EU Tax Observatory Working Paper No. 32

Safely Opening Pandora’s Box: A Guide for Researchers Working with Leaked Data

May 2025

Annette Alstadsæter
Skatteforsk and EU Tax Observatory

Matthew Collin
EU Tax Observatory and Skatteforsk

Andreas Økland
Skatteforsk and EU Tax Observatory

The views expressed here are those of the author(s) and not those of the EU Tax Observatory. EU Tax Observatory working papers are circulated for discussion and comment purposes. They have not been subject to peer-review or to the internal review process that accompanies official EU Tax Observatory publications.

Safely Opening Pandora's Box: A Guide for Researchers Working with Leaked Data[*]

Annette Alstadsæter (Skatteforsk and EU Tax Observatory)
Matthew Collin (EU Tax Observatory and Skatteforsk)
Andreas Økland (Skatteforsk and EU Tax Observatory)

May 22, 2025

Abstract

The use of leaked data is becoming increasingly common in empirical research, particularly in public finance. Although these data can be an enormously powerful tool for investigating behavior that is otherwise difficult to measure, their use also generates substantial ethical and legal risks. In this paper, we (i) present the growing body of social science research that relies on leaked data, (ii) discuss the ethical, legal and privacy hurdles faced by projects relying on such data and (iii) offer a practical roadmap for researchers looking to enter the space.

[*] Corresponding author: Annette Alstadsæter, Norwegian University of Life Sciences (NMBU), e-mail: annette.alstadsater@nmbu.no. This paper builds on insights from Alstadsæter's keynote, "Navigating the Hidden Currents: The Evolution and Measurement of Offshore Wealth in the Age of Data Leaks", presented at the 80th Annual Congress of the International Institute of Public Finance (IIPF) in Prague, 2024. It also incorporates ideas and discussions from the special session "Using Leaked Data in Academic Research" at the same congress, including presentations by Collin and Økland. We are grateful to the IIPF for providing a platform to explore these issues, and to the session participants for their enthusiastic engagement and valuable feedback. Further, we thank the editor of this journal issue, Dominika Langenmayr, an anonymous referee, Ron B. Davies, Aïcha Payet, and Hanne Pernille Gulbrandsen for valuable comments and suggestions. Any remaining errors are our own. We gratefully acknowledge funding from the Research Council of Norway, grants No. 341289 and 352151.

1 Introduction

From the Offshore Leaks to the Pandora Papers, leaked data are becoming increasingly accessible for researchers. Today, data from hundreds of different leaks can be found online, hosted by investigative reporting consortia, whistleblower organizations, or on the dark web.[1] This growing cache of information affords academics an opportunity to answer questions of enormous public benefit. Leaks frequently capture behavior that is socially harmful but often hard to measure, such as tax evasion, corruption, or other forms of illicit activity. This has led to a surge in academic work in fields such as economics, political science and finance: roughly forty papers using leaked data have been published or released as working papers in the past two decades, the majority in the past few years (Figure 1).

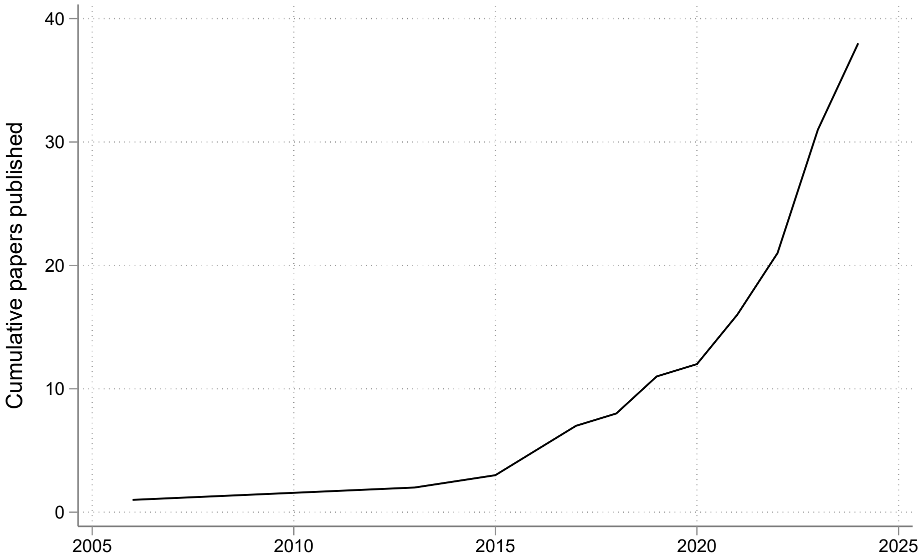
However, the use of leaked data in research raises important ethical and legal questions around privacy and the lack of consent from the individuals involved. Existing work relying on leaked data rarely addresses these issues in a consistent manner (Thomas et al., 2017), raising concerns that researchers are acting on an ad hoc basis, without clear guidance or institutional support. Without developing clear, transparent standards, leaked data research may face a legitimacy crisis in the future, leading academic and publishing institutions to withdraw their support for otherwise-impactful work. For example, journals managed by the American Economic Association recently adopted a Data Legality Policy, subjecting work using data that was not "legally obtained" to further scrutiny.[2]

Drawing upon our own experience running multiple projects, the goal of this paper is to provide a guide for researchers embarking on projects using leaked data by demonstrating how they can navigate ethical issues such as harm and consent as well as data protection and privacy concerns such as those recently introduced by the General Data Protection Regulation (GPDR) in Europe. Our target audience includes academics starting a new project, but also extends to those who may make decisions affecting a project's progress, including Institutional Review Boards (IRBs), university data protection officers, and journal editors. We argue that, when managed with rigor and transparency, research using leaked data can be enormously impactful, while still aligning closely with the ethical and legal principles that govern the use of personally identifiable information

[1] The transparency organization Distributed Denial of Secrets (DDOS) hosts more than 350 leaks, and together the International Consortium of Investigative Journalists (ICIJ) and the Organized Crime and Corruption Reporting Project (OCCRP) host dozens more.

[2] See "AEA Data Legality Policy and Explanations" https://www.aeaweb.org/journals/data/data-legality-policy

Figure 1: Total number of papers using leaked data over time



Notes: Figure 1 shows the cumulative number of papers in our literature review (see Section 3) that use leaked data, both published and in working paper form. Our search was restricted largely to economics, accounting and political science, broadly dealing with issues involving illicit behavior.

from traditional sources.

