Reuters Institute Fellowship Paper

University of Oxford

Global Database Investigations: The role of the computer-assisted reporter

by Alexandre Léchenet

Michaelmas Term 2014

Sponsor: Le Monde

Acknowledgements

This work could not have been done by one person alone.

I would like to thank the Reuters Institute for the Study of Journalism and its team for their presence, presentation and work. I would also like to thank Jonathan Bright from the Oxford Internet Institute for his patience and careful reading.

As well, I would like to thank Brigitte Alfter, Mar Cabra, Nicolas Kayser-Bril, Nils Mulvad – the journalists that kindly accepted to answer my questions. I would like also to thank Michael Keller and Sylvain Parasie for their meticulous reading and advice.

I would like to thank my sponsor, *Le Monde*, and especially Natalie Nougayrède, Sylvie Kauffmann and Sandrine de Laclos.

I would also like to thank all my Whatsapp friends from here and there that were most of the time available for me and my doubts.

Table of Contents

1. Introduction

With globalisation, crime and politics have gone global as well. It needs then a global journalistic response. Even though there are examples of collaboration in journalism, it is not deeply rooted in the culture of traditional journalists. The common practice of 'lone wolf' reporters, the difference in work habits across the globe, and the high competition in the industry are all obstacles.

However, in recent years the most impressive investigations have been a result of global collaboration. The 2013 Pulitzer Prize was awarded to two newspapers for their coverage of the National Security Agency (NSA) leaked documents, which was in essence two cross-border collaborative stories. The story started in a room in Hong-Kong with an American journalist based in Brazil and working for a British newspaper, an American journalist and filmmaker living in Germany, and a British journalist. And their work was published in different newspapers in Germany, USA and Great Britain.

Besides globalisation, the world is facing digitalisation. More and more information, including about our personal lives, is recorded in huge databases. Computer-assisted reporting techniques to investigate these databases have been around for several years, but in recent years more tools and more material sources and data-driven journalism have appeared underpinned by a strong philosophy of openness and transparency.

The following research paper focuses on the role of the computer-assisted reporting techniques in these investigations. The first part is about defining and summarizing the history of computer-assisted reporting. The second will lay out the different kinds of collaboration and explain the need for it.

The following three chapters contain an analysis of three reasonably successful collaborations – one about farm subsidies, another about tax havens and the last about migrants coming to Europe – to draw some lessons from them and understand how collaborating to analyse databases helps to produce successful news stories.

2. **A brief history of computer-assisted reporting**

Computer-assisted reporting (CAR) was first used to mean the application of social science methods to do journalism. It then described more broadly the use of computers to investigate databases. Nowadays, with the advance of digitalisation and the frequent use of computer, data journalism has become common place. In this section, I'll define computer-assisted reporting and some of their techniques.

a. **The origin of CAR**

Computer-assisted reporting was born in the American newsroom in the late 1960s, inspired by the social sciences and helped by technological advances. Journalists first used a computer to help in the reporting process in covering the 1952 US national elections to predict the outcome of the vote[1]. It took another fifteen years for reporters to utilise computing power in investigations when Philip Meyer, a journalist at the Detroit Free Press used a computer when covering the Detroit riots in 1967. Through a survey made with the help of an IBM 360, he was able to sketch a profile of the rioters.

This Pulitzer-winning story was the first attempt by a journalist to use analytical methods from sociology and similar domains in the context of a newsroom. This combination, although novel, was no accident: in addition to being a trained reporter, Meyer also studied political science and the social sciences as a Nieman Fellow at Harvard University in 1966. In his book, Precision Journalism, published in 1969, he defines this new way of doing journalism as "the application of social and behavioral science research methods to the practice of journalism" (Meyer, 1979). He developed statistical and quantitative methods to investigate the effect of public policies. He suggested a scientific approach to each story: collect facts, build a theoretical model and then confront the model with the reality. Computers at that point were only tools that facilitated the application of a new method.

[1] Cindy Carpien (2012), "The Night A Computer Predicted The Next President", NPR, http://www.npr.org/blogs/alltechconsidered/2012/10/31/163951263/the-night-a-computer-predicted-the-next-president

b. **Evolution of CAR and networking**

The work Meyer started evolved in the following years. In the late 1980s, precision journalism, re-cast as CAR, appeared in the newsroom (Coddington, 2014). The number of reporters practising the techniques have grown, in part thanks to a strong professional network, built especially with the National Institute for Computer-Assisted Reporting (NICAR) founded in 1989 (Fink and Anderson, 2014). CAR at that time became mainly "the data gathering and statistical analysis descended from Meyer's precision journalism" and focused on investigation (Coddington, 2014). But the environment has changed in key ways. Whereas information was then rare, today data have become so widespread as to become a challenge.

c. **Data-journalism and the rise of data**

Data journalism or data-driven journalism is a broader use of the data for journalism. It appeared in the late 2000s, helped by cheaper access to computers, the appearance and use of the Web as a source and a publication support, and the open-data movement. As Philip Meyer coins it, (Data-Journalism Handbook, 2012) the aim of data driven journalism is to "bring sense and structure out of the never-ending flow of data".

