RECORDS, TERMINOLOGY AND TECHNOLOGY IN THE PANAMA PAPERS

Oydin Berdiyeva

Junior Research Fellow at the University
of Journalism and Mass Communications of Uzbekistan

Abstract: The Panama Papers leak of April 2016 exposed 11.5 million internal records of the Panamanian law firm Mossack Fonseca, covering the period 1977–2015 and totalling 2.6 terabytes. Beyond its political and financial consequences, the leak is a rare, fully documented example of an organisation's complete record collection being processed, searched and interpreted by outsiders. Building on the case study proposed by Chabin (2017), this article asks what the Panama Papers can teach records management (RM) about terminology, retrieval, information metrics and email governance.

Keywords: Panama Papers; records management; archival terminology; diplomatics; information metrics; email governance; search engines; data leaks

1. Introduction

On 3 April 2016, the International Consortium of Investigative Journalists (ICIJ) and more than one hundred partner media organisations began publishing stories based on a leak that became known as the Panama Papers. The source was the internal record collection of Mossack Fonseca, a Panama-based law firm that had created and administered offshore companies for clients around the world. The leak was delivered by an anonymous source, calling themselves "John Doe", to journalists Bastian Obermayer and Frederik Obermaier of the German daily *Süddeutsche Zeitung*, who then shared it with ICIJ (Obermayer & Obermaier, 2016). According to ICIJ, the material comprised 11.5 million records and 2.6 terabytes of data and was analysed by 376 journalists for about a year before publication (ICIJ, 2016).

The consequences were immediate and far-reaching. Iceland's Prime Minister Sigmundur Davíð Gunnlaugsson resigned within days of the first stories. In Pakistan, the leak triggered court proceedings that ended with the disqualification of Prime Minister Nawaz Sharif in July 2017. In the United Kingdom, Prime Minister David Cameron was forced to explain his holdings in an offshore fund set up by his late father. Mossack Fonseca itself announced its closure in March 2018. ICIJ later reported that governments had recovered more than 1.3 billion US dollars in taxes and penalties linked to the investigation. The project received the Pulitzer Prize for Explanatory Reporting in 2017.

Most academic writing on the Panama Papers belongs to economics, law, political science and journalism studies. It examines tax avoidance, the offshore industry, beneficial ownership and the ethics of using leaked material (e.g. Zucman, 2015; O'Donovan, Wagner & Zeume, 2019). Much less attention has been given to what the case reveals about records themselves: how they are named, how they are found, how their volume is measured, and how they escape their creators. This gap is surprising. The Panama Papers are, at bottom, the complete archive of a professional services firm. They were created as evidence of transactions, kept for legal and business purposes, and later mined by outsiders with the help of advanced search technology. Few events illustrate so clearly the life

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR

Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

cycle of records, from creation to unintended disclosure.

The French records management theorist Marie-Anne Chabin was among the first to treat the leak as a case for information science. In a short paper published in the *Brazilian Journal of Information Science*, she raised four questions: which words the media used for the leaked material; whether search engines have made classification schemes obsolete; why the published volume figures caused such public astonishment; and what role email played in the leak (Chabin, 2017). Her paper is deliberately brief and essayistic. It opens lines of inquiry rather than closing them.

The present article takes up Chabin's four questions and develops them in a structured, empirical form. It extends her English–French comparison to Russian and Uzbek, two languages that matter for records professionals in Central Asia; it recalculates her volume estimates and tests them with alternative assumptions; it compares the Panama Papers with other major leaks; and it adds a wider set of examples on search technology and email governance. The research questions are:

- **RQ1.** How do professional and media vocabularies in English, French, Russian and Uzbek name the material of the Panama Papers, and what do the differences reveal about records management terminology?
- **RQ2.** How did search and analysis technologies change access to the Mossack Fonseca records, and what is the remaining role of structure, context and completeness?
- **RQ3.** How can the volume of the leak be expressed in metrics that are meaningful to non-specialists and to records professionals?
- **RQ4.** What does the role of email in the leak suggest about information governance in organisations?

2. Materials and Methods

2.1 Research design

The study is a single, instrumental case study in the sense of Stake (1995) and Yin (2018): the Panama Papers are examined not for their own sake but because they illuminate broader questions of records management. The design is mixed. Qualitative terminological and diplomatic analysis is combined with descriptive quantitative analysis of published figures. No leaked documents were consulted directly; all data come from published, publicly accessible secondary sources. This choice avoids the legal and ethical problems of handling leaked personal data and keeps the study reproducible.

2.2 Materials

Four groups of materials were used.

- **The ICIJ overview article** "Giant leak of offshore financial records exposes global array of crime and corruption" (ICIJ, 2016), close to 6,000 words long. This is the reference English text used by Chabin (2017) for her word counts.
- **French-language press coverage.** The corpus of about twenty online articles analysed by Chabin (2017), including *Le Monde*, *Les Echos*, *L'Obs*, *Le Matin*, *RTL*, *LCI*, *La Dépêche* and *Vanity Fair France*.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

- **Russian- and Uzbek-language usage.** Standard terminology from the Russian national standard GOST R ISO 15489-1-2019, from Russian archival usage, and from the Uzbek legal texts on archives and electronic document management (the Law "On Archival Affairs", 2010; the Law "On Electronic Document Management", 2004), together with typical phrasing in Russian and Uzbek news reports on the leak.
- **Quantitative data.** The file-type breakdown published by *Süddeutsche Zeitung* and reproduced by *Les Echos* (Figure 1 in Chabin, 2017); ICIJ's published totals; and ICIJ's or the publishing organisations' totals for eight other leaks between 2010 and 2021 (Cablegate, Offshore Leaks, LuxLeaks, Panama Papers, Paradise Papers, FinCEN Files, Pandora Papers and, for contrast, the Enron email corpus).

