PANAMA PAPERS: A CASE STUDY FOR RECORDS MANAGEMENT?

Marie-Anne Chabin

Archive17, 80 rue Saint-Denis, 75001 Paris France, mac@marieannechabin.fr

Abstract

The international financial and political crisis named Panama Papers (2016) provided a rather good material for information studies, particularly for digital diplomatics. First, the comments, in some languages, allow to analyze how media and newspapers use the words information, data, document, file and record, each with its own culture: some papers are directly written with the own words of the journalist, others are translations, showing notably the differences between English and French professional and common vocabulary.

Second, it was also an opportunity to check the place of email in those cases of disclosure of sensitive data in the "society of information". The point is that the fragility of the digital files themselves (format, preservation) is not the worse for information and records management; the risk is more linked to the fragility of the international network and the ability of search engines to reach these records, wherever they may be stored.

Keywords: Panama papers; Records management; Terminology

1 Introduction

The findings of the so-called "Panama papers" generated a lot of comments. I have been interested first as much as anybody else around the world. Then I went further and looked for more information about the data processing within the investigation: records management terminology, search engines, information metrics and email.

2 Records management terminology

My first point deals the words used, both in English and in French, to describe the documentary material collected for this investigation. I went through the original paper posted by ICIJ (International Consortium of Investigative Journalists) on the website (1).

From the beginning, I wondered why it was called "Panama papers" and not Panama Leaks or Panama Gate? The explanation was given there: it is the name chosen by ICIJ, most likely because the term "papers" is shared with lawyers.

The ICIJ paper is rather long (near 6000 words). The findings of the investigations are displayed factually [French medias are known to mix easily facts and comments]. The material of the leak is mainly described with the three following terms: document (used 36 times), record (31 times) and file (27 times): *A massive leak of **documents** exposes… The cache of 11.5 million **records** shows… The **files** reveal… The leaked **records** come from… The **files** contain… The secret **documents** suggest…*

In this digital era where the use of the term" document" declined in favor of "data", I was a little bit surprised to find out that information and data are less used (10 and 5 times respectively) for example in theses sentences: "The information from this unremunerated whistleblower documents transactions as far back as the 1970s and eventually totaled 2.6 terabytes of data" and "The data includes emails, financial spreadsheets, passports and corporate records"; *Les Echos*, inspired by the *Süddeutsche Zeitung,* shows the documents types as follow:

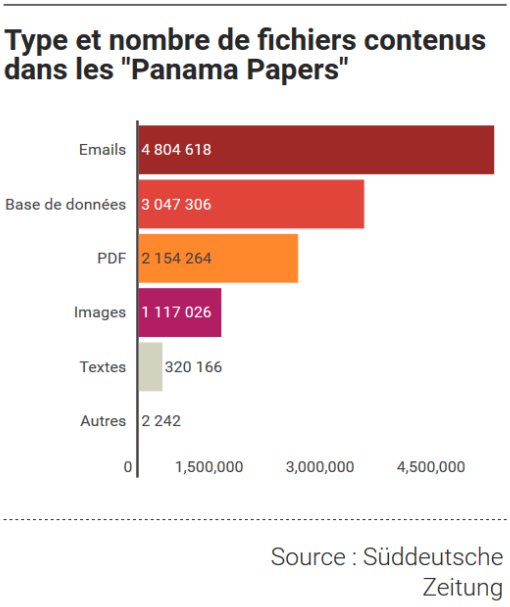

Figure 1. *Types and number of files in Panama Papers (2)*

Source: Süddeutsche Zeitung (2016)

As a French records management theoretician, I looked methodically through the French speaking online

newspapers to find out which words have been chosen by those medias (near twenty articles) for "document", "record" and "file" translation. The results are: the word "document" predominates but "*fichier*" and "archives" are also regularly used:

> "Mes collègues et moi avons analysé des archives contenant 11,5 millions de documents internes provenant des fichiers de Mossack Fonseca (3) […] La fuite contient la quasi-totalité des archives et données informatiques d'une seule et même société" (4) […] des millions de documents confidentiels, baptisés "Panama Papers" (5) […] 11 millions et demi de fichiers, les archives du cabinet, depuis 1977 jusqu'à 2015. Plus de 2,6 téraoctets de données (6) […] millions de documents et données provenant des archives de Mossack Fonseca (7) […] près de 11,5 millions de documents, piratés dans les archives (8) […] 11, 5 millions de documents des archives du cabinet panaméen Mossack Fonseca (9)".