Computer-assisted reporting and data driven journalism are often described as two different practices, although the latter derived from the first. Data journalism tends to show more the process of the story, using data-visualisation and opening it up to the readers (Coddington, 2014). As Liliana Bounegru, at the time working for the European Journalism Center, explains in the Data Journalism Handbook (written by 70 journalists from various countries and media organisations):

> "Data journalists have an important role in helping to lower the barriers to understanding and interrogating data, and increasing the data literacy of their readers on a mass scale."

Bounegru concludes that the two communities (CAR journalists and data-journalists) should work closer together to learn "more innovative ways of delivering the data and presenting the stories" and to understand "the deeply analytical approach"[2]. In 1999, Philip Meyer suggested

[2] Liliana Bounegru, "Data Journalism in Perspective", Data Journalism Handbook, http://datajournalismhandbook.org/1.0/en/introduction_4.html

that the concept of computer-assisted reporting should be abandoned. He described it as "an obsolete concept that is starting to get in the way of our understanding of the real battles that we need to fight and win if we are to save the soul of journalism"[3].

Philip Meyer might be right about the fact that nowadays every journalist is using a computer in his or her work. However, a lot of journalists only use the computer for access to the Internet and for writing their stories. Therefore, I'll use the term "computer-assisted reporting" in the next paragraph to describe an advanced use of computers and databases to investigate and uncover stories.

d. **Education and culture**

Data-driven journalism is a peculiar practice to teach because it covers such a broad range of subjects and skills. Journalists who work with data can play a lot of different roles, from scraping to visualisation or machine-learning as well as project management; no one path exists that leads people to data journalism. Some journalism schools are now starting to propose courses or specialisation. As described in "The Art and Science of Data-Journalism", most of the data driven journalists self-teach, with the help of networks like NICAR and its mailing list, meetings like Hacks/Hackers, or from informal peer-to-peer learning[4].

Either journalists or developers, the new types of workers arriving in the newsroom bring with them a new culture derived from open-source and hacker culture. After a visit to the *New York Times* multimedia team, Cindy Royal believes that "part of the success" of the team comes "from the culture that has been developed in the department, the roots of which are in creativity and innovation, driven in many ways by the open-source or hacker culture"[5].

This culture leads developer-journalists to interrogate data published by governments, and try to "disclose truth (...) through the accessing, combination and processing" of data, as Sylvain

[3] Philip Meyer, "The Future of CAR: Declare Victory and Get Out!", When Nerds and World Collide: Reflections on the Development of Computer Assisted Reporting", http://www.unc.edu/~pmeyer/carfuture.doc
[4] Alexander Benjamin Howard, "The Art and Science of Data-driven Journalism", http://towcenter.org/blog/the-art-and-science-of-data-driven-journalism/
[5] Royal (2010): "The Journalist as Programmer: A Case Study of *The New York Times* Interactive News Technology Department"

Parasie and Eric Dagiral concluded after interviewing journalists and programmers in the Chicago area[6].

e. Open-data and access to public information

Among the various resources available to data-driven journalists, data produced by governments is precious. Some administrations, pushed by activists and journalists, followed open-data strategies. They built platforms to host and share some of the data they produce. But those datasets rarely have important news inside them. Journalists rather try to combine different datasets.

When the official data are not published, reporters can make use of the law to access public information, when it exists. This can be a very efficient process in the US or for the European Union institutions, but the law is less powerful in France for example. Differences come from the obligation or not to hand out the data, and the range of data available.

In Europe, the term "wobbing", which comes from the Wet Openbaarheid van Bestuur (WOB) (open government act) in the Netherlands, defines the use of such law. US journalists would call it FOIAing, deriving from the Freedom of Information Act.

f. Open-source intelligence

Among the tools available in the computer-assisted reporting techniques, there is open-source intelligence (OSINT), which is the collection of already published information to find news. The Internet makes accessible a lot of different information material, either from national or local newspapers, public data or social media. The technique is often used for economic forecasting but can also be a source material for journalists[7].