2.3 Methods of analysis

Terminological analysis (RQ1). Following the frequency counts reported by Chabin (2017), relative frequencies per 1,000 words were calculated for the five key terms in the ICIJ text. Equivalents in French, Russian and Uzbek were then mapped to the records management concepts defined in ISO 15489-1:2016 (*record*, *records system*, *aggregation*, *metadata*). Each equivalent was classified as (a) a direct conceptual match, (b) a partial match, or (c) a false friend.

Technological workflow analysis (RQ2). Published descriptions of the ICIJ workflow (Cabra & Kissane, 2016; Chabin, 2017; Obermayer & Obermaier, 2016) were broken down into stages: ingestion, text extraction, indexing, entity resolution, graph analysis and collaborative verification. Each stage was compared with the corresponding traditional archival function, from arrangement and description to reference service.

Metric analysis (RQ3). Chabin's (2017) paper equivalence model was recalculated step by step. A sensitivity analysis varied the two main assumptions: pages per digital file (2 to 5) and pages per archive box (600 to 1,000). Average file size and average number of records per client entity were also computed.

Comparative analysis (RQ3, RQ4). The Panama Papers were set beside other leaks by number of files, volume and the share of email, to test whether the case is exceptional or typical.

Diplomatic reading (RQ2, RQ4). Diplomatics studies the genesis, forms and transmission of records in order to assess their authenticity (Duranti, 1998). Its basic elements of a transaction (date, place, persons, action) were used as the analytical frame for judging the structure of the Mossack Fonseca collection and the evidential weight of email.

2.4 Limitations

The study relies on published counts, not on direct inspection of the leak. Word counts in the ICIJ text are taken from Chabin (2017) and may vary slightly with later edits of the online article. Russian and Uzbek examples illustrate typical usage and do not form a full corpus. Finally, the entity count differs between sources: Chabin cites about 240,000 companies, while ICIJ's searchable database lists 214,488 offshore entities. Both figures are reported where relevant.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

3. Results

The results are presented in four parts that follow the research questions. Each part combines the data with worked examples, because the value of the case lies in concrete detail.

3.1 Naming the material: terminology across four languages (RQ1)

3.1.1 Word frequencies in the reference English text. In the ICIJ overview article, three terms carry almost the whole description of the leaked material. Table 1 shows Chabin's (2017) counts converted into relative frequencies.

Table 1. Key terms in the ICIJ overview article (approx. 6,000 words)

Term	Occurrences	Per 1,000 words	Share of the five terms
document	36	6.0	33.0%
record	31	5.2	28.4%
file	27	4.5	24.8%
information	10	1.7	9.2%
data	5	0.8	4.6%
Total	**109**	**18.2**	**100%**

Source: counts from Chabin (2017); relative values calculated by the author.

The three "object" terms (*document*, *record*, *file*) account for 86.2% of uses; the two "content" terms (*information*, *data*) for only 13.8%. This is notable in a period when public discourse increasingly speaks of "data" rather than "documents". The journalists described the leak as a body of evidence made of discrete, authored objects, not as an undifferentiated stream of data. Typical phrasings include:

1. *A massive leak of **documents** exposes…* - the document as a unit of disclosure.
2. *The cache of 11.5 million **records** shows…* - the record as a unit of evidence.
3. *The **files** reveal…* - the file as a unit of storage or of a client case.
4. *The **data** includes emails, financial spreadsheets, passports and corporate records* - here *data* is the umbrella term, and *records* is one category inside it.

The last example deserves comment. In ISO 15489-1:2016, a record is "information created, received and maintained as evidence and as an asset" by an organisation. In the ICIJ sentence, by contrast, "corporate records" are only one type among emails, spreadsheets and passports. A records manager would call all of these records, because all were kept by Mossack Fonseca in the course of business. The example shows how professional and everyday meanings of *record* diverge even within English.

3.1.2 The name itself. The choice of "Papers" rather than "Leaks" or "-gate" is itself a

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

terminological decision. Chabin (2017) links it to the vocabulary of lawyers, for whom "papers" are the documents of a case. A second, journalistic lineage is just as relevant: the *Pentagon Papers* published by *The New York Times* in 1971. The name places the leak in a tradition of documents of public interest, not of hacking. Table 2 shows how the name travelled into other languages.

Table 2. The name of the leak in five languages

Language	**Usual name**	**Literal meaning**	**Records concept evoked**
English	Panama Papers	Panama papers	legal case documents
French	Panama Papers (borrowed)	-	none; English loan
German	Panama Papers (borrowed)	-	none; English loan
Russian	Панамское досье (Panamskoe dos'e); also Панамские документы	Panama dossier; Panama documents	the dossier, an aggregated case file
Uzbek	Panama hujjatlari; also Panama arxivi	Panama documents; Panama archive	the document; the archive as a whole body

The Russian name is particularly interesting. *Dos'e* (from French *dossier*) describes an organised file about a person or matter. It therefore evokes exactly the archival aggregation that Chabin found missing in French press usage, where *dossier* was rare. The Uzbek *Panama hujjatlari* translates *Papers* by *hujjat* (document), the standard legal and archival term, and so stays closer to the ICIJ wording than either French or Russian.