My two comments are:

Fichier means file. What is obvious here is that the English word has two different translations in French: the old word "*dossier*" points to the records arranged as a unit in a file/folder; and the IT word that is a computer file (a resource for storing information, which is available to a computer program and is usually based on some kind of durable storage, according to Wikipedia). It was strange to note that the word "*dossier*" is rarely used. If the leak had started in France, no doubt that the word "*dossier*" would have been used more, in addition to "*fichier*", because the expressions "business file" and "customer file" are normally translated by "*dossier d'affaire"* and "*dossier client*", which are also translated in English by "business record" and "customer record".

The French word "*archives*" is normally used by journalists to name the Mossack Fonseca records as a collection of documents and everything is very clear for everybody. The English word "archives" does not appear in the English text because, in English, archives means historical records and it is not history the journalists are taking about! But in French, the word "*archives*" always had and still have the meaning of "records". This fact shows that the recommendations of the French Standardization Organization (AFNOR) regarding the translation of records are out of the real life and inoperative. French *archives* and *archivage* DO NOT FIT with English archives and archiving.

3. The classification scheme is dead; long live the search engine!

There are some papers which tell more about the working methods of the consortium for the investigation (*le Monde or la Dépêche*) (10).

The role of technologies in the leak and the disclosed information is quite clear: fuzzy search mechanisms, OpenRefine software for data cleanup, *Neo4j*, *Linkurious* software graph-visualization.

As I often say to my clients and students, explaining records management methods and tools, classification scheme for research is out of date; keywords selected manually, metadata added by users and so on was relevant at the end of the last century. The power of search engines is so huge that spending months and years to built classification schemes and thesauri is nowadays useless. The machine sorting (in numeric or alphabetic order, or with any preselected criteria) is faster than human being and is more reliable. A huge step has been reached.

Nevertheless, the efficiency of these tools and algorithms is linked to the quality of the material, to the quality of the data. This quality of data refers, from my point of view, to two major concepts in records management, archival science and diplomatics: structure and completeness.

If a mass of data is processed and mined in to provide evidence (records are first created and retained for evidence), the best the record group or collection is structured, the more reliable are the results. This structure depends on some basic elements inherent to any recorded transaction and activity: a date, a place, one or more actors and a verb to tell the nature of the transaction: to buy, to look at, to come in, to go out, to allow, to refuse, to inform etc.

From this point of view, the record collection of the Mossack-Fonseca firm is well structured, as the records in any law firm are, in a logical, hierarchical and rigorous manner. The register/list of customers acts as the spine of the whole, from which start customers files/records made of both legal documents (corporate deeds) and, in chronological order, series of operations (financial or anything else) and mail/emails exchanged between the firm and each customer.

If the files came from a panel of different places and firms, related to many different topics the list of which is unknown, inquiry will be more difficult or, in other terms, the results of the inquiry will be more difficult to understand and to check.

So, from a qualitative perspective, the leap allowed by technology is not so significant than it is from the quantitative aspect. Technology is amazing, of course; what I mean is that human/manual analysis and algorithm are not to be set against; there not at the same level. They are more complementary than opposite.