OSINT can take many forms. One example was an investigation by Elliot Higgins, known by his pen name Brown Moses, to verify videos allegedly shot in Syria. Working from his British home and using clues available online such as Google Maps and Google StreetView or other

[6] Parasie and Dagiral (2012): "Data-driven journalism and the public good", New Media Society, http://nms.sagepub.com/content/early/2012/11/15/1461444812463345

[7] Alexandre Léchenet, "Collecter pour mieux analyser : les bases de données et le journalisme", http://data.blog.lemonde.fr/2014/10/07/collecter-pour-mieux-analyser-les-bases-de-donnees-et-le-journalisme/

websites, he was able to authenticate the videos and document the civil war, in zones not accessible to regular journalists. His work was used by newspapers and organizations to help them understand the conflict and many relied on his investigations[8].

The blogger has now started a website called Belling Cat, where, with the help of its readers, he carries on investigations on war zones from Leicester, a city in the UK. For example, he followed the Buk convoy in Russia and Ukraine. The Buk missile was suspected of being responsible for the shooting of a civil airplane in July 2014. With the help of satellite imagery and social media content, his investigation proved that it was possible for the pro-Russia fighters to use such weapons[9].

Building a database with openly available information can help to quantify a problem. For example, ten journalists across Europe gathered the notification of migrants' deaths in the newspapers or in lists built by non-governmental organizations. Their investigations, The Migrants files, helped to assess the danger of each route taken by the migrants and quantified the deaths because they are not counted by EU officials[10].

[8] Matthew Weaver, "How Brown Moses exposed Syrian arms trafficking from his front room", March 21, 2013, http://www.theguardian.com/world/2013/mar/21/frontroom-blogger-analyses-weapons-syria-frontline
[9] Geolocating the MH17 Buk Convoy in Russia", https://www.bellingcat.com/resources/case-studies/2014/09/29/geolocating-the-mh17-buk-convoy-in-russia/
[10] The Migrants Files, https://www.detective.io/detective/the-migrants-files/

3. Collaboration

An open-source and collaborative culture lends itself well to projects that seek to untangle a story hidden in a large dataset. Finding information as well as writing complete stories afterwards can be complicated for one person alone. Collaboration with others – both inside and outside the newsroom – becomes inevitable. Collaboration can help in many ways: it can make sense of the amount of data, tackle a story, and find interesting highlights. This section will focus on the different types of collaboration and their benefits for reporting.

a. Collaboration with the audience: crowd-sourcing

Outside collaboration can for example, be done with the public via crowd-sourcing. *The Guardian* chose in 2009 to ask for the help of its readers. The expenses of all the Members of Parliament were published in June 2009 by House of the Commons. But the amount of files, and the way they were edited led the journalists to build a microsite allowing readers to read and annotate the 700,000 documents.

The main goal was to enable "users to fully investigate the documents and track what they – and other users – have found", Janine Gibson, editor in chief of *The Guardian* website, explained in a press release. The microsite, she says, would help uncover new expenses, make readers "use the information" and make "this information as transparent as possible"[11]. *The Guardian* experiment worked well because readers could identify and look for their own Member of Parliament.

But crowdsourcing has its limits, and various attempts have not been successful. For example, to analyse the Afghan war logs published by Wikileaks, OWNI, a French publication covering Web culture, proposed a platform for crowdsourcing. OWNI readers could investigate the logs and point out any interesting stories. The application received a lot of visitors, but only 250 persons contributed. And among those 250 readers, 70 percent analysed only one document and then

[11] "MPs' expenses: The Guardian launches major crowdsourcing experiment", http://www.theguardian.com/gnm-press-office/crowdsourcing-mps-expenses

stopped[12]. The other 30 percent were valuable contributions, but were mainly done by specialists. A balance has to be found between journalists, experts and the crowd: database investigations often require background knowledge.

b. Collaboration between emerging and legacy media

The publication of the Wikileaks files was an opportunity for another type of collaboration. What Lisa Lynch call "megaleaks" were published thanks to an "alliance" between emerging and legacy media outlets[13]. Wikileaks needed the help of specialists and journalists to analyse the hundreds of thousands of files – Iraq war logs, Afghan war logs and diplomatic cables – and journalists wanted the stories in the files brought by Wikileaks. For the diplomatic cables publication – what is known as Cablegate – Wikileaks gave access to five newsrooms.