This paper proceeds as follows. In Section 2 we discuss what we mean by "leaked data," detailing the types of data leaks that are ordinarily made available to researchers, along with the varying levels of pre-processing they entail. In Section 3 we present the results of our literature review of existing work in social science that relies on leaked data, showcasing the breadth of questions that can be answered using this information. In Sections 4 and 5, we highlight the main ethical and data privacy considerations that researchers working with leaked data must grapple with. Combining these lessons, in Section 6 we summarize and provide a checklist of best practices for how researchers can engage with such data in a responsible manner.

2 What is leaked data and what forms can it take?

The general increase in availability of large-scale datasets has widened the scope and increased the precision of empirical research in recent decades. Researchers can, after undergoing thorough review processes and data anonymization, analyze sensitive data on topics like employment, education and health at the individual or company level. However, some questions cannot be answered with the data willingly shared by companies and public authorities or with the answers gathered in a survey. In situations where these unanswered questions are important enough, a researcher may justify the use of leaked data. In this section, we outline the different types of leaked data that are available, as well as the different levels of pre-processing that may have occurred before landing on a researcher's desk, with implications for how much personal data they might have access to. We also review the existing works in social science that relies on leaked data.

There is no definitive definition of leaked data, but we consider all data obtained or released against the will of the original people or entities who control the data as leaked data. This can include, for instance, situations where those within an organization intentionally share sensitive information with external parties, as was the case with the Panama Papers. It also includes data illicitly obtained by external parties through actions such as hacking, aggressive data scraping or physical theft.[3] Additionally, leaked data can include information inadvertently published online or disclosed through legal proceedings, such as court filings.[4] While the ethical challenges associated with using intentionally or forcefully disclosed data are often more pronounced, all forms of leaked data present heightened ethical and legal considerations for researchers. We have intentionally adopted a broad definition of leaked data so we can explore the whole gamut of issues that arise when researchers choose to use these types of data.

Both the structure and level of fidelity provided by leaked data can vary greatly, depending on its origin. We describe four typical scenarios below and examples of how they are used by researchers. We then summarise in a separate section how leaked data have contributed to recent advancements in several social sciences literatures.

2.1 Unstructured leaked data from original sources

Leaked data are - by definition - rarely produced with the purpose of being fed into empirical research. This means that when researchers access leaked data directly, not via an intermediary (such as a media organization or nonprofit) that has already processed it, the data usually need extensive processing and cleaning before it can be used for research. For example, the Panama Papers consists of 11.5 million financial and legal records from the Panamanian law firm Mossack Fonseca. These documents were shared

[3] The *Luanda Leaks* and leak from the Cayman National Bank were both the result of hacking. By contrast, a leak of 70,000 profiles from the dating website OK Cupid was the result of aggresive data scraping by a Danish researcher.

[4] Reporting by the OCCRP on the scale of Dominica's Citizenship by Investment (CBI) scheme relied on the transcription of physical copies of official government gazettes. By contrast, Barake et al. (2024) rely on data from legal filings from the bankrupcy of the crypto platform Celcius for the data used in their work.

with journalists in Süddeutsche Zeitung by a whistleblower. The full data leak covers close to 5 million emails, more than 2 million PDFs and 1 million images, but also extensive excerpts of an internal Mossack Fonseca database (Obermaier et al., 2017).[5]

The size and format of the data makes it evident that a group of researchers would have to put in a lot of effort to understand, verify and structure this dataset before they could start any meaningful analysis. It also requires substantial legal and security processes to handle this type of sensitive data, which we will discuss in detail in later sections. In addition, the source of the data needs to be protected, something neither researchers nor the infrastructure around them are used to handling. While institutions that operate under the GDPR often have robust procedures for managing sensitive information—such as health or administrative records—these protocols typically do not account for the unique challenges posed by leaked data. Researchers are rarely prepared for the fact that such data may have reached them through actions that, while possibly motivated by whistleblowing or public interest, may also involve legal or ethical breaches by the party providing access.

This is why these types of "raw" leaks are more likely to be used by journalists than academics. Although most leaks start out as an unstructured data dump, researchers are more inclined to use them after they have been pre-processed. This can change in the future as more raw leaks are being posted openly on the internet, where researchers can access them directly. This means that more researchers might be inclined to invest the time and effort required to analyze these leaks.

One example of such a leak that is used for research purposes is the data that was obtained when the Isle of Man subsidiary of the Cayman National Bank was hacked. The material from this leak was handed over by the hacker to the journalist collective Distributed Denial of Secrets (DDOS), who made this material public. The material was shared with the public in two ways: The first way was through a searchable database of files and e-mails. The second was the leaked data in its entirety: the contents of several dozen hard drive images taken from bank's servers. This material was first analyzed by Collin (2021) and is further analyzed in Bomare and Collin (2025).

2.2 Processed leaks from intermediaries (made public)

Processed leaks are often more accessible to researchers because they have already been structured by a third party who gained access to the leak first. The best-known example

[5] https://panamapapers.sueddeutsche.de/articles/56febff0a1bb8d3c3495adf4/

of this is the ICIJ Offshore Leaks Database, which compiles information from multiple data leaks into an easily searchable format. The database covers more than 800,000 offshore companies (shell companies) and includes information about incorporation, relevant intermediaries and related individuals like shareholders, beneficial owners, directors and trustees.

Importantly, the ICIJ Offshore Leaks Database does not include the original documents the database was built on. This means that information about the offshore bank accounts of the respective shell companies is not public. Likewise, the passport information and other supporting documents that have helped journalists verify the data and report on the information remains secret to the researchers. The processing has thus limited the fidelity of the personally identifiable information available to the researchers, reducing some of the risks we will highlight in Section 4 and 5, while also potentially impacting the quality of the data.

Londoño-Vélez and Ávila-Mahecha (2021) and Londoño-Vélez and Avila-Mahecha (2025) are examples of projects that use the data in the ICIJ Offshore Leaks Database together with micro data from tax records. The authors show in two affiliated papers how Colombian individuals in the Panama Papers leak seems to be engaged in offshore tax evasion and that the leak led many of them to disclose their offshore assets to the Colombian tax administration.