3.1.3 One English word, several French, Russian and Uzbek concepts. Chabin's (2017) analysis of about twenty French articles showed that *document* predominated, while *fichier* and *archives* were also common and *dossier* was rare. Typical French phrases were "11,5 millions de documents internes provenant des fichiers de Mossack Fonseca" and "près de 11,5 millions de documents, piratés dans les archives". Table 3 extends this mapping to Russian and Uzbek.

Table 3. Equivalents of key records terms (match type: D = direct, P = partial, F = false friend)

Concept (ISO 15489 sense)	**English**	**French**	**Russian**	**Uzbek**	**Match notes**

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR

Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

A single record kept as evidence	record; document	document d'activité (AFNOR); document	документ (dokument)	hujjat	D in Russian and Uzbek; P in French, where *document d'activité* is rarely used outside standards
An aggregation of records about one matter	file; case file	dossier	дело (delo)	ish; yig'majild	D, but rarely used by media in all three languages
A computer file	file	fichier	файл (fayl)	fayl	D; English *file* is ambiguous between this row and the one above
The body of an organisation 's records	records	archives	документы организации; архив	hujjatlar; arxiv	F: French *archives* covers current records, English *archives* does not
Records of historical value	archives	archives historiques	архивные документы	arxiv hujjatlari	D
Records management	records management	gestion des documents d'activité; archivage	делопроизводство; управление документами	ish yuritish; hujjatlarni boshqarish	P: *deloproizvodstvo* and *ish yuritish* include office paperwork routines
Data	data	données	данные (dannye)	ma'lumotlar	D
Information	information	information	информация	axborot; ma'lumot	P: Uzbek *ma'lumot* means both data and information

Several examples from the table illustrate concrete problems of translation.

Example 1: "customer file". In a law firm, a customer file is the full set of records for one client. French renders it naturally as *dossier client*, Russian as *дело клиента* (delo klienta), and Uzbek as *mijoz ishi* or *mijozning yig'majildi*. When media translate "the files" as *fichiers* or *fayllar*, the reader imagines computer files, not organised client cases. The structure that made the collection searchable (one file per client entity) disappears from the language used to describe it.

Example 2: "archives" as a false friend. The French sentence "les archives du cabinet, depuis 1977 jusqu'à 2015" means the whole record collection of the firm, including current records. A literal English translation, "the firm's archives", suggests old, inactive material of historical interest, which

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

is why the English ICIJ text never uses the word. Russian and Uzbek behave more like English here. Russian *архив* and Uzbek *arxiv* usually refer to the institution or repository that holds records after they leave active use. A Russian reader of "архивы Mossack Fonseca" might therefore understand historical holdings rather than a live business system.

Example 3: a third meaning in information technology. In Uzbek and Russian, *arxiv* and *архив* also mean a compressed computer file (a ZIP or RAR "archive"). A sentence such as "jurnalistlarga arxiv yuborildi" ("an archive was sent to the journalists") can therefore mean a compressed file transfer, a body of records, or a historical collection. Context alone decides.

Example 4: one ICIJ sentence in four languages. The sentence "The cache of 11.5 million records shows how banks, law firms and other intermediaries…" is typically rendered as follows:

5. French: *Les 11,5 millions de documents montrent comment…* - *records* becomes *documents*.
6. Russian: *Массив из 11,5 млн документов показывает…* (Massiv iz 11,5 mln dokumentov) - *cache* becomes *massiv* (array, mass), a data-oriented word.
7. Uzbek: *11,5 million hujjatdan iborat to'plam … ko'rsatadi* - *cache* becomes *to'plam* (collection), an archival-sounding word.

In each language the evidential sense of *record* is replaced by the more general *document*, while the word for the whole body shifts between data vocabulary and collection vocabulary.

3.1.4 Standards versus usage. Chabin (2017) argued that the French standardisation body's recommended equivalent for *record*, *document d'activité*, is out of touch with real usage. The press corpus supports this: the term does not appear once. A similar gap exists in Uzbekistan. Legal texts use *hujjat* and *elektron hujjat* (electronic document) as the core concept, and the profession uses *ish yuritish* for records work. Yet there is no settled Uzbek term that carries the specific evidential meaning of the English *record*. Translators of ISO 15489 into Uzbek, and teachers of archival English at language universities, face exactly the choice that the Panama Papers made visible.

3.2 From classification schemes to search engines (RQ2)

3.2.1 The investigative workflow. The Mossack Fonseca material arrived in batches over many months, in heterogeneous formats: emails, database exports, scanned PDFs, images of passports and certificates. Published accounts of the investigation (Cabra & Kissane, 2016; Obermayer & Obermaier, 2016) and the French press sources cited by Chabin (2017) describe a chain of tools. Table 4 sets each stage against the traditional archival function it replaced or supported.