The point is how you control the context of information. Traditionally, the understanding of context is provided physical contact with the records, the possibility to appreciate in a look the shape and the nature of thousands and millions of documents, arranged in file folders or boxes, or just lying on the shelves. The know-how of the expert or of the researcher allowed

him/her to capture visual and sensorial clues within a couple of seconds and to evaluate what is the "density" of the records, how they have been created and accumulated, if a part is missing, if the whole is reliable or not.

My feeling (a lack of experience with search engines?) is that this kind of context information is lost when dealing with a huge amount of data (and it is worse with what is called "big data"), lost and not yet replaced by some equivalent technological tool. I guess policemen have the same feeling when comparing the forensic science and the old fashioned practices of those police commissioners looking at human being's behavior before at physical evidence.

Regarding completeness, I want to underline that the concept of completeness should be applied to the resources before to be applied to the results. If you do not control the completeness of the record collection, how can you be sure of the quality of the result? If the "Panama papers" collected and analyzed by ICIJ fit with the complete records of Mossack Fonseca, the fact that the name of a politician does not appear in the results means that this person is not a customer of this firm. But it does not mean that this person is not the customer of another exotic law firm. This is obvious in a way but it is useful to remind it and some newspapers should be more careful in their conclusions. The record group concerned by the leak is complete; the results as the list of offshore companies and fraudsters are not complete.

4. A volume of data to give vertigo to some people

The third point of this analysis is related to the volume of data involved and to the metrics for information volume.

Some journalists (I mean commentators, not investigative reporters) and those who read newspapers have a feeling of vertigo with the figures given by the Panama papers: 11,5 million files/documents (4,8 million emails, 3,1 million spreadsheets, 2,1 millions PDFs etc.), 2,6 TB of data.

Some comments let you think people are very impressed by the volume of information, speaking of decades necessary to go through all the data, of "peeling" the files, comparing the 2,6 TB with an incredible number of TV series:

> "Le journal allemand possède son lot d'experts du data-journalisme, mais à lui tout seul, il lui aurait fallu des dizaines d'années pour éplucher l'entièreté des données collectées" (11) […]"… selon plusieurs médias ayant épluché les fichiers provenant des archives du cabinet d'avocats basé au Panama" (12) […] 2,6 tera octets, soit l'équivalent de 34 665 épisodes de séries télévisées ! (13)".

But information professionals, whose job is to deal every day with business information, records, documents, files, emails records and data created, received and managed in public or private organizations, as evidence and information, are not surprised by those figures! The volume of the leak is impressive; the volume of data is not!

Such comments in front of 2,6 TB of data show that digital technologies are changing the way human beings appreciate data, that digital technologies atomize information; that digitalization increases the granularity of information. And the society has not yet formulated the proper metrics to appreciate digital information.

Let us think about it. There are 240.000 companies concerned in the Mossack Fonseca records made of 11,5 million documents. It means that a file contains in average 48 documents: no surprise here. Of course, some files are very thin while others have extended to 3.000 elements, *Les Echos* reports; nothing extraordinary either (14).

Further. If an electronic file is equivalent to a paper document of 3 or 4 pages (normal hypothesis), the Mossack Fonseca records contains 40 million pages. In average, a storage box (10 cm wide) contains 800 pages; so, 40 million pages require 5 linear kilometers of paper storage. Mossack Fonseca firm has 500 employees; that means two filing cabinets near every employee should be enough to store all the records in paper form from 1977. Nothing special again.

Imagine these 5 linear kilometers or records are dispatched between the 376 journalists involved in the investigation; each journalist would be in charge of going through 14 linear meters (two large filing cabinets); that is not unmanageable. The main difficulty lies in sharing findings and crossing results with data from other databases.

Without relevant indicators to measure the equivalence between the information object and the potential knowledge inside, it is easy to give the way to fanciful comments. Personally, I meet every day organizations dealing with 2 TB of data, and more than that.

What is impressive is not the volume of data in the leak but the fact they are shared out of the originated addressees and managers, due to the power of digital technologies. Maybe we speak about data too much and not enough about networks.