Bomare and Le Guern Herry (2024) and Lafitte (2024) are examples of research that use the Offshore Leaks Database to understand the patterns of tax haven use without linking the leaked data to specific individuals. Instead, they investigate the patterns in the data to uncover which tax havens are popular with the residents of different countries and whether there is a gravity relationship. Bomare and Le Guern Herry (2024) use the composition of residence country of shell company owners in different jurisdictions to assign the tax haven jurisdictions into treatment and control groups based on whether they tend to host shell companies for individuals from countries that implemented the Common Reporting Standard (CRS), an international tax compliance cooperation. Lafitte (2024) estimates a gravity model on the Offshore Leaks database, where the number of links between two jurisdictions (for instance that an individual from country a is the owner of a shell company in jurisdiction b) are the dependent variable.

2.3 Processed leaks from intermediaries (non-public)

The ICIJ Offshore Leaks Database is made available to researchers and others after the ICIJ journalists are done with their reporting on the material. An alternative to this model is a closer alliance between researchers and the intermediaries. The work on Dubai real estate data, which all three authors of this paper have been involved in, is an example of this. The academic output of this collaboration is available in Alstadsæter et al. (2022) and Alstadsæter et al. (2024).

C4ADS (Center for Advanced Defense Studies), an American think tank with long experience working on security issues, has received several dumps of information about real estate ownership in Dubai from undisclosed sources. Both C4ADS and media collaborators published stories and cases based on the information in the first datasets they received (Page, 2020). Starting in 2020, the C4ADS decided to formally collaborate with researchers in order to better analyze a new leak of data, which included extensive information about individual ownership of real estate in Dubai. The researchers got exclusive access to the data and were able to develop security procedures in cooperation with C4ADS, who have long experience in handling this type of data. Working with an intermediary also allows for a good understanding of where the data comes from, without the hands-on experience of working with the source. The third part of this collaboration were journalists. The C4ADS shared the leaked real estate data with selected media partners, who reported on individual cases in the material. This again created new synergies, like independent verification of the data, discussion about data patterns, etc.

This collaboration helped C4ADS process the data, understand potential deficiencies and unveil patterns they might not have been able to see themselves. One example of this is the work done by us to construct a model that puts a monetary value on each property, something that was not included in the leaked data (Alstadsæter et al., 2022).

Another example of this type of leaked data is data handed over to authorities, such as the Swiss Leaks, which was provided to tax administrations and journalists and subsequently shared with researchers. Alstadsæter et al. (2019) utilized this data to study global wealth and tax evasion. Processed leaks significantly lower the barrier for researchers, as much of the work to organize the data has already been done.

2.4 Processing openly-available data against the desire of the data holder (such as scraped data)

While the unstructured data discussed in 2.1 is handed over to researchers by external, closed sources, this subsection discusses data the researchers accessed through open sources.

There are many reasons why the availability of data points in open sources does not translate into easily accessible datasets. On the one innocent end of the scale is data published in an unstructured way mainly for user interface or design reasons. This might be data from social media or other web platforms, although terms of use of the service might still regulate the use. Another example is data disclosed as a byproduct of other processes. For instance, the Celsius cryptocurrency platform depositor list, used by Barake et al. (2024), was published during bankruptcy proceedings, offering a unique window into the platform's operations.

On the other end of the scale, we find data that is intentionally published in an unstructured way. Multiple administrative data sources are designed like this, for instance. The purpose is often to prevent external users from accessing large-scale registries in bulk or performing broad queries. Instead, users are typically restricted to looking up entries based on a single, predefined variable, such as a registration number or an address. An example is how registries let you look up the owners of cars or properties by registration number or address, while not allowing users to look up the total ownership of a given person or company. Many beneficial ownership registries are for instance designed like this. The leak labeled "OpenLux" is an example of this type of leak. Journalists with the French newspaper Le Monde used scraping techniques to build a registry of the beneficial owners of all companies in the Luxembourg beneficial ownership registry.

A somewhat different example of gathering, structuring and analysing openly available data in a way that is clearly against the will of the data owners is the work by Ederer et al. (2024). They analyse the users of the website Economics Job Market Rumors (EJMR), an online forum for economists, using only publicly available data. The website let users post pseudonymously, by generating encrypted usernames based on the IP addresses. The authors recover 47,630 distinct IP addresses of EJMR posters by unlocking the formula that generate the pseudonymous usernames from the IP addresses. They then geolocate the posts based on the likely IP address of the poster and analyse the language used on the forum by posters connected to different specific geographies and institution.

Researchers who collect information about individuals from open sources against the

the will or desire of the data holders need to handle this data with the same care as they would do with other types of data covering individuals. The ethical standards imposed on researchers do not distinguish based on how the data was obtained. The collection, structuring and analysis of the data therefore also require the same careful consideration as the use of the types of data described in prior sections.

3 Existing research using leaked data

In this section, we demonstrate that leaked data are becoming more common in applied empirical work, and is being used to answer a wide range of academic questions. As part of our literature review, we used both firsthand knowledge and extensive search (via Google Scholar) to identify around forty research papers using leaked or illicit data over the past twenty years. We largely focused on work within the fields of economics, finance, accounting and political science, broadly dealing with issues involving illicit behavior, including tax evasion and avoidance, corruption, or the use of tax havens.[6] It is not illegal to interact with or in tax havens, but the opacity services provided there can be used for illegal or harmful activities. Thus, this is not a comprehensive stock take of all empirical work with leaked data, but what we think would be of the most interest to economists working on these topics.[7]

Roughly half of these relied on a single source: the ICIJ's Offshore Leaks Database which we introduced above,[8] the most prominent leak being the Panama Papers. The popularity of this database is driven in part by the scale of the data, but also the fact that the data are presented in a downloadable, machine-readable format that is easy for researchers to use. The rest are a mix of different sources and means of acquisition: hacks of offshore banks hosted on a transparency website, sensitive government data leaked to think tanks (or directly online), and transaction data from cryptocurrency exchanges or central banks leaked directly online.

One common strand of work is the use the Offshore Leaks Database to study efforts to circumvent multilateral policy efforts to crack down on offshore secrecy and tax evasion or to enforce sanctions. Caruana-Galizia and Caruana-Galizia (2016) use an early version

[6] See Griffin and Kruger (2024) for a thorough review of the growing field of 'forensic finance', a subfield of finance that analyze potential illegal or immoral activity, where leaks is a commonly used data source.

[7] Examples of other uses of leak data that do not appear here include a large body of political science/international relations work using Wikileaks data.