Table 4. Stages of the ICIJ workflow and their archival equivalents

Stage	Tools reported	What the stage did	Traditional archival equivalent
1. Ingestion and processing	Nuix forensic software	Opened millions of files of many formats and extracted text and metadata	Accessioning; physical survey of a transfer

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Stage	Tools reported	What the stage did	Traditional archival equivalent
2. Text recognition	OCR within the processing platform	Made scanned contracts, passports and certificates searchable	Item-level description of unindexed paper
3. Indexing and search	Apache Solr and Apache Tika, with a Blacklight interface	Full-text and fuzzy search across the whole collection	Finding aids, card indexes, registers
4. Data cleaning	OpenRefine	Normalised names, dates and addresses	Authority control of names
5. Relationship analysis	Neo4j graph database; Linkurious visualisation	Linked clients, intermediaries, entities, officers and addresses	Provenance research; reconstructing an archival bond
6. Collaborative verification	A secure online forum (the Global I-Hub)	376 journalists shared leads and checked findings	Reading-room reference service and peer consultation

At no stage did the consortium build a classification scheme or a thesaurus. Access was organised entirely through search and through relationships discovered automatically. This confirms Chabin's (2017) provocative claim that "the classification scheme is dead" as a tool for retrieval in very large collections.

3.2.2 Examples of what search made possible. Four examples illustrate the gain.

Example 5: fuzzy matching of names. A person's name may appear in many spellings, especially across alphabets. An Uzbek surname such as *Xo'jayev* can occur as *Khodjaev*, *Khodzhaev*, *Hodjaev*, *Xojayev* or the Cyrillic *Ходжаев*, depending on the passport, the bank and the decade. A classification scheme would file the person under one heading chosen by an archivist. Fuzzy search, which tolerates one or two differing characters and can be combined with transliteration rules, retrieves all variants in one query. For collections from Central Asia, where Latin, Cyrillic and older transliteration systems coexist, this is a decisive advantage.

Example 6: lists of politicians matched against the leak. Journalists ran automated checks of lists of heads of state, ministers and other politically exposed persons against the whole index. A manual review of 11.5 million records for each of thousands of names would have been impossible; batch searching reduced it to a review of hits.

Example 7: graph analysis of nominee structures. In a graph database, each client, company, intermediary, officer and address is a node, and each role (shareholder, director, agent) is an edge. A query such as "find all companies whose director is also director of more than 100 other companies" exposes nominee directors at once. This is how investigators could show that a small group of

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

professional nominees served as directors of very large numbers of entities. The same structure, followed in the other direction, revealed reported networks of companies linked to the Russian cellist Sergei Roldugin, a long-time friend of President Vladimir Putin, through which transactions worth up to about two billion US dollars were said to have passed (ICIJ, 2016).

Example 8: dates and chronology. Because extracted dates were indexed, journalists could reconstruct timelines automatically, for instance the sequence in which an entity was incorporated, changed its shareholders and was dissolved after a sanctions announcement. In a paper archive, such a timeline would require reading every document in a client file in order.

3.2.3 Why structure still mattered. The speed of search did not remove the need for the qualities that archivists have always valued. Chabin (2017) identifies two: *structure* and *completeness*.

The Mossack Fonseca collection was well structured because a law firm keeps records in a logical and hierarchical way. Its likely arrangement can be described as follows:

- **Client register** - the spine of the collection, listing each client entity.
- **Entity file** for each company, foundation or trust - the archival aggregation (*dossier, delo, ish*).
- **Constitutive records** - articles of incorporation, registers of directors and shareholders, share certificates, powers of attorney.
- **Operational records in chronological order** - invoices, bank documents, resolutions, changes of officers.
- **Correspondence** - letters and, increasingly after the late 1990s, emails with clients and intermediaries, often carrying the constitutive records as attachments.

Each record in such a structure contains the basic elements that diplomatics identifies in any documented transaction: a date, a place, one or more persons, and an action. An illustrative (fictional) incorporation record can be analysed as shown in Table 5.

Table 5. Diplomatic elements of an illustrative incorporation record

Element	**Example value**	**Search use**
Date	12 May 2005	Timelines; sanctions and statute checks
Place	Panama City; registered in the British Virgin Islands	Jurisdiction analysis
Persons	Registered agent; client; intermediary bank; nominee director	Graph nodes; name matching
Action	Incorporate a company; issue bearer shares	Classification of the transaction

When these four elements are present and consistent, machine extraction is reliable. When

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

records come from many unrelated sources, each with its own logic, the same tools return results that are harder to interpret and to verify.

3.2.4 Completeness of the source versus completeness of the result. Chabin (2017) makes a distinction that is easy to state and often forgotten: completeness must be judged first for the source collection, then for the results. If the leak reproduces the complete records of Mossack Fonseca, the absence of a politician's name means only that this person was not a client of this firm. It does not mean that the person holds no offshore assets.

Later leaks confirm the point with concrete examples. The Paradise Papers (2017) came mainly from the law firm Appleby and named people and companies absent from the Panama Papers. The Pandora Papers (2021) drew on records from fourteen different offshore service providers and revealed undisclosed holdings of serving leaders, such as King Abdullah II of Jordan, who had not appeared in the 2016 leak. A newspaper that had concluded in 2016 that a public figure was "clean" because the Panama Papers did not mention them would have drawn a false inference from an incomplete result. In records management terms, the *record group* was complete; the list of offshore owners derived from it was not.