In large part, the editorial schedule forced the media into collaboration: the newsrooms had to work together to find stories about their and others' countries and to coordinate the whole publication so as not to spoil each other's stories. "If media groups did not learn to work across borders on stories, the stories would leave them behind", explained David Leigh and Luke Harding who worked on the files for The Guardian[14].

c. Global collaborations

The publication of large databases, and the globalisation of policies and problems like organised crime or corruption, needs an effective answer from journalists. In a blog post on the World Association of Newspaper (WAN-IFRA) blog, Knight chair in international journalism Professor Rosental Calmon Alves highlighted the need for global collaboration[15]. Even if "the cultural change is still in the works (...) cross-border collaboration is here to stay":

---

[12] Nicolas Kayser-Bril, "Journalisme et crowdsourcing", https://prezi.com/-myadcmtqdiu/journalisme-et-crowdsourcing/

[13] Lisa Lynch (2013): "Wikileaks after megaleaks", Digital Journalism, http://dx.doi.org/10.1080/21670811.2013.816544

[14] David Leigh and Luke Harding (2011), "WikiLeaks: Inside Julian Assange's War on Secrecy", Guardian Books

[15] Rosental Calmon Alves, "The growing importance of global collaborative investigative journalism", http://blog.wan-ifra.org/2014/08/07/trends-in-newsrooms-7-the-growing-importance-of-global-collaborative-investigative-journa

> "We are just scratching the surface of immense opportunities to create more cross-border journalistic projects that would be unimaginable years ago, before the spread of digital communication tools that facilitate journalists located in different countries or continent to work together in an efficient, virtually costless way."

Global collaboration is made easy because high levels of competition don't exist between news organisation in different countries or across different languages. That's how the International Consortium of Investigative Journalists (ICIJ) – the organization behind Offshore Leaks – can work on its investigation across the globe, asking 110 reporters to work on their own countries. The publication of their findings, on the same day, helped to create a large impact and offer a wider range of coverage on the issue in question.

Collaborative work can be difficult when there is not any material, as Vadim Makarenko explained in his report about the supplement Europa published by five media organisations across Europe[16]. The collaboration is made much easier when journalists can work around a single database, accessible from anywhere.

d. **Database investigation**

Digitalisation allows a lot of documents to be made accessible in a database format. If all the files written by PricewaterhouseCoopers (PwC) from the Luxembourg leaks investigation published in October by ICIJ and its partners were printed, they would fit into thousands of boxes whereas, scanned, they could be taken out of the company in a small USB key. Similarly, WikiLeaks used a secure internet connection to transfer hundreds of thousands of files and Edward Snowden was able to download similar amounts.

Sometimes, the first task for the computer-savvy journalist is to make his or her way through the data. To work with the Snowden files, James Ball, a Guardian journalist, developed a search engine to dig into the thousands of documents.

[16] Vadim Makarenko (2011), "Europa : changing the way Europe is reported", Reuters Institute for the Study of Journalism

Having the source material in a database format is a boon for global collaboration because the same data cleaning, storage and analysis can be used on multiple investigations. In her article "Call to arms to databases researchers", Sarah Cohen explains[17]:

> "The original story may choose to focus on a particular time, location, entity, or way of looking at data. But given the data source and the query, we can generalize the query as a parameterized template, and try other instantiations of it on the data to see if they lead to other stories."

ProPublica, a non-profit news organization often works with other media organisations, mostly regional media, to expand the audience for its investigations. On their website they publish a recipe of tips to help the local reporter write a version of a national investigation for their own town, state or region. Eric Umansky explains that their main goal is not to get a lot of hits on their story but to get the story to the people it affects. "The best way to do that is not to do one story on your own, it's to provide other news organizations the tools so that they can write their own stories", he said after the publication of both a story and a recipe about restraints in US schools[18].

A set of stories about farm subsidies in Europe followed a similar trajectory. After finding out how Common Agricultural Policy (CAP) money was spent in Denmark, Nils Mulvad helped journalists from other countries to collect the same data. He also taught other journalists how to analyse this data and turn it into stories for their countries. He then built a platform to host all the data and give the stories a European overview.

e. A need for cross-border investigation

Computer-assisted journalism and data-driven journalism are, if not new practices, still very specialised ones, done by only a few journalists. In a digital world where problems like corruption are global, cross-border investigations into databases are essential. Amplifying

[17] Sarah Cohen and al. (2013): "Computational Journalism: A Call to Arms to Database Researchers", http://db.cs.duke.edu/papers/cidr11-CohenLiEtAl-cjdb.pdf

[18] Joseph Lictherman, "How ProPublica uses a "reporting recipe" to cook up collaboration", Nieman Lab, http://www.niemanlab.org/2014/08/how-propublica-uses-a-reporting-recipe-to-cook-up-collaboration/

coverage and multiplying findings, collaboration is essential and computer-assisted reporters can play a key role. By now looking at three recent cross-border investigations on farm subsidies, migrants' deaths and offshore banking, we will get a better sense of the different missions of the computer-assisted reporter, from data analysis to project management.

