Embargoed until Friday June 14 at 10 pm EST // 4:00 CET on Saturday June 15

**How La Nación Costa Rica developed ICIJ's**

**Application to visualize offshore companies**

By Giannina Segnini

An art restorer is able to rebuild a brokensculpture after thoroughly studying its multiple unconnected pieces in order to decode how they related in the past, to return them to their original place and to discover its full shape again.

Just like art restorers, the Investigative Unit at *La Nación Costa Rica* received in November 2012 a device with millions of data in different formats. The relational databases came scattered over more than 320 tables and without an original dictionary to explain their relations.

These databases were parts of two larger separated databases that had been fed for nearly30 years by two companies: Singapore-based Portcullis Trust Net (PTN), and Commonwealth Trust Limited (CTL), based in the British Virgin Islands (BVI).

Both firms specialize in setting up offshore financial structures. They have helped tens of thousands of people create offshore companies and trusts, as well as hard-to-trace bank accounts.

The data were obtained by the International Consortium of Investigative Journalists (ICIJ), which chose The Investigative Unit at *La Nacion* to process it and to develop the interactive application of the most ambitious cross-border investigative project in history.

The task did not start from zero. In the preceding months, UK journalist Duncan Campbell and programmer Matthew Fowler had made progress in understanding and documentingpart of those relations.

Between January and April, the Investigative Unit's computer science engineer, Rigoberto Carvajal, thoroughly analyzed the data and, with

advice from the UK team and data journalist Mar Cabra, applied reverse engineering processes to reveal the original relations between tables, fields, codes, and, ultimately, hundreds of thousands of records of companies and people.

As he started work,Rigoberto found himself faced with a disorganized and scattered structure, which for years enabled an insufficient and incomplete feeding of data, duplications, void values, unneeded repetitions, missing data and poorly solved relations.

There were thousands of names of people and companies which were duplicated because they had minimal variations in some character, abbreviations, typing errors, or a slightly different order of the elements.

If the data remained that way, the true links and relations of each separate element would have never been disclosed through visualization. It would have been something similar to varnishing, without first sanding them, the dirty pieces of a disassembled sculpture.

Part of the solution consisted in integrating the databases to bring together their similar entries and then organizing them in such a way that the structure would become practical for visualization.

In order to do so, the *La Nación* team used the Talend Open Studio for Data Integration, an open source tool for ETL (Extraction, Transformation and Load).

Talend hosted all of the processes: extracting the databases tables, organizing their structure to combine similar records, converting them into a node and link structure and, finally, loading up the unified nodes into a sole database which would feed the public application.

Procedures and algorithms to de-duplicate the data were applied. In this task, alibrary developed by the Massachusetts Institute of Technology (MIT) as a result of a project named Vicino, played an important role.This library was added to the Talend Open Studio tool to apply functions in the data flow.

Also relevant was the use of the SIMILfunction, which estimates the percentage of similarity between two chains of text, based on the number of sub-strings they have in common.

With these algorithms, Carvajal merged several thousand separate recordswhichwere the same persons or companieswith a total degree of certainty and which had exactly the same addresses.

Following this merge, the links associated to each of those entities were finally related.

**Design and interface**

While these cleaning processes were underway, work was also done on the design and functionality of the interactive application. The goal was to conceive a simple and versatile solution for the visual exploration of the data.

*La Nación*'s web designer,Marco Hernández, captured his graphic proposal based on ICIJ and the Investigative Unit's requirements. Marco required that everyone involved in the project simplify their bulky and complex suggestions.

His solution prevailed almost intact during the whole approval process and allowed the development of a clean interface with simple and unified full text searches in one field.

The design of the application and subsequent adjustments were made through MockFlow, a collaborative web tool which allows the creation of original digital sketches, sharing them with multiple users, and allow them to modify or make comments on them from anywhere around the world.

The tool was perfect for an international project such as this one, in which members from more than five countries participated in the definition of the design.

Matthew Caruana-Galizia, the Investigative Unit's web developer, undertook the development of the web application.

To Matthew, who is also a journalist, the idea behind the project was clear: to build an interface in which any user could easily explore the data of offshore companies.

He took charge of visualizing the data in nodes (or circles), which represent companies or people, and lines, which represent the links between each of them.

He also programmed the site to generate an independent page for each person or company consulted, in such a way that users could access a permanent URL for each of them, and then share it or go back and consult it later.

The application enables the user to expand any node, starting at the visualization, to discover its relations, if it has any. The system allows users to create their own relation map, but with the advantage of being able to go back in case they expanded an entity with lots of connections. The “undo” key returnsthe navigation to the preceding stage of the visualization, when it was still legible.

One of the major challenges during the development of the web interface was what Matthew called the “hairball” visualization – a configuration of lines that looks like something your cat produced. This effect was produced when visualizing nodes that had thousands of connections with other companies or persons.

The answer to the problem, which also overloads the computer processing, was preventing or restricting the visualization of companies with over 100 relations or connections. For these cases, the application generates the information in tables.

For many of the nodes, however, those with 20 to 100 connections, the application considers several other functionalities to avoid the “hairball” effect.

One of them is known as the fisheye lens, which amplifies the data upon which the cursor is set. This works by distorting the area around the mouse pointer so that the selected data seem closer and easy to read. The distorted context remains visible, however, contrary to the magnifying glass effect.

The data visualization was programmed using sigma.js, which was the library with the most and best options to produce a personalized design with very good possibilities for exploration.

Beyond datarestoration, the applicationyou can navigate today also creates a visual interface from which the users can explore, in a friendly environment, thousands of relations between persons, companies, and groups which until recently remained hidden.

**GianninaSegnini is editor of the Investigative Unit at *La Nación Costa Rica and an ICIJ member.***

PCIJ