[8] The Offshore Leaks Database comprises multiple leaks that occurred during the 2010s, including the Pandora Papers, Paradise Papers, Panama Papers, Bahamas Leaks and Offshore Leaks, but all are presented in a similar, machine-readable structure.

to show that EU residents responded to the introduction of the 2005 EU Tax and Savings Directive by opening up shell companies in non-European haven jurisdictions. Omartian (2017) finds a similar response using the Panama Papers, also showing that efforts to strengthen the EU Tax and Savings Directive so that European beneficial owners of shell companies would also be covered led to decline in their use. Bomare and Le Guern Herry (2024) investigate the impact of the Common Reporting Standard (CRS) on offshore investment into non-reportable UK real estate, using the Offshore Leaks Database to identify tax havens with a greater share of companies owned by investors being targeted by the CRS and thus more likely to be used as vehicles for real estate investment. Similarly, Collin et al. (2023) use this database to identify tax havens most likely to respond to a beneficial ownership transparency measure introduced by the UK. Kavakli et al. (2023) show that individuals living in countries targeted by sanctions are more likely to incorporate shell companies in tax havens as a response, presumably in an effort to circumvent the sanctions regime.

Another set of studies have used leaked data to study the location, characteristics and behavior of those who own offshore assets or engage in tax evasion. Alstadsæter et al. (2019) merge both data leaked by HSBC Switzerland and the Panama Papers to Scandinavian tax returns to show that the propensity to hold an unreported Swiss bank account or a shell company rises sharply at the very top of the wealth distribution. Londoño-Vélez and Ávila-Mahecha (2021) and Londoño-Vélez and Avila-Mahecha (2025) use the Panama Papers data to demonstrate a similar pattern for Colombian taxpayers, finding that they are more likely to incorporate offshore companies following a hike in the wealth tax rate. Furthermore, those disclosed in the leak were subsequently more likely to disclose hidden wealth to the tax authority. Bachas et al. (2024) conduct a similar matching exercise for Ecuador, Honduras and Senegal, finding that taxpayers at the top of the income distribution were more likely to appear in the ICIJ's Offshore Leaks Database. Barake et al. (2024) uncover weak tax compliance behavior among cryptocurrency owners by comparing the list of cryptocurrency owners on the platform Celsius, which was compiled and made public in the wake of the bankruptcy proceedings for the platform, to the information reported in Norwegian tax records. Collin (2021) uses data leaked from a bank in the Isle of Man to show that the users of offshore accounts disproportionally come from wealthier countries, and that are likely to be from the upper end of both the income and wealth distribution themselves. Finally, Chernykh and Mityakov (2017) use leaked transaction data from the Russian Central Bank to show that firms who do business with banks that transact more with offshore financial centers are more likely to engage in tax evasion.

Other work uses leaked data to come up with credible macroeconomic estimates of the holders of offshore assets, to validate estimates derived using other data, or to characterize the network structure of offshore holdings. Both Johannesen et al. (2022) and Bomare and Le Guern Herry (2024) use the Offshore Leaks Database, combined with other data sources, to derive country-by-country estimates of the offshore ownership of real estate in England and Wales. Alstadsæter et al. (2022) use leaked microdata on the ownership of real estate in Dubai to construct similar estimates. By contrast, Alstadsæter et al. (2019) use the distribution of beneficial ownership across tax havens from the Panama Papers to help validate their estimates of bilateral distribution of ownership of offshore financial wealth. Finally, both Fernando and Antoine (2022) and Chang et al. (2023) use the Offshore Leaks Database to analyze the network of offshore ownership, in order to identify which jurisdictions or intermediaries might be prime targets for policy countermeasures.

A final strand of public finance work uses the release of leaks themselves as a shock, to study the impact of that revealing information about a multinational firm's use of subsidiaries in tax havens might have on its stock price, its reputation, and its subsequent operations and reporting (O'Donovan et al., 2019; Schmal et al., 2023), or investment and effective tax rates for private firms (Ortiz and Imbet, 2023). Others have used the shock of the Panama Papers leak to study its impact on the use Panama as a conduit for trade (Figueroa et al., 2024).

In contrast to work focused on tax evasion, a large number of studies have used leaked data to study corruption and financial malfeasance by firms. For example, both Andersen et al. (2022), Marcolongo and Zambiasi (2024) and How Choon et al. (2024) use offshore incorporations from the Offshore Leaks Database as a proxy for elites siphoning off foreign aid disbursements or taking kickbacks for the issuing of oil exploration licenses or from autocratic leaders, respectively. Other work uses leaked microdata, largely for Russia, to create credible risk indicators for corruption. Mironov and Zhuravskaya (2016) use firm-level banking transactions leaked from the Russian central bank to show that firms that repeatedly win government procurement contracts are more likely to engage in "tunneling", a process of using related parties which withdraw large amounts of cash, and that this activity intensifies around municipal elections when firms are more likely to be engaging in bribery.[9] Szakonyi (2023) and Szakonyi (2024) uses leaked information on

[9] Mironov (2013) uses the same data to look at the characteristics of tunneling firms more generally.

vehicle ownership in Russia to identify likely-corrupt politicians through their ownership of luxury cars. Braguinsky and Mityakov (2015) use the same data to study wage under-reporting by firms in Russia, finding that domestic firms engage in a higher degree of under-reporting relative to foreign-owned ones.

Finally, several researchers have used datasets to study the online illicit economy, both in the space of cryptocurrency and in cybercrime. Aloosh and Li (2024) and Saggese et al. (2023) use leaked internal trading records from the Tokyo-based Bitcoin exchange Mt. Gox to study bitcoin wash trading and arbitrage, respectively. Cong et al. (2023b) and Cong et al. (2023a) use information on the perpetrators and victims of ransomware attacks, leaked onto the dark web, to study how characteristics of both have evolved over time. Nershi and Grossman (2023) also use a dark web-leaked dataset of ransomware attacks to analyze whether Russian-affiliated groups increase their activity around elections in western democracies.

The breadth of work described here indicates that leaked data has the potential to address a wide range of important research questions. But the source of data in each of these studies raises unique ethical and data-protection concerns. In the next section we will discuss how a researcher pursuing a similar project might grapple with these, based on our own experiences and learning-by-doing on these issues.

4 Ethical concerns using leaked data

Most empirical social scientists are already familiar with well-defined compliancy routines regarding research ethics for working with data. Those working directly with human subjects, typically in the form of survey work or in the running of randomized field experiments, normally have a standardized IRB process they must undertake at the outset. Those working with secondary data are often working with data that has already been rendered anonymous, are working with an institutional partner with a significant stake in protecting the identity of the data subjects, or can handle the sensitive processing itself (as is often the case when researchers work with tax authorities, for example).