4. Case study 1: farmsubsidy.org

Starting in 2005, a collaborative investigation by journalists from across Europe uncovered previously unknown data and information about the Common Agricultural Policy (CAP). They used 'wobbing' to access the subsidies given by Europe to farmers in member states and build a network to gather all the data in a single place.

a. Origin of the project

The CAP was launched in 1964 to improve agricultural productivity. Twenty years later, the goal is more than reached: production exceeds the need. The main mechanism CAP used to achieve this goal was by subsidizing farmers directly. Other strategies tried to address market subsidies, but as long as the farmer fulfilled the requirements of "look[ing] after farmland" and meeting certain standards, they received money[19].

This approach was radical at the time and the program became one of the most expensive policies in the European Union: costing around 50 billion euros a year. While journalists and activists in some countries had managed to investigate the program in order to find how the money was allocated, in 2008, *La Tribune*, a French economic publication characterized the recipients' names in that country a state secret[20].

b. "Wobbing" the data and analysing it

It took almost one year for the Danish journalist Nils Mulvad and his colleague to receive the file he asked for[21]. Using the Danish Freedom of Information law, they requested the data but had to prove it existed in a certain format in order for officials to acknowledge they could transfer it. In April 2004, they published the first story detailing CAP benefactors in Europe, revealing that

[19] European commission, "A partnership between Europe and farmers", http://ec.europa.eu/agriculture/cap-overview/2012_en.pdf
[20] "Secret d'Etat", La Tribune, September 26th 2008, http://www.latribune.fr/journal/archives/editorial/200511036hrs8m/secret-detat.html
[21] interview with Nils Mulvad, November 2014

four Danish ministers were receiving farm subsidies, and that a Swedish company received far more than everyone else – around 100 million euros a year[22].

Some months later, in the United Kingdom, Jack Thurston teamed up with *The Guardian* journalists to request and publish the same information about British farmers. Their investigation revealed that the Queen and the Prince of Wales were among the major landowners "receiving the largest subsidies from the taxpayers"[23].

Efforts to analyse the data for the entire EU were initially stymied: Brigitte Alfter, a Danish-German journalist correspondent in Brussels, asked the European commission for EU-wide data. After repeated requests, the EU denied her request citing technical difficulties. She then focused on Germany and published a story about German farmers, revealing again that former nobility was among the top recipients. After these efforts, the three journalists decided to work together to go after the more comprehensive data at a European level.

Alfter, Mulvad and Thurston developed a network, mainly composed of journalists, around the farm subsidies data. Using local Freedom of Information laws in each country, they asked for the dataset. The law and the level of transparency differ in each country. Sometimes, only company names were given, other benefactors being redacted in the name of privacy.

The network operated as follows: one journalist would "wob" the data in his or her country, then publish the story and share the data on farmsubsidy.org. Others could then access all the subsidies acquired so far and perform a broader analysis.

As the main editorial work is done separately, the main role of the computer-assisted reporter, Nils Mulvad was to create the tools and define methodology. He would also help on the cleaning of the data. "The data was often very dirty, on a massive PDF file. He knew how to scrape and reconcile the data", Alfter said. The data on PDF had to be transformed into exploitable data, such as a spreadsheet.

---

[22] Brigitte Alfter, Nils Mulvad, and Jack Thurston, "Digging the dirt on farm subsidies" in "Bursting the Brussels Bubble", http://www.alter-eu.org/sites/default/files/documents/bursting-the-brussels-bubble.pdf

[23] David Hencke and Rob Evans, "Royal farms get £1m from taxpayers", http://www.theguardian.com/uk/2005/mar/23/eu.freedomofinformation1

To work on the data, the team of journalists from the whole Europe gathered together. In May 2010, they "locked" themselves up in a room in Brussels to wob, analyse and "network to find the European or cross-border aspects of the materials", said Alfter[24].

c. Publication and outcomes

The publication of the stories was done following the data availability. Even some work was done in collaboration, as the fact that the subsidies had to be obtained in each country didn't allow them to publish every story at the same time. Although some cross-border stories were produced, most of them applied to a single country. The data was gathered on the website and the stories would be collected on the media coverage sections.