5. In the beginning was electronic mail

I was wondering how the leak started. It began with the hack of email server at Mossack Fonseca; the emails lead to attachments, and that makes a lot of information (at least since the last twenty years). I am interested in learning more but I could not find anything else until now (15).

Email is actually symptomatic of information governance in the organizations. Governance is needed because of the risk of non-governance. The risk comes often from the lack of control in emailing, both regarding content and security of transactions. That is exactly the message of the CR2PA, French club for records management, through its publications, manifestations and MOOCs (16). But the behavior of executives and managers, facing information governance issues, remind more cicadas than ants as the French poet La Fontaine depicted them three centuries ago (17). The cicadas missed everything when winner came while the ants had anticipated the troubles to come.

6 Conclusion

This paper aims to demonstrate that information management is specific discipline that requires a specific and clear metrics (to be adapted to digital era); that also must reconcile concepts and methods of diplomatics, archival science and records management with information technologies (networks, cloud, search engine); and that will failed if the machines, whatever their power and utility, are considered more important than human beings.

In the Panama Papers leak, the protagonists are (honest) journalists pursuing (supposed) fraudsters, in the name of the common good. The point is that technologies and methods would be the same if, on the other way round, honest citizens were pursued by dishonest people.

Let's pay attention to this exciting digital world!

Notes

(1) https://panamapapers.icij.org/20160403-panama-papers-global-overview.html

(2) http://www.lesechos.fr/monde/enjeux-internationaux/021815669450-panama-papers-les-coulisses-de-lenquete-1211437.php

(3) http://rue89.nouvelobs.com/2016/04/08/enquete-mossack-fonseca-cabinet-panama-papers-263707

(4) http://www.lematin.ch/monde/Une-enorme-fuite-revele-les-dessous-du-monde-offshore/story/20889986

(5) http://lci.tf1.fr/monde/evasion-fiscale-une-affaire-planetaire-de-fraude-revelee-8730174.html

(6) http://www.rtl.fr/actu/societe-faits-divers/panama-papers-11-5-millions-de-fichiers-du-cabinet-mossack-fonseca-a-explorer-7782670471

(7) http://www.lemonde.fr/panama-papers/article/2016/04/06/panama-papers-le-cabinet-mossack-fonseca-assure-avoir-ete-pirate-de-l-exterieur-et-a-porte-plainte_4896575_4890278.html

(8)http://tempsreel.nouvelobs.com/monde/20160403.OBS7744/c-est-le-plus-gros-scandale-d-evasion-fiscale-de-l-histoire-que-sont-les-panama-papers.html

(9) http://www.lesechos.fr/tech-medias/medias/021814397453-la-finance-offshore-mondiale-cible-des-revelations-des-panama-papers-1211183.php

(10) http://data.blog.lemonde.fr/2016/04/08/panama-papers-un-defi-technique-pour-le-journalisme-de-donnees/ & http://www.ladepeche.fr/article/2016/04/07/2320269-panama-papers-linkurious-start-up-aide-eplucher-donnees.html

(11) http://www.vanityfair.fr/actualites/international/articles/-panama-papers-comment-les-journalistes-ont-decrypte-la-plus-grande-fuite-de-lhistoire/37123

(12) http://www.nicematin.com/faits-de-societe/panama-papers-le-nom-de-dmitry-rybolovlev-apparait-dans-le-dossier-38586

(13)http://www.lesechos.fr/monde/enjeux-internationaux/021815669450-panama-papers-les-coulisses-de-lenquete-1211437.php

(14) http://www.lesechos.fr/monde/enjeux-internationaux/021815669450-panama-papers-les-coulisses-de-lenquete-1211437.php

(15)http://www.theregister.co.uk/2016/04/05/email_server_hack_led_to_mossack_fonseca_leak/

(16) http://blog.cr2pa.fr/

(17)http://www.frenchtoday.com/french-poetry-reading/poem-la-cigale-et-la-fourmi-la-fontaine-audio