By contrast, working with leaked, stolen or otherwise illicitly obtained data are, to be frank, an ethical wild west. While there are some circumstances where there is a third party "data controller" who will enforce some semblance of ethical discipline, most researchers are in practice left without a pre-existing roadmap to follow in these particular cases. As we have discussed in earlier sections, this has not stopped researchers from carrying on with leaked data research, resulting in discussions about the ethical implications of such work in fields such as political science and computer science (Thomas et al., 2017; Boustead and Herr, 2020; Ienca and Vayena, 2021; Darnton, 2022).

While a researcher's first stop when designing a project around leaked data should be their IRB, in practice IRBs may not always be willing to weigh in on such projects. In some circumstances, they may consider them exempt if the data are already in the public domain.[10] Even if the researcher is working at an institution that has an IRB willing to review a project, neither party may have much experience with these types of data and so the process may be fraught with misunderstanding, particularly if the researcher has not come to the conversation having already anticipated some of the main ethical considerations.

In this section, we lay out the main ethical concerns that arise with working with leaked data and discuss ways that a researcher can accommodate these at the outset of a new project.

4.1 Informed consent

Informed consent is one of the cornerstones of research ethics, allowing data subjects a say in accepting the risks inherent in a new project. With most data gathering exercises, particularly ones that involve interventions that will affect the population of interest, data subjects must be given an opportunity to understand the nature of the underlying research and agree to take part in it. By contrast, as most leaked data has been obtained without the consent of the data subjects (or the consent of the data controller itself), then its use threatens this ethical principle.

One possible solution to this problem would be, post-leak, to seek informed consent from those who were featured in it. Yet there are a number of reasons why this would not only be impractical from the perspective of the researcher, but also seriously threatens the quality of the research. The first is that the primary advantage of leaked data is that it allows the researcher to study behavior that expected to remain hidden at the time it was conducted. This means that ex-post informed consent would likely be biased, due to the fact that those who would have the most interest in keeping their behavior hidden may be less likely to give that consent. Second, depending on the size and complexity of the leak, informed consent can often be impractical, due to either a very large sample size or incomplete contact information.

[10] This is true in the United States: note that section 46.104 of the CFR Common Rule for IRBs exempts research when "The identifiable private information or identifiable biospecimens are publicly available." (45 CFR 46.104)

Finally, informed consent in the context of leaked data work may not only jeopardize the quality of the research (due to selective non-response or withholding of consent by certain data subjects), but in some circumstances it may present a personal risk to the researcher as well. While many journalists have substantial training in how to protect themselves physically, legally and financially during investigative work on people engaged in illicit activity, academics rarely do.[11]

It is worth mentioning, though, that the lack of informed consent for using personal data also applies to all studies using administrative micro data, such as secondary data collected by the authorities. For such data it is accepted that the sheer number of involved parties makes it impossible to receive informed consent prior to utilizing the data, even though they were collected for other purposes than research.

4.2 Maximize benefits and minimize harms

Another tenet of research with human subjects is that a project is on firmer ethical grounding when the risks to subjects is minimized and what risks remain are justified given the public benefit of the research.[12] Much of applied economics research, including large swaths of public finance, is conducted under the implicit assumption that there is public benefit, largely due to its influence on policy. But the implication of harm minimization puts the onus on researchers to demonstrate higher levels of potential public benefit than would normally be required.

For maximizing benefits, there are several things the research team will need to consider before embarking on a project with a new leaked dataset. First, they need to ask themselves: does this data source offer a unique insight into the question at hand or is it just a convenient, novel source of data? Leaked data research may not pass this test if it only replicates a well-established result, or is not being used to address a question with significant social welfare implications (Ienca and Vayena, 2021)? Second, they will need to ascertain whether the leaked data itself would allow the researcher to engage in high-impact science: if the type of data prevents the research team from using robust methods or if the research team has no real outlet for their research, it will be hard to establish that there will be a significant public benefit.

In context of leaked data research, the risks born by data subjects are largely due to the possibility that public identifying information is further divulged by the research

[11] Researchers interested in such defensive measures can find useful guidance on the https://gijn.org/stories/legal-help-for-journalists/.

[12] This is espoused in the Common Rule (45 CFR 46.11)

team, either due to a data breach or through the research itself. Many projects now not only process existing data, but further enhance it: using multiple data sources, both public and private, to create a more comprehensive profile of both the behavior and portfolio of wealth of certain data subjects, which could cause significant harm if leaked. This creates significant tension between the two main goals as maximizing the public benefit of research typically involves much greater levels of both access and processing and thus greater risks posed to data subjects. A descriptive analysis of companies or persons in the Panama Papers data, for example, is less useful than one which merges that information with tax admin data to understand the correlates of offshore ownership (as in Alstadsæter et al. (2019) or Londoño-Vélez and Ávila-Mahecha (2021)), but the latter involves significantly higher risks for the data subjects.

Another consideration is the degree to which the data subjects have already been harmed by the data leak (Boustead and Herr, 2020). In situations where all or most of the data has been made freely available online for anyone to access and use, the bulk of the harm has already, immutably occurred. This is particularly the case when the data uncovers malfeasance of some kind: for example, following the Panama Papers, the combined effect of media reporting, government access to the underlying data, and the data available online has not only lead to several investigations and prosecutions but also the recoup of significant amounts of revenue by tax authorities around the world (McGoey, 2021).

Relatedly, if the main risk posed to data subjects is identity theft or fraud, most will likely have taken countermeasures following the leak (such as changing passwords, identity documents, or company names), minimizing the potential for additional harm caused if a researcher accidentally re-leaked the same information. However, sometimes researchers have unique access to data that is held privately by a third party, typically journalists, investigative outlets, or leaked data outlets. In such situations, the additional potential harm to data subjects is clearer.

4.3 Other considerations

There are other, related concerns that come up less frequently, but may still arise as a research team navigates a leaked data project. The first is the concern that the use of leaked data itself condones and perhaps even incentivizes the act that created the leak (Ienca and Vayena, 2021). For example, if more and more researchers start using data coming from illegal hacks, then this may lead to further hacks in the future. However, there is scant empirical evidence that this is the case, and in most cases where someone has leaked personal private information, their first port of call has not been academic researchers, but journalists, with the goal of getting news coverage. Academic research could then be justified as being an attempt to extract the most good out of a leak motivated by other factors. This would be much harder to justify in instances where researchers pay for hacked or stolen data, particularly if they knew it was illicitly-obtained at the time of aquisition.