But the network failed to be the main resource on the issue. In France, for example, farm subsidies received great coverage in *Le Parisien* in May 2009[25], when the paper found that one of the biggest recipients was the Prince of Monaco. Even though the findings were made after the publication of others stories in Europe and would follow the same trends – industrialists and noble families would be on the top of the list of recipients of the subsidies – the European aspect was not present, nor was farmsubsidy.org mentioned.

The creation of a single website and an annual meeting of the journalists working on harvesting the data helped build a network of European journalists and activists. Even though farm subsidies were still in their mind, the network grew bigger. The annual general meeting was re-named Dataharvest and then DHplus since 2014. As previously discussed, networking and peer-to-peer learning were essential aspects to the data-driven journalism community, and farmsubsidy.org helped build the European network of data-driven journalism[26].

One of the goals of the Journalismfund.eu, the foundation organising the event, is to promote cross-border investigations. Brigitte Alfter, chairwoman of the board explains that "there is so much data that are the same for all of us on important policies" in Europe that journalist should "share the data". Publications don't need to worry about competition because although the "basic

[24] Brigitte Alfter, "Follow the subsidy-money", http://blogs.euobserver.com/alfter/2010/05/05/follow-the-subsidy-money/
[25] Emeline Cazi, "Ceux qui profitent du pactole de Bruxelles", Le Parisien, May 28th 2009
[26] History of Dataharvest, http://www.journalismfund.eu/what-dataharvest

analysis is the same, the way of telling the story is different", she said[27]. The farmsubsidy.org project is now maintained by the Open Knowledge foundation.

[27] interview with Brigitte Alfter, November 2014

5. Case study 2: OffshoreLeaks

Using leaked information, the International Consortium of Investigative Journalists (ICIJ) with the help of more than 110 journalists revealed some hidden facts about offshore companies and the names of politically vulnerable people – entrusted with a prominent public function – who were using such companies for money laundering or for avoiding tax in their countries.

a. Origin of the project

Gerard Ryle, an Australian journalist, now head of the ICIJ, received a hard drive containing 260 gigabytes of information, concerning more than 100,000 companies operating under tax havens. The information came from two companies: Portcullis Trustnet and Commonwealth Trust Limited which specialised in providing help to set up offshore companies.

The ICIJ usually hires journalists on a per-project basis, but for the OffshoreLeaks investigations, they decided to form a partnership with media organisations. They undertook this strategy because they had more leverage - they already possessed the source material to convince the journalists. Mar Cabra, a journalist working for ICIJ, thinks that involving both the newspaper and the journalism at the beginning is a good strategy as it secures publication, as journalists feel it's their story.

The project faced two big challenges. The consortium first had to organise and index all the data. They also had to manage the work of all the partners as well as the communication between them. By its size, ICIJ said that the Offshore Leaks investigation was the biggest ever involving more than a hundred journalists from 47 different countries.

a. Data cleaning and analysis

The first step of the investigation was getting the data into the right format. The hard drives contained 2.5 million files of account details, passports and emails, but 40 percent of the files were duplicates. During the whole investigation, no less than fifteen journalists and programmers worked on the data aspect, representing approximatively 10 percent of the total number of journalists working on the investigation. Their main goal was to organize the files and help the

journalists analyse them. A lot of other tools were conceived and programmed, to find duplicates, to match names and addresses or to reconstruct some particular database.

The next step, taken by two New Zealand journalists, was to manually assign at least one country to each file to get a sense of the work for the journalist in each newspaper. It "provided an initial look at the scope and range of clients"[28] and helped figure out which countries the ICIJ journalists should focus on and which media outlets should be partnered with.

The next step was to understand the whole picture. Working with Optical Character Recognition (OCR) software, they managed to identify several names on passports, letters or contracts. They also pre-indexed the names using NUIX, a free-text retrieval" software for easy keyword searching across a group of documents – similar to an online search engine.

The goal was to build a structured database, with names of benefactors and companies linked together. So, a tool was developed to look into the names and the files. The database was complex to use and because of the sensitive nature of the data, the ICIJ decided "not to give the data to everybody, but only to one person that the ICIJ trusted" explained Mar Cabra[29]. She is a Spanish data journalist and was responsible for finding the data for each journalist[30], looking through the database.