3.3 Measuring the leak: volume and meaningful metrics (RQ3)

3.3.1 Composition by file type. The file-type breakdown published by *Süddeutsche Zeitung* classifies 11,445,622 files, very close to the rounded total of 11.5 million. Table 6 adds the percentage shares.

Table 6. Panama Papers by file type

File type	**Number of files**	**Share**
Emails	4,804,618	41.98%
Database files	3,047,306	26.62%
PDF documents	2,154,264	18.82%
Images	1,117,026	9.76%
Text documents	320,166	2.80%
Other	2,242	0.02%
Total classified	**11,445,622**	**100%**

Source: Süddeutsche Zeitung (2016), via Les Echos and Chabin (2017); shares calculated by the author.

Two observations follow. First, emails alone account for more than two in five items, and emails plus PDFs, most of which were scanned legal instruments, make up over 60%. Second, images (often scans of passports, utility bills and certificates) were unreadable to a search engine until processed by OCR. Roughly 3.3 million PDFs and images, nearly 29% of the collection, therefore depended on a

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

technical conversion step before they could be searched.

3.3.2 Why the figures caused "vertigo". Press commentary treated the numbers as almost unimaginable. French outlets wrote that a single newspaper would have needed decades to go through the data, or compared the 2.6 terabytes to 34,665 television episodes (Chabin, 2017). This last comparison illustrates the problem of inappropriate metrics. Dividing 2.6 TB by 34,665 gives about 75 MB per episode, a measure based on compressed video. A single email of the leak occupies a few kilobytes. Byte counts mix very different kinds of content and say almost nothing about how much there is to read.

The average size of an item in the leak is about 226 KB (2.6×10^{12} bytes ÷ 11.5×10^{6} items). That is the size of a short scanned document or a small email with an attachment. In 2016 the whole leak would have fitted on a single portable hard drive of 3–4 TB costing well under 200 US dollars.

3.3.3 Paper equivalents: recalculating Chabin's model. Chabin (2017) proposed translating the leak into paper terms that records managers handle every day. Her chain of reasoning, recalculated here, is:

- 11.5 million items × 3.5 pages per item (midpoint of her 3–4 page assumption) = about **40 million pages**.
- 40.25 million pages ÷ 800 pages per standard 10 cm archive box = about **50,300 boxes**.
- 50,300 boxes × 0.10 m = about **5,030 linear metres**, i.e. roughly **5 km** of shelving.
- 5,030 m ÷ 376 journalists = about **13.4 m per journalist**, roughly two large filing cabinets each.
- 5,030 m ÷ 500 employees of Mossack Fonseca = about **10 m per employee**, over almost four decades.

The same figures can be expressed per client and per year:

8. **Per client entity:** 11.5 million ÷ 240,000 entities (Chabin's figure) = about 48 items; using ICIJ's database figure of 214,488 entities, about 54 items. For a law firm, a client file of about 50 documents over its life is entirely ordinary. Chabin notes that some files are very thin while the largest contain around 3,000 items.

9. **Per year:** the records span 1977–2015, i.e. 39 years. 11.5 million ÷ 39 ≈ 295,000 items a year, or about 1,180 items per working day (assuming 250 working days). Shared among 500 employees, this is about 2.4 items per employee per day. In reality the volume grew sharply in the email era, but even a tenfold concentration in later years produces figures that any records manager of a mid-sized firm would recognise.

3.3.4 Sensitivity of the model. The estimate depends on two assumptions. Table 7 varies them.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR

Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Table 7. Shelf-length equivalent under different assumptions

Pages per item	Total pages	Linear metres (800 pages per box)	Metres per journalist (376)
2.0	23.0 million	2,875	7.6
3.0	34.5 million	4,313	11.5
3.5	40.3 million	5,031	13.4
4.0	46.0 million	5,750	15.3
5.0	57.5 million	7,188	19.1

With 3.5 pages per item, a fuller box of 1,000 pages gives about 4,025 m, and a lighter box of 600 pages about 6,708 m. Under every reasonable assumption the leak lies between 3 and 7 km of shelving. For comparison, national archives routinely hold hundreds of kilometres of records. The volume of the Mossack Fonseca collection is large for a leak but modest for an archive.

3.3.5 Comparison with other leaks. Table 8 places the Panama Papers among other well-known disclosures.

Table 8. The Panama Papers among other large leaks and disclosures

Year	Leak or release	Source of records	Reported size
2021	Pandora Papers	14 offshore service providers	11.9 million records; 2.94 TB
2020	FinCEN Files	US Treasury's financial intelligence unit	Over 2,100 suspicious activity reports
2017	Paradise Papers	Appleby law firm and corporate registries	13.4 million records; 1.4 TB
2016	Panama Papers	Mossack Fonseca law firm	11.5 million records; 2.6 TB
2014	LuxLeaks	PwC tax rulings in Luxembourg	About 28,000 pages; 548 rulings
2013	Offshore Leaks	Two offshore service providers	About 2.5 million files; 260 GB
2010	Cablegate (WikiLeaks)	US diplomatic cables	251,287 cables
2003	Enron email corpus (official release)	Enron Corporation, via US regulator FERC	About 0.5 million emails

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Sources: ICIJ project pages; Klimt & Yang (2004) for the Enron corpus. Sizes are as reported by the publishing organisations.