Similarly, there may also be concerns that working with these types of data may erode the relationships researchers have with institutions that are often subject to these leaks. For example, research using data leaked from a tax administration might make other tax administrations more hesitant in general (particularly after large-scale leaks to investigative reporters, such as IRS tax return data that was turned over to ProPublica in 2021).

For both of these, researchers will need to keep these reputational concerns in mind, making an active, transparent argument for why the project is still worthwhile given these risks (or how those risks themselves might be mitigated, by pre-committing not to work with certain types of data).

4.4 Characteristics of a good project

In practice, the best way to grapple with these questions is publicly, typically in the form of an ethical statement, to be drafted at the outset of the project and included either in the main draft of the research or as part of an easily accessible online appendix.[13] That statement should lay out the researcher's position on why the public benefit of the research was worth any additional risk posed to the research subjects as well as the reasoning behind the side-stepping of any requirements (such as consent). Ideally, this statement should be explicit about the trade-offs between data processing and privacy. While it is easy for a researcher to make a general case that a given research question is worth undertaking despite the risks to data subjects, a good ethics statement should also explain why the level of processing being undertaken is essential.

Other elements of the ethical statement should include a discussion of the uniqueness of the data for the question at hand and explain why this data can better answer the research question than other, more legitimate data sources. Related, researchers should

[13] For a recent examples of an ethical statement, see Ethics and data privacy statement for research conducted using data from the C4ADS Dubai Property Registry for the academic paper "Who Owns Offshore Real Estate? Evidence from Dubai"

explain why informed consent was not obtained in this specific case. The public ethics statement can also form the basis of submissions to an IRB (should your institution have one) and work on any data protection approvals, which we discuss in the next section.

5 Legal formalities and considerations

The use of leaked data in research introduces a host of legal considerations that parallel many of the ethical concerns already discussed. Just as ethical evaluations focus on balancing the potential societal benefits of research with the risks to data subjects, legal formalities require a detailed analysis of how personal data are handled, processed and protected. For this, the process of developing a Data Protection Impact Assessment (DPIA) is a critical tool and required for researchers within the extended EU region. Personal data refers to any information that can directly or indirectly identify a natural person, such as names, identification numbers, location data, or online identifiers.[14] While much of this section focuses on data protection law under the GDPR, researchers must also be aware of other legal concerns that go beyond privacy. These include potential exposure to criminal liability, such as receiving, handling, or publishing data that was obtained through illegal means, and civil liability, for instance due to breaches of intellectual property law. In some jurisdictions, accessing or reusing data that is known or suspected to originate from a criminal offense may itself be unlawful. Where uncertainty exists, particularly around source provenance, international access, or the presence of sensitive data, researchers are advised to seek legal counsel and consult national guidelines. Ethical rigor is necessary, but not always sufficient, to ensure that research involving leaked data proceeds within legal boundaries.[15] Without legal review, researchers may inadvertently violate existing provisions, exposing themselves and their institutions to significant legal and financial risks.

5.1 General formalities

One of the primary frameworks governing the use of personal data in research is the General Data Protection Regulation (GDPR), enforced within the European Union and the European Economic Area since 2018 to safeguard individuals' privacy and regulate

[14] The concept of personal data is defined under the GDPR, Article 4.1.

[15] Guidance from national data protection authorities can help delineate these boundaries. For example, in France, the Commission Nationale de l'Informatique et des Libertés (CNIL) encourages researchers to assess not only compliance with data protection principles, but also whether the source and dissemination of a dataset may itself involve a criminal offense or legal prohibition.

the processing of personal data. This is the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. The primary purpose of these regulations is to enhance individuals' control over their personal data, ensure transparency and accountability in data processing, and harmonize data protection laws within the EU. GDPR sets strict requirements for the lawful collection, storage, and use of personal data, emphasizing principles such as data minimization, purpose limitation, and security.

When available, it is advisable that researchers consult those in their institution tasked with maintaining GDPR compliance or data protection more generally, typically Data Protection Officers, legal counsel, and their IT department. Researchers outside the EU may still, and in most cases should still, seek legal counsel to ensure their processing of sensitive information is permitted. In some cases, GDPR will also apply to researchers outside the EU, if they process personal data about data subjects in the EU. The regulations and guidelines stipulated for the GDPR may also serve as a helpful guide for researchers based outside of the EU.

Personal data refers to any information that can directly or indirectly identify a natural person. This includes names, identification numbers, location data, and online identifiers, among others. Special categories of personal data, as defined in GDPR Article 9(1), include sensitive information such as political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data used for unique identification, health data, and data concerning a person's sex life or sexual orientation. Due to the heightened risks to individuals' rights and freedoms, the processing of such data is generally prohibited unless specific exceptions apply, such as explicit consent, legal obligations, or substantial public interest under strict safeguards. Data considered anonymous falls outside the scope of the GDPR. Anonymous data are defined as information that cannot be linked to an identified or identifiable person, even if theoretical re-identification is possible, as long as such re-identification would require disproportionate effort. Pseudonymized data such as data with encrypted personal identifiers, while masked with no name or actual identification number, is still considered personal data under GDPR if there is a risk of re-identification.[16]

Particular emphasis is placed on the personal data with high risk to the rights and freedoms of natural persons, that is, when data processing could lead to significant harm,

[16] See Henriksen et al. (2024) for a thorough definition of the different data types, as well of the concept personal data.

such as discrimination, identity theft, financial loss, or emotional distress.[17] These risks are particularly acute when processing sensitive data or linking datasets that increase re-identification potential.

When researchers plan to utilize personal data with high risk, a Data Protection Impact Assessment (DPIA) must be initiated. This is a systematic process designed to identify, assess, and mitigate risks to the privacy and rights of individuals when personal data are processed.[18] If is uncertain whether a DPIA is necessary, it may be wise to conduct one regardless, as it serves as a valuable tool for ensuring compliance with data protection laws.

