

Case Study



The International Consortium of Investigative Journalists (ICIJ)

Neo4j Enables Pulitzer Prize-Winning Investigation into Global Tax Evasion

INDUSTRY

Media & Publishing

USE CASE

Fraud Detection & Graph-Based Search

GOAL

- Expose tax evasion by uncovering people's assets hidden in offshore companies

CHALLENGE

- Search and analyze 11.5 million source documents for concealed links

SOLUTION

- Fraudulent practices exposed using Neo4j graph database and Linkurious data visualization tool

RESULTS

- Arrests and resignations of high-profile politicians and others
- Financial rules tightened globally, including in the U.S., EU and UK

With the Panama Papers investigation, the ICIJ encountered an overwhelming 40 years' worth of confidential documents to research and find connections. They used Neo4j to analyze at least 11.5 million documents, exposing tax evasion ranging from country Presidents and Prime Ministers to drug traffickers.

The Organization

The [ICIJ](#) is a global network of nearly 400 journalists, drawn from over 80 countries and 100+ media organizations such as the *BBC* and the *New York Times*. The ICIJ's core team of 12 staff, based in Washington, D.C., coordinate this network to investigate major cases of crime, corruption and other cross-border. The ICIJ has won numerous press awards, including the biggest honor of all, the Pulitzer Prize, for its Panama Papers exposé.

The Challenge

The [Panama Papers](#) investigation has been the biggest data leak in the ICIJ's 21-year history – and the biggest data leak of all time. In 2015, an anonymous leak of 11.5 million documents from Panama-based law firm Mossack Fonseca revealed the illicit use of offshore bank accounts by the world's rich and famous.

The material comprised 40 years' worth of confidential documents relating to over 200,000 companies in 21 tax havens, ranging from Switzerland and Hong Kong to Nevada in the US. These hideaways are used by individuals to conceal their true wealth from the tax authorities, behind a web of shell companies and accounts registered to front men or close relatives.

Initially, though, the ICIJ's journalists struggled to sift through this trove of emails, financial spreadsheets, passports and corporate records, written in English, French, Spanish, Russian, Mandarin and Arabic.

"It was a shock at first," said the ICIJ's Data Editor Mar Cabra.

She recognized the ICIJ needed accessible technology to analyze this interconnected dataset and uncover the complex web of connections. And Cabra's past experience suggested graph technology could be the answer.

The Solution

The ICIJ had already deployed a graph system in 2013 to publicly present the findings from its Offshore Leaks inquiry.

"This graph was the most successful product the ICIJ had ever used," Cabra said. "You could enter a name and just double click, and the networks would expand. Millions of people had gone into it. So in the Panama Papers investigations, we knew that we needed graphs to understand the data better."

Case Study



"The graph allowed you to explore these networks in a very, very easy way. My journalists were amazed. We felt like we had superpowers."

- Mar Cabra,
Data Editor, ICIJ

To tame the 2.6 terabytes of Panama Papers data, the ICIJ extracted the document metadata using Apache Solr and Tika, then connected all the information together in a [Neo4j graph database](#), accessed by the Linkurious data visualization tool. Alongside this, its member journalists used the OXWALL open source social platform to share their findings, tips, leads – and threats – relating to the investigation.

The ICIJ's developers built the Neo4j graph around the leaked data's key entities such as companies, their clients and officers. This enabled the journalists to uncover relationships between these core nodes – matching, say, bank accounts to people who had the same address, family ties or business links, or who regularly emailed each other.

The graph comprised 840,000 nodes and 1.3 million relationships, but the reporters could simply type in an individual's name and instantly reveal their web of connections. They could also dig deeper into the data through advanced Cypher queries.

The result was a huge leap forward from previously used technologies.

"The graph allowed you to explore these networks in a very, very easy way that anybody could understand," Cabra said. "My journalists were amazed. We felt like we had superpowers, because the reaction was, 'Oh my God, I did not see these connections before by looking through documents, I'm finding more stories.' To them, this was magic. With graph databases you're basically able to find connections that you couldn't see before when working in an SQL database."

The Results

Since April 2016, the ICIJ has published a series of Panama Papers exposés that have triggered hundreds of investigations and audits in more than 82 countries, and driven new financial rules in the U.S., EU, UK and Switzerland.

Using Neo4j, the ICIJ revealed that some 500 banks had registered nearly 16,000 shell companies for Mossack Fonseca clients – including 14 current or former world leaders, billionaires, drug traffickers, celebrities, sports stars and a network of people close to Russian President Vladimir Putin who shuffled some \$2 billion around the world.

These revelations have prompted arrests and resignations of high-profile figures worldwide. They included Iceland's President Ólafur Grímsson, who stepped down after the graph connected his wife to an undeclared offshore company holding millions of dollars in bank bonds. Former Pakistan Prime Minister Nawaz Sharif was also jailed for corruption. In 2017, Mossack Fonseca's founders themselves were arrested on money laundering charges, and in March 2018 the company shut down.

For all this, the ICIJ was rewarded with the [2017 Pulitzer Prize for Explanatory Reporting](#).

"This was the biggest collaborative investigation in journalism history, and Neo4j and Linkurious played a big role," Cabra concluded.

Neo4j is the leader in graph database technology. As the world's most widely deployed graph database, we help global brands – including [Comcast](#), [NASA](#), [UBS](#), and [Volvo Cars](#) – to reveal and predict how people, processes and systems are interrelated.

Using this relationships-first approach, applications built with Neo4j tackle connected data challenges such as [analytics and artificial intelligence](#), [fraud detection](#), [real-time recommendations](#), and [knowledge graphs](#). Find out more at [neo4j.com](#).