The comparison shows that the Panama Papers were about ten times larger than Offshore Leaks by volume (2.6 TB against 260 GB) and about 4.6 times larger by number of files. Yet only five years later the Pandora Papers were larger still. The "record" size of 2016 was quickly normalised. What changed between leaks was less the volume than the number of sources combined and the capacity of networks and tools to process them.

3.3.6 Summary of the metric problem. None of the usual units - bytes, files or television episodes - tells a reader how much meaningful content exists. Bytes are dominated by images and attachments; file counts treat a one-line email and a 200-page contract equally; entertainment equivalents confuse storage with reading time. Records management offers better units: the aggregation (client file), the transaction, the linear metre, and the retention category. These units connect volume with function and therefore with meaning.

3.4 "In the beginning was electronic mail": email as the point of entry (RQ4)

3.4.1 How the leak started. Contemporary technical reporting, cited by Chabin (2017), linked the leak to a compromise of Mossack Fonseca's email server. Security commentators at the time also reported that the firm's client portal and public website ran outdated, unpatched software and that its email traffic was not encrypted. Whatever the exact technical path, the result is clear from Table 6: the largest single category in the leak was email, and emails led to their attachments, which carried the contracts, certificates and passports.

3.4.2 Email in diplomatic terms. An email is an unusual record. Its header supplies several diplomatic elements automatically, but it often depends on other records for its full meaning. Table 9 compares a typical email with a formal corporate record from the same client file.

Table 9. Diplomatic elements of an email compared with a formal corporate record

Element	Email (illustrative)	Board resolution (illustrative)
Date	Automatic, precise to the second	Written by the author; may be backdated
Place	Implied only by server or signature block	Stated (e.g. "Panama City")
Persons	From, To, Cc fields; may include personal accounts	Named directors with signatures
Action	Often indirect: "please find attached…"	Explicit: "resolved to appoint…"
Form	Free, informal, threaded	Fixed, formal template
Evidential role	Transmits and contextualises	Creates or changes a legal situation

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

A typical (illustrative) message - *"Dear Sir, please find attached the signed nominee director declaration for Company X. Kindly keep the original in your files and do not forward."* - shows this dual role. The email is the record of transmission and of intent; the attachment is the dispositive record that changes the legal situation. Investigators needed both. The email linked a real person (the sender or recipient) to an entity whose official documents named only nominees. Here email supplied exactly the context that the formal records were designed to conceal.

3.4.3 Comparable cases. The Panama Papers are not the only case in which email became the decisive record.

Example 9: the Enron corpus. After the collapse of Enron in 2001, the US Federal Energy Regulatory Commission released about half a million emails from some 150 employees. The corpus became a standard research dataset for email classification and network analysis (Klimt & Yang, 2004). It showed how an organisation's informal email record can reveal decision-making that its formal records omit.

Example 10: the Sony Pictures hack (2014). Attackers published thousands of internal emails. The damage came less from trade secrets than from candid remarks by executives, a clear illustration of Chabin's point that the risk lies in uncontrolled email content.

Example 11: the private email server of a US Secretary of State (2015–2016). The use of a private server by Hillary Clinton during her time as Secretary of State raised questions under US federal records law about whether official emails had been captured and preserved. The case turned not on a leak but on the failure to bring email inside the official records system.

Example 12: managing email by role. In response to such problems the US National Archives introduced the "Capstone" approach in 2013. Instead of asking every employee to classify every message, Capstone preserves entire email accounts of senior officials permanently and applies shorter retention to other accounts (NARA, 2013). It is an example of records management adapting to email volume by changing the unit of appraisal from the item to the account.

3.4.4 Email and governance. Chabin (2017) compares executives' attitude to information governance to the cicada in La Fontaine's fable, who sings all summer and has nothing when winter comes, while the ant prepares. The examples above suggest three recurring failures that the fable describes: no policy on what may be sent by email; no technical protection such as encryption and patched servers; and no integration of email into the formal records system. In each case the absence of governance became visible only after the records left the organisation.

For organisations in Uzbekistan the lesson has practical relevance. The Law "On Electronic Document Management" (2004) gives legal force to electronic documents signed with an electronic digital signature. Yet in daily practice much business is still conducted by ordinary email and messaging applications outside any document management system. Such communications are records in the ISO 15489 sense, but they are rarely appraised, retained or protected as records.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

4. Discussion

4.1 Terminology: words carry archival concepts

The terminological results confirm and extend Chabin's (2017) observation that recommended equivalents of *record* do not match living usage. In all four languages examined, the media preferred *document* (*document*, *dokument*, *hujjat*) to any specialised term for *record*. This preference is not merely stylistic. The concept of a record rests on three properties: it is evidence of an activity, it is bound to its creator, and it belongs to an aggregation (Duranti, 1998; Yeo, 2007). The word *document* conveys only the first loosely and the other two not at all.

The disappearance of the aggregation is the most important loss. English *file*, French *dossier*, Russian *delo* and Uzbek *ish* name the archival unit that gives each document its context. In news language, this unit is replaced by the computer file (*fichier*, *fayl*), a unit of storage with no archival meaning. Readers of the coverage could therefore easily imagine the leak as 11.5 million unrelated objects, while in fact it was about 214,000–240,000 well-organised client files. The Russian name *Панамское досье* is an interesting exception: by calling the whole leak a "dossier", it preserves the idea of an organised case file, even if at the scale of the entire collection.