The process of carrying out a DPIA is useful for identifying and evaluating any risks to individuals' privacy and rights arising from data processing activities. As part of the process, researchers must document that the data processing is necessary for its intended purpose and that the level of processing is proportionate to the benefits, and describe implemented measures to mitigate risks, such as such as data pseudo-anonymization, encryption, or restricted access. For this process, researchers can lean on the ethical statement process we described in Section 4.

An important part of the legal compliance process is to thoroughly describe accurately both the data and variables, the data flow, and how the project adheres to the principle of data minimization, limiting the use of leaked data or variables to only what is strictly necessary to achieve research objectives. From experience, a clear description of this is also necessary and helpful prior to any constructive dialogue with legal counsel. An important advice for this process is that writing a DPIA is a legal process, not a scientific process within a given scientific field. That means that the descriptions of the dataset and variables need to answer questions that researchers might not be accustomed to answer, but that are relevant when assessing compliance with privacy requirements, like what programs are used to process the data, where are the different data sources stored, who can access the different datasets and what is the nature of the data.

Researchers must also articulate the steps they will take to minimize the risks to data subjects. This involves implementing robust security protocols to prevent the research product or intermediary data from being leaked or misused, such as the use of secure

[17] See guidelines here on Guidelines on the process of Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679. https://ec.europa.eu/newsroom/article29/items/611236

[18] A detailed explanation of the DPIA requirements and process is available here: https://gdpr.eu/data-protection-impact-assessment-template/, with a starting template for a DPIA available here: https://gdpr.eu/wp-content/uploads/2019/03/dpia-template-v1.pdf. Many academic institutions also have available samples of previously approved DPIAs that can be very helpful in this process.

servers or zero-knowledge encryption file sharing. Further, public outputs from the research, such as papers or presentations, are carefully anonymized to avoid re-identifying individuals or exposing sensitive information. Finally, researchers should conduct only the necessary degree of data processing to address the research question. For example, if linking to external datasets is not essential for answering the question, researchers should refrain from doing so.

A common misperception, also frequent among lawyers, is that a project cannot be conducted if there are any risks remaining and that the purpose of the DPIA process is to remove all risk. That is not the case.[19] GDPR does not demand the elimination of all risks, but instead requires proportionate measures to reduce risks to an acceptable level, balancing the potential societal or organizational benefits of processing against the likelihood and severity of harm. Based on this, the data controller (the academic institution, on the advice of the Data Protection Officer at the institution) can make an informed decision on whether the potential societal gains of the project exceed the potential risk to the involved data subjects.[20] Again, much of this groundwork can be covered during the ethical review.

The DPIA process should be documented for transparency and accountability, and the team must monitor and update the DPIA regularly as part of the ongoing works. It is quite common for researchers to adjust their research plan as they develop new processing strategies, add new people to the research team, or need access to new datasets. A continuous evaluation of these issues is necessary throughout the project to evaluate the classification of data (e.g., anonymous vs. identifiable) as linking datasets or further processing may change its legal status.

5.2 Specific data privacy issues related to leaked data

Analysis of leaked data often offer unique opportunities to make progress on issues of great public interest. This in turn provides useful insights for policy makers, tax administrations, and researchers worldwide, and can justify a deviation from data privacy requirements. For example, the secrecy provided by tax havens can contribute to the erosion of the tax base. A major problem with tax evasion is that capital's location is

[19] See the recent Skatteforsk-Note (in Norwegian) by Aanestad et al. (2025) for a more thorough documentation and discussion of common misperceptions regarding GDPR interpretation that hinders the public sector's utilization of own data.

[20] However, if a Data Protection Impact Assessment (DPIA) concludes that a high risk remains despite the safeguards implemented, the institution is to consult with the relevant Data Protection Authority before proceeding with the data, as follows from Article 36 of the GDPR.

hidden in countries with limited public information about capital and its owners. Neither the owners of capital nor the authorities of tax havens have an interest in transparency regarding ownership. It is, therefore, difficult to obtain information about such matters through ordinary channels.

Analogous to researchers setting up experiments like randomized control trials (RCTs), researchers using leaked data must think carefully about their data collection, data processing, and analysis in advance due to the sensitivity of the data. The formal legal procedures required, like the DPIA in the EU, are also great avenues for the structuring and executing of this thinking. In some ways, the DPIA is the pre-analysis plan of the researcher working on leaked data, except with a focus on data protection and privacy rather than a focus on specified hypotheses.

The DPIA is the best place to discuss the roadmap for processing the data, allowing the researcher to discuss key concerns and tradeoffs and highlight any legally acceptable reasons for deviating from legal requirements. One example is informed consent, the requirement to inform affected parties about the processing of their personal data. As discussed before, this is often not possible due to the large number of individuals in the data and because of incomplete contact. One solution is announcing the research project online, with an invitation to people who know that they are in the data to reach out.[21] This is, however, not possible in all cases, both for practical and security reasons for the research team and others involved with the leak, as was the case in our previous work with the C4ADS Dubai data we describe above.

A DPIA for a leaked data project should also discuss the additional burden that processing might have on subjects of the study, similar to the potential harm discussion from the previous section. This burden will depend on the nature of the leak, if it is widely known and publicly available and on the behavior that is revealed. The focus of a DPIA will be on reducing this additional burden as much as possible by, for instance, pseudo-anonymizing the data. Some leaks are already widely known, like the Panama Papers, and research on these might even relieve some of the burdens on the subjects of the leak. For instance, highlighting lawful, justified reasons of being named in a leak, something which is not normally in the interest of media organizations and other groups, can achieve this.

An explicit Data Management Plan can serve as the foundation for the DPIA, including a comprehensive plan for managing the leaked data, covering acquisition, storage,

[21] See for instance: https://www.nmbu.no/forskning/skatteforsk/forsker-pa-celsius-data

cleaning, and analysis. This should detail data flow, tools, and access controls. For this it is important to understand the data generating process, insofar as it is possible for leaked data, which often are surrounded by a cloud of opacity on the provenance, for good reasons, often to protect the source. In the cases where the source is unknown or cannot be made public, it is essential to confirm the relevance and reliability of the data from other sources.

An additional challenge when working on leaked data is that the subjects of study are identifiable in the material. While researchers normally resort to working on anonymized or pseudo-identified datasets when utilizing large, administrative data sources, this preprocessing is (with a few exceptions) not available to researchers who utilize leaked data. Data owners may then be hesitant to share data for research, and the personally identifiable information nature of the data can trigger strong obligations for data protection and prevent sharing. One solution can be for the researchers to sign Memorandums of Understanding and Non-Disclosure Agreements with the data owners to regulate data protection, access, use, and ownership.