Because Cabra was quickly overwhelmed, the team had to find another solution. Programmers in Costa Rica and in the United Kingdom worked on a database allowing anyone to access the files. More than 28,000 online searches were made and 53,000 documents were downloaded at the end.

a. **Publication and outcomes**

One strategy Offshore Leaks followed, which had not been employed by farmsubsidies, was coordinating publication dates internationally. This coordination has the potential benefits of

[28] Duncan Campbell, "How ICIJ's Project Team Analyzed the Offshore Files", http://www.icij.org/offshore/how-icijs-project-team-analyzed-offshore-files

[29] interview with Mar Cabra, November 2014

[30] Anne Michel, "Comment "Le Monde" a enquêté sur le scandale des paradis fiscaux", Le Monde, April 4th 2014, http://www.lemonde.fr/economie/article/2013/04/04/revelation-sur-le-scandale-des-paradis-fiscaux_3153318_3234.html

"creating a wave" says Cabra. So, on April 4th, 2013, the stories were published at the same time across the globe.

Choosing a perfect publication date is a complex process especially when partner media outlets span different time zones. To make matters more difficult, the weekly magazines, newspapers and broadcast network shows might not be publishing on the same day.

Ahead of publication, the ICIJ team and the partner newspapers made their stories available to each other. In that way every journalist was aware of the main findings and could also choose among all the papers some to be translated and published in their own newspaper.

Several weeks after the publication, in June 2013, the ICIJ chose to publish a platform allowing readers to search names of persons and offshore companies[31]. The website does not publish the whole dataset, but rather allows access to the structured database, "a careful release of basic corporate information", so that interested readers can look for politically vulnerable people and eventually "strip away secrecy"[32]. "ICIJ believes many of the best stories may come from its readers when they explore the database", explains Marina Walker Guevara, ICIJ deputy director. This disclosure of the information echoed the open-source culture and the aim for transparency mentioned in the first part of this paper.

The Offshore Leaks investigation was the largest project of its kind at the time. Cabra explains that the ICIJ learned a lot from it and it helped them with their next investigation on China and Luxembourg: a single publication date, the publication of the database, and the tools built to discuss online and organise data.

In November 2014, the organisation published the result of a long investigation into tax avoidance schemes made by PricewaterhouseCooper (PwC). They simultaneously published not only the stories but also the complete database, allowing readers and others journalists to investigate the PwC files. After one week, out of the 1.6 million visits the website received,

[31] ICIJ Offshore Leaks Database, http://offshoreleaks.icij.org/search

[32] Marina Walker Guevara, "ICIJ Releases Offshore Leaks Database Revealing Names Behind Secret Companies, Trusts", June 14, 2013, http://www.icij.org/offshore/icij-releases-offshore-leaks-database-revealing-names-behind-secret-companies-trusts

600,000 were for the interactive giving access to the documents. This helped other media to find new tax rulings and illustrate the LuxLeaks stories with their own examples.

The Offshore Leaks investigations and those that followed inspired ICIJ to build secure tools to help journalistic collaboration. They used open-source tools to build two platforms: one to search the data and one to ease communication between journalists on a forum-like website. The tools behind the two website will be merged and developed in 2015 thanks to a Knight Foundation grant[33].

[33] Hamish Boland-Rudder, HICIJ to build a global I-Hub, a new secure collaboration tool, July 17, 2014, http://www.icij.org/blog/2014/07/icij-build-global-i-hub-new-secure-collaboration-tool

6. Case study 3: the Migrants Files

A team of data-journalist across Europe used publicly available information to build an exclusive database of all the migrants' deaths at European Union's borders to analyse the dangers of different migration routes.

a. Origin of the project

The European Union sees millions of migrants entering its border each year. But not all migrants arrive safely; thousands die on their way. Although the European Agency for the Management of Operational Cooperation at the External Borders – also called Frontex – is charged with maintaining border security, they consider migrants' deaths outside their jurisdiction. "Frontex's work is the fight against illegal immigration, not their rescue. And those persons are dead, they are not migrants anymore", explained one European official[34].

In order to shed light on this area, a team of a dozen of journalists decided to measure the number of migrants dying who try to come to Europe, and to analyse their routes. Their goal was to build a single database for all such deaths across Europe, with the help of previous work done by the OWNI organization, which in 2011 had used data from an organization called United against Racism amongst others.

b. Data collection and cleaning

The team used open-source intelligence to build the database. All the information was available but spread across a variety of data sources from news report to spreadsheets maintained by NGOs. The main task was to gather the information, clean it and construct a single database. At the beginning, the journalists worked mainly on data standardisation, checking for duplicates and verifying incidents.

[34] Jean-Marc Manach, 31 March 2014, Le Monde Diplomatique, http://www.monde-diplomatique.fr/carnet/2014-03-31-morts-aux-frontieres

As they wanted to map the deaths, they used geo-location to add to each record geographical coordinates. The geographic information had to be the most precise possible to be automatically geo-located.

