16. Infrastructuring Collaborations Around the Panama and Paradise Papers

Emilia Díaz-Struck, Cécile Schilis-Gallego and Pierre Romera

Abstract

How the International Consortium of Investigative Journalists (ICIJ) makes digging through gigantic amounts of documents and data more efficient.

Keywords: data leaks, text extraction, radical sharing, cross-border investigation, data journalism, International Consortium of Investigative Journalists

The International Consortium of Investigative Journalists (ICIJ) is an international network of journalists launched in 1997. Journalists who are part of ICIJ's large collaborations have diverse backgrounds and profiles. There is a wide range of reporters with different skills, some with strong data and coding skills, others with the best sources and shoe-leather reporting skills. All are united by an interest in journalism, collaboration and data.

When ICIJ's director Gerard Ryle received a hard drive in Australia with corporate data related to tax havens and people around the world as a result of his three-year investigation of Australia's Firepower scandal, he couldn't at that time imagine how it would transform the story of collaborations in journalism. He arrived at ICIJ in 2011 with more than 260 gigabytes of data about offshore entities, about 2.5 million files, which ended up turning in a collaboration of more than 86 journalists from 46 countries known as Offshore Leaks (published in 2013).¹

¹ <https://www.icij.org/investigations/offshore/how-icijs-project-team-analyzed-offshore-files/>

After Offshore Leaks came more investigative projects with large data sets and millions of files, more ad hoc developed technologies to explore them, and more networks of journalists to report on them. For instance, we recently shared with partners a new trove of 1.2 million leaked documents from the same law firm at the heart of the Panama Papers investigation, Mossack Fonseca.² This was on top of the 11.5 million Panama Papers files brought to us in 2015 by the German newspaper *Süddeutsche Zeitung* and 13.6 million documents that were the basis of the subsequent Paradise Papers probe.³

If a single journalist were to spend one minute reading each file in the Paradise Papers, it would take 26 years to go through all of them. Obviously, that's not realistic. So, we asked ourselves, how can we find a shortcut? How can we make research more efficient and less time consuming? How can technology help us find new leads in this gigantic trove of documents and support our collaborative model?

In this chapter we show how we deal with large collections of leaked documents not just through sophisticated “big data” technologies, but rather through an ad hoc analytical apparatus comprising of: (a) international collaborative networks, (b) secure communication practices and infrastructures, (c) processes and pipelines for creating structured data from unstructured documents, and (d) graph databases and exploratory visualizations to explore connections together.

Engaging With Partners

The ICIJ's model is to investigate the global tax system with a worldwide network of journalists. We rally leading reporters on five continents to improve research efforts and connect the data dots from one country to another.⁴

Tax stories are like puzzles with missing pieces: A reporter in Estonia might understand one part of the story; a Brazilian reporter might come across another part. Bring them together, and you get a fuller picture. ICIJ's job is both to connect those reporters and to ensure that they share everything they find in the data.

We call our philosophy “radical sharing”: ICIJ's partners communicate their findings as they are working, not only with their immediate co-workers, but also with journalists halfway around the world.

2 <https://www.icij.org/investigations/panama-papers/>

3 <https://www.icij.org/investigations/paradise-papers/>

4 <http://www.icij.org/journalists>

In order to promote collaboration, ICIJ provides a communication platform called the Global I-Hub, building on open-source software components.⁵ It has been described by its users as a “private Facebook” and allows the same kind of direct sharing of information that occurs in a physical newsroom. Reporters join groups that follow specific subjects—countries, sports, arts, litigation or any other topic of interest. Within those groups, they can post about even more specific topics, such as a politician they found in the data or a specific transaction they are looking into. This is where most of the discussion happens, where journalists cross-check information and share notes and interesting documents.

It took ICIJ several projects to get reporters comfortable with the I-Hub. To ease their way onto the platform and deal with technical issues, ICIJ’s regional coordinators offer support. This is key to ensuring reporters meet the required security standard.

Encrypting Everything

When you conduct an investigation involving 396 journalists, you have to be realistic about security: Every individual is a potential target for attackers, and the risk of breach is high. To mitigate this risk, ICIJ uses multiple defences.

It is mandatory when joining an ICIJ investigation to setup a PGP key pair to encrypt emails. The principle of PGP is simple.⁶ You own two keys: One is public and is communicated to any potential correspondent who can use it to send you encrypted emails. The second key is private and should never leave your computer. The private key serves only one purpose: To decrypt emails encrypted with your public key.

Think of PGP as a safe box where people can store messages for you. Only you have the key to open it and read the messages. Like every security measure, PGP has vulnerabilities. For instance, it could easily be compromised if spyware is running on your computer, recording words as you type or sniffing every file on your disk. This highlights the importance of accumulating several layers of security. If one of those layers breaks, we hope the other layers will narrow the impact of a breach.

5 See <https://www.icij.org/blog/2014/07/icij-build-global-i-hub-new-secure-collaboration-tool/>. For a different perspective on journalistic platforms such as the I-Hub, see Căndeia’s chapter in this book.