The false friend *archives* has practical consequences for translation and teaching. Since the records continuum model (Upward, 1996) and ISO 15489, the English-speaking profession has tried to reduce the separation between "records" and "archives". French usage has always joined them; Russian and Uzbek usage, shaped by the Soviet state archival system, separates them by institution rather than by concept. Translators working between these traditions need to decide case by case whether *archives* means the current records of an organisation, the repository that holds them, or a compressed computer file. For language universities such as UzSWLU, the Panama Papers offer an authentic corpus for teaching these distinctions in courses on specialised translation and terminology.

4.2 Search engines and the human sense of context

The workflow analysis supports the claim that full-text search, fuzzy matching and graph analysis have replaced the classification scheme as the primary means of retrieval in very large collections. No archivist could have built a subject index for 11.5 million items in time to be useful, and none was needed. In this sense the Panama Papers belong with other demonstrations of computational methods in archives and e-discovery, where technology-assisted review has been accepted in courts as more consistent than manual review (Grossman & Cormack, 2011).

The results also support Chabin's (2017) qualification: the quality of machine results depends on the quality of the source. The Mossack Fonseca collection was searchable because it was structured by a law firm's discipline: a client register, entity files and constitutive documents with clear dates, places, persons and actions. Technology exploited an order created by records management decades earlier. In this light, classification has not died; it has moved. It no longer lives in a scheme maintained by archivists for retrieval, but in the structure of the records system at the moment of creation, which later supports automated processing.

What technology has not yet replaced is the experienced professional's ability to sense the "density" of a collection at a glance: whether parts are missing, how it accumulated, whether it is

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

reliable. Chabin (2017) compares this to the intuition of an old-style police detective, set against forensic science. In the digital setting, equivalent signals could be provided by dashboards that show gaps in date ranges, unusual distributions of file types, or client files with abnormally few or many items. Such "contextual metrics" are a promising area where records management and data science could meet.

4.3 Completeness and the ethics of inference

The distinction between a complete source and an incomplete result is perhaps the most practical lesson for journalists, researchers and compliance officers. A search that returns no hits proves only that the searched collection contains no hits. The Paradise and Pandora Papers showed that people absent from the Panama Papers could appear in other firms' records. The same principle applies inside organisations. An internal investigation that searches only the official document management system, and not email or messaging applications, cannot conclude that a decision was never recorded. Records managers can contribute by documenting the boundaries of every collection: what it contains, which period it covers, and which sources it excludes.

4.4 New metrics for digital information

The public's "vertigo" before 2.6 TB reflects a missing vocabulary rather than an extraordinary volume. Digital technology increases granularity: a paper letter with an enclosure becomes an email, an attachment and several metadata records, each counted separately. Byte counts and file counts therefore inflate apparent size without increasing meaning. The calculations in Section 3.3 show that the collection corresponds to about 5 km of shelving, around 50 items per client and 2–3 items per employee per working day. Expressed in these units, the leak becomes ordinary.

The profession therefore needs metrics that relate volume to function. Candidates include the number of aggregations (client files, case files), the number of documented transactions, the share of records by retention category, and paper-equivalent shelf length as a communication device for non-specialists. Such metrics would help journalists report leaks accurately and help organisations understand their own information assets.

4.5 Networks, email and the real risk

The final point of the discussion concerns where the risk to records lies. Debates about digital preservation often focus on the fragility of formats and media. The Panama Papers show a different risk: records that are perfectly preserved can escape their creator through networks. As Chabin (2017) notes, what was remarkable was not the volume of the data but the fact that it was shared outside its intended recipients and managers. Email is the main channel for such escape, because it combines high volume, informal content, attachments with legal force, and weak integration into records systems.

The ethical dimension is symmetrical. In the Panama Papers, honest journalists used these technologies to expose alleged wrongdoing. The same methods - server intrusion, mass indexing, graph analysis - could be used against honest citizens or organisations. Records management therefore has a dual responsibility: to make records accessible for legitimate accountability and to protect them from unauthorised disclosure. Both depend on governance, not on technology alone.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

5. Conclusion

The Panama Papers are usually read as a story about money and power. Read as a story about records, they offer four lessons, one for each research question.

First, **terminology shapes understanding**. In English, French, Russian and Uzbek alike, the media described the leak as "documents" and "files" and rarely as records organised into client files. The archival aggregation that gave the collection its meaning almost vanished from public language, with the notable exception of the Russian *Панамское досье*. Standard equivalents such as the French *document d'activité* remain unused, and Uzbek still lacks a settled term for the evidential sense of *record*. Bilingual and multilingual glossaries grounded in real usage are needed, and language universities are well placed to develop them.

Second, **search engines have replaced classification schemes for retrieval, but not structure for meaning**. Fuzzy search, OCR and graph databases made 11.5 million items usable by 376 journalists in about a year. They worked because the source was a well-structured, complete record collection. The completeness of a source must never be confused with the completeness of the results drawn from it.

Third, **the volume was large for a leak but ordinary for an archive**. Translated into records management units, the 2.6 TB correspond to about 40 million pages, 5 km of shelving, around 50 items per client file and two to three items per employee per working day. The public's astonishment points to the absence of meaningful metrics for digital information, not to an unprecedented quantity.