Although the team was comprised of data-savvy journalists, their work did not require technical skills, per se. “It was regular journalistic work”, says Nicolas Kayser-Bril, journalist and project manager.

However, one skill that was required was collaboration. The international nature of the team helped for instance, in identifying stories of migrants who died on their way to Greece by working more easily with the appropriate Greek agency.

c. Publication

On April 2th 2014, the team published the database and the articles on nine websites across Europe. They could assess with the help of the database which migrations routes were the most dangerous and how the migration policy affected the popularity of each route.

To publish the database, the team used Detective, a tool developed by Journalism++ to help journalists make their own databases. Even though the tool is designed to help journalists through the whole investigation, they used it only to publish the database. It was complicated to use it before as the tool makes it difficult to prevent duplicates in complex data, such as an event’s description.

The investigation proved that leaked information wasn’t required to carry out investigations at a European level. The use of OSINT allowed the team to build a new database with new findings. The database and the number of dead migrants number was used by other organisations such as the International Office for Migration or Amnesty International, mainly because each event was fact-checked with the source and because it was the most complete database available.

7. Conclusions

“There is nothing that journalists like more than learning to code than to collaborate with each other”, said Emily Bell with some sarcasm during the Reuters Memorial lecture in November 2014. Despite the difficulties inherent in international collaborations, when done well, the results can create a larger impact than the work of any one organisation and contribute to the public good (Parasie and Dagiral, 2012). In that context, computer-assisted reporters are a great resource in investigation, particularly cross-border ones.

The computer-assisted reporter plays an important role analysing the data and organising the team. He or she has to make sure the data is accessible, navigable and readable to other journalists, sometimes by developing a custom interface or search tool. He can also organise the data and enhance it. But by being sometimes the only person to master the data, he becomes as well a project manager, assigning tasks and helping the story come out.

Global investigations are needed because crimes and crises are global too. Although collaboration is a great tool for better coverage, because of the impact of simultaneous publication and the wider breadth of coverage possible, the involvement of news websites could become more complex as newspapers tend to go more global and increase the competition between media that previously could work together.

The simultaneous publication of the stories but also, if possible, of the database, and the source code all help the investigation to multiply and to uncover new stories.

The collaborative culture is not rooted in journalism practice. But the computer-assisted reporting techniques often require the journalists to collaborate, either with the readers or with other journalists, and the culture is shifting. “Cross-border journalism often develops on empirical stage”, Alfter said. “We need to systematise its spread by teaching it in journalism schools”, she believes, as she is working on such program.

From my research, one reason why collaboration might have worked is because the investigations revolved around a database. The data serves as common ground for the

investigation and helps build the stories. As we've seen, data can multiply and journalists can share their recipes. The stories and findings can be adapted to different languages and cultures, so that it can have a bigger impact. The database also makes it easier to find new stories, even after publication. It can be easily updatable and become a source for NGO or other media.

Offshore Leaks and Migrants Files proved that a strong project management, by people understanding the data well, is crucial in carrying out a big investigation of this kind. The data-journalist can assume this role by being the main contact with other journalists who are not so good with the data.

The three investigations discussed above were also the occasion to build or consolidate a network. ICIJ used it again for its next investigations. The Farm subsidies project turned into a yearly event that reunites journalists and ‘wobbing’ activists from all over Europe and is part of a foundation that funds cross-borders projects. And the Migrants files journalists are still connected and have worked since on different projects.

The next years will see more global collaborations, certainly around databases. Even though it is possible to imagine collaborations without data, the databases make it easier and on a bigger scale. The challenge will be to find enough collaborating media when the trends tend to concentration and newspaper sites become more global.

Bibliography

Coddington, M., “Clarifying Journalism’s Quantitative Turn”, *Digital Journalism*, 2014

Gray, J., L. Bounegru and L. Chambers, *Data Journalism Handbook*, O’Reilly, 2012

Howard, A.B., *The Art and Science of Data-driven Journalism*, Tow Center for Digital Journalism, 2013

Fink K. & C. W. Anderson (2014): “Data Journalism in the United States, Journalism Studies”

Leigh, D., L. Harding, *WikiLeaks: Inside Julian Assange's War on Secrecy*, Guardian Books, 2011

Meyer, P., *Precision journalism: a reporter's introduction to social science methods*, Indiana University Press, 1979

Parasie, S. and E. Dagiral, “Data-driven journalism and the public good”, *New Media Society*, vol. 15, 2012