6 <https://www.gnupg.org/>

To ensure the identity of its partners, ICIJ implements two-factor authentication on all of its platforms. This technique is very popular with major websites, including Google, Twitter and Facebook. It provides the user with a second, temporary code required to log in, which is usually generated on a different device (e.g., your phone) and disappears quickly. On some sensitive platforms, we even add third-factor authentication: The client certificate. Basically, it is a small file reporters store and configure on their laptops. Our network system will deny access to any device that doesn't have this certificate. Another noteworthy mechanism ICIJ uses to improve its security is Ciphermail. This software runs between our platforms and users' mailboxes, to ensure that any email reporters receive from ICIJ is encrypted.

Dealing With Unstructured Data

The Paradise Papers was a cache of 13.6 million documents. One of the main challenges in exploring them came from the fact that the leak came from a variety of sources: Appleby, Asiaciti Trust and 19 national corporate registries.⁷ When you have a closer look at the documents, you quickly notice their diverse content and character and the large presence of “non-machine readable” formats, such as emails, PDFs and Word documents, which cannot directly be parsed by software for analyzing structured data. These documents reflect the internal activities of the two offshore law firms ICIJ investigated.

ICIJ's engineers put together a complex and powerful framework to allow reporters to search these documents. Using the expandable capacity of cloud computing, the documents were stored on an encrypted disk that was submitted to a “data extraction pipeline,” a series of software systems that takes text from documents and converts it into data that our search engine can use.

Most of the files were PDFs, images, emails, invoices and suchlike which were not easily searchable. Using technologies like Apache Tika (to extract metadata and text), Apache Solr (to build search engines) or Tesseract (to turn images into text), the team built an open-source software called

7 <https://www.icij.org/investigations/paradise-papers/paradise-papers-exposes-donald-trump-russia-links-and-piggy-banks-of-the-wealthiest-1-percent/>, <https://www.icij.org/investigations/paradise-papers/appleby-offshore-magic-circle-law-firm-record-of-compliance-failures-icij/>, <https://www.icij.org/investigations/paradise-papers/roll-roll-asiacitis-u-s-marketing-tour/>

Figure 16.1. Inside the leak.

Extract with the single mission of turning these documents into searchable, machine-readable content.⁸ This tool was particularly helpful in distributing this now-accessible data on up to 30 servers.

ICIJ also built a user interface to allow journalists to explore the refined information extracted from “unstructured data”: The hodgepodge of different types of documents from various sources. Once again the choice was to reuse an open-source tool named Blacklight which offers a user-friendly web portal where journalists can look into documents and use advanced search queries (like approximate string matching) to identify leads hidden in the leak.⁹

Using Graphs to Find Hidden Gems Together

ICIJ published its first edition of the Offshore Leaks database in 2013 using graph databases to allow readers to explore connections between officers and more than 100,000 offshore entities. This has grown to over 785,000 offshore entities at the time of writing, including from subsequent leaks such as the Panama and Paradise Papers.

ICIJ first attempted to use graph databases with Swiss Leaks, but it was with the Panama Papers that graph databases started playing a key role during the research and reporting phase. To explore 11.5 million complex financial and legal records amounting to 2.6 terabytes of data was not an easy task. By using network graph tools such as Neo4J and Linkurious, ICIJ

8 <https://github.com/ICIJ/extract/>

9 <https://github.com/projectblacklight/blacklight>, https://en.wikipedia.org/wiki/Approximate_string_matching

was able to allow partners to quickly explore connections between people and offshore entities.

Our data and research teams extracted information from the files, structured it and made data searchable through Linkurious. Suddenly partners were able to query for the names of people of public interest and discover, for instance, that the then Icelandic prime minister, Sigmundur Gunnlaugsson, was a shareholder of a company named Wintris. The visualization with this finding could be saved and shared with other colleagues working on the investigation in other parts of the world.

One could then jump back into the document platform Blacklight to do more advanced searches and explore records related to Wintris. Blacklight later evolved to the Knowledge Center in the Paradise Papers. Key findings that came through exploring data and documents were shared through the Global I-Hub, as well as findings that came from the shoe-leather reporting.

Graph databases and technologies powered ICIJ's radical sharing model. "Like magic!" several partners said. No coding skills were needed to explore the data. ICIJ did training on the use of our technologies for research and security, and suddenly more than 380 journalists were mining millions of documents, using graph databases, doing advanced searches (including batch searches), and sharing not only findings and results of the reporting, but also useful tips on query strategies.

For the Panama Papers project, graph databases and other ad hoc technologies like the Knowledge Center and the Global I-Hub connected journalists from nearly 80 countries working in 25 languages through a global virtual newsroom.

The fact that structured data connected to the large number of documents was shared with the audience through the Offshore Leaks database has allowed new journalists to explore new leads and work on new collaborations like the Alma Mater and West Africa Leaks projects. It has also allowed citizens and public institutions to use them independently for their own research and investigations. As of April 2019, governments around the world have recouped more than USD1.2 billion in fines and back taxes as a result of the Panama Papers investigation.

Since the first publication of the Panama Papers back in 2016, the groups of journalists using ICIJ technologies has grown and more than 500 have been able to explore financial leaked documents and continue to publish public interest stories linked to these millions of records.

About the Authors

Emilia Díaz-Struck is research editor and Latin American coordinator at the International Consortium of Investigative Journalists (ICIJ).

Cécile Schilis-Gallego is a journalist at Forbidden Stories and she previously worked as a data journalist at the ICIJ.

Pierre Romera is Chief Technology Officer at the ICIJ.