During a Webinar months later, Mar Cabra, ICIJ Data Editor, would say in reference to the efficiency that Neo4J, the specific software they used, brought to the project: "We felt we [had] superpowers".

What the Panama Papers mean for journalism today

This case of collaborative journalism shows how investigative journalism increasingly relies on data mining and big data techniques as an essential tool in their investigations. It is also an example of how journalists are gradually learning new techniques, or incorporating data specialists to their teams. A further transformation to the ever-changing landscape of newsrooms?

All the information from the Panama Papers is available online, both as database and standardised information – and anyone can go 'fishing' in it for new leads.

For more on other collaborative journalism projects taking place right now, check this piece on Journalists around the world are working together more than ever. Here are 56 examples.

Clara Aguirre Hernando is currently a candidate of the MSc in Media and Communication (Data & Society) at the London School of Economics and Political Science.