Fourth, **email is the weakest point of information governance**. Emails formed 42% of the leak and opened the way to the attachments that carried legal force. The Enron, Sony and US federal email cases show the same pattern: risk arises where informal communication is not governed as a record.

From these findings, several practical recommendations follow for records professionals, including those in Uzbekistan:

- Capture email and messaging records inside managed records systems, and apply role-based retention in the spirit of the Capstone approach.
- Protect record systems as networks, not only as storage: patch servers, encrypt communications and monitor access.
- Design records at creation with clear diplomatic elements (date, place, persons, action) so that they remain reliable under automated analysis.
- Document the scope and gaps of every collection so that searches can be interpreted correctly.
- Report volume in functional units (aggregations, transactions, linear metres) alongside bytes and file counts.
- Develop and teach multilingual records terminology (English–Russian–Uzbek–French) based on authentic texts such as the Panama Papers coverage.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR
Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Records management, as Shepherd and Yeo (2003) argue, exists to keep records trustworthy and usable over time. The Panama Papers show that this task now extends from the shelf and the server to the network itself. Technology is indispensable, but, as Chabin (2017) concludes, information management will fail if machines are treated as more important than the people who create, manage and interpret records. Future research could test the metrics proposed here on other leaks, build a parallel corpus of Uzbek, Russian and English coverage, and examine email governance practices in Uzbek organisations empirically.

References

1. Cabra, M., & Kissane, E. (2016). *Wrangling 2.6TB of data: The people and the technology behind the Panama Papers*. International Consortium of Investigative Journalists.
2. Chabin, M.-A. (2017). Panama papers: A case study for records management? *Brazilian Journal of Information Science: Research Trends, 11*(4), 10–13. https://doi.org/10.36311/1981-1640.2017.v11n4.03.p10
3. Duranti, L. (1998). *Diplomatics: New uses for an old science*. Scarecrow Press.
4. GOST R ISO 15489-1-2019. (2019). *Sistema standartov po informatsii, bibliotechnomu i izdatel'skomu delu. Informatsiya i dokumentatsiya. Upravlenie dokumentami. Chast' 1: Ponyatiya i printsipy* [Information and documentation. Records management. Part 1: Concepts and principles]. Standartinform.
5. Grossman, M. R., & Cormack, G. V. (2011). Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, *17*(3), 1–48.
6. International Consortium of Investigative Journalists. (2016, April 3). *Giant leak of offshore financial records exposes global array of crime and corruption*. https://panamapapers.icij.org/20160403-panama-papers-global-overview.html
7. International Consortium of Investigative Journalists. (2017). *Paradise Papers: Secrets of the global elite*. ICIJ.
8. International Consortium of Investigative Journalists. (2021). *Pandora Papers*. ICIJ.
9. International Organization for Standardization. (2016). *ISO 15489-1:2016 Information and documentation - Records management - Part 1: Concepts and principles*. ISO.
10. Klimt, B., & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004* (Lecture Notes in Computer Science, Vol. 3201, pp. 217–226). Springer.
11. Les Echos. (2016). *Panama Papers : les coulisses de l'enquête*. http://www.lesechos.fr/monde/enjeux-internationaux/021815669450-panama-papers-les-coulisses-de-lenquete-1211437.php
12. National Archives and Records Administration. (2013). *NARA Bulletin 2013-02: Guidance on a new approach to managing email records*. NARA.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR

Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

13. Obermayer, B., & Obermaier, F. (2016). *The Panama Papers: Breaking the story of how the rich and powerful hide their money*. Oneworld.
14. O'Donovan, J., Wagner, H. F., & Zeume, S. (2019). The value of offshore secrets: Evidence from the Panama Papers. *The Review of Financial Studies, 32*(11), 4117–4155.
15. O'zbekiston Respublikasining "Elektron hujjat aylanishi to'g'risida"gi Qonuni [Law of the Republic of Uzbekistan "On Electronic Document Management"]. (2004, April 29). No. 611-II. https://lex.uz
16. O'zbekiston Respublikasining "Arxiv ishi to'g'risida"gi Qonuni [Law of the Republic of Uzbekistan "On Archival Affairs"]. (2010, June 15). No. O'RQ-252. https://lex.uz
17. Shepherd, E., & Yeo, G. (2003). *Managing records: A handbook of principles and practice*. Facet Publishing.
18. Stake, R. E. (1995). *The art of case study research*. Sage.
19. Süddeutsche Zeitung. (2016). *About the Panama Papers*. Süddeutsche Zeitung.
20. Upward, F. (1996). Structuring the records continuum, part one: Postcustodial principles and properties. *Archives and Manuscripts, 24*(2), 268–285.
21. Yeo, G. (2007). Concepts of record (1): Evidence, information, and persistent representations. *The American Archivist, 70*(2), 315–343.
22. Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage.
23. Zucman, G. (2015). *The hidden wealth of nations: The scourge of tax havens*. University of Chicago Press.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 15 Issue: 01 in January-2026
https://www.gejournal.net/index.php/IJSSIR

Copyright (c) 2026 Author (s). This is an open-access article distributed under the terms of Creative Commons Attribution License (CC BY). To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/