Data protection requirements can make linking the data to other administrative data, which often are necessary to get the necessary background information, a painstaking process within the current regulations. Governments around the world have become increasingly open to sharing administrative data with researchers, and they have, in some countries, also allowed for the linking of these administrative data with non-administrative data sources, like survey results or data from private companies. This linking can, in theory, be done for leaked data as well, but it requires thorough preparations. The key consideration is the risk of indirectly identifying the individuals hiding behind the encrypted IDs in the large, linked administrative data sets, and establishing a data flow with clear limits on who access the identifiable information in the leaks vs. the pseudo-anonymized administrative data.

Working with leaked data requires even more emphasis on security. The project must implement robust data security measures, such as encryption and restricted access, to prevent unauthorized use or security breaches. It is vital that the identity of the original sources of the leak is protected when applicable, and ideally the researchers should not know their identity. the project must also comply with institutional and regulatory standards for storing the data, and ideally also enforce even stronger security measures. Throughout the project, it is important to clearly document how data flows between systems or collaborators.

Leaked data often come in a messy format, so cleaning and processing is a crucial part of the research process. It is then important to continuously address inconsistencies, errors, or gaps in the leaked data while maintaining its integrity and to clearly document the cleaning process. The researchers should also confirm, as far as possible, the authenticity of the leaked data and cross-check, and document, its quality against other sources to enhance reliability. Another factor to keep in mind is to only use institutionally approved tools for analysis, particularly for sensitive tasks like entity resolution or name classification. For example, tools for name analysis should be vetted for compliance with security and ethical guidelines. Throughout the process, it is crucial to maintain comprehensive documentation detailing the rationale, processes, and safeguards applied to using leaked data.

In some cases working with leaked data may entail security threats also for the researchers, and personal security and potential measures for the team must be taken into account at initial stages and throughout the project. For instance, when the original source of the leak is unknown to the researchers, it is recommendable to make this very clear in the public space when presenting the project and results.

6 Summary: the responsible use of leaked data in research

The use of leaked data in research offers unparalleled opportunities to address pressing societal questions, uncover hidden patterns, and inform evidence-based policy-making. However, it also introduces significant challenges, requiring researchers, institutions, and journals to navigate complex legal, ethical, and methodological landscapes. Balancing the potential benefits of such research with the risks to individuals, institutions, and society necessitates a clear framework to guide responsible data use.

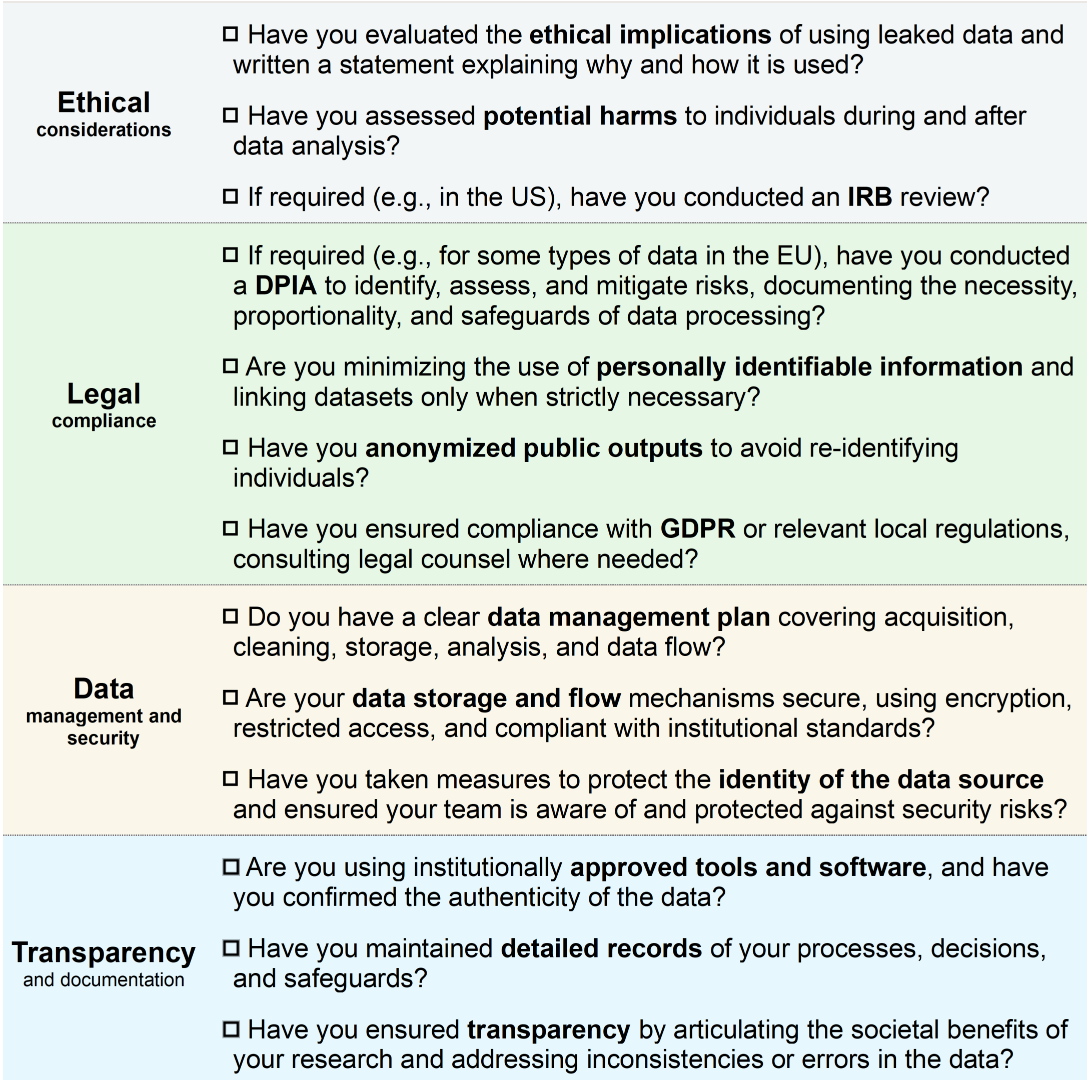
We here suggest a summary of the best practices outlined in this paper, presented as a checklist in Figure 2 to aid researchers, administrators, personal protection officers at academic institutions, and journal editors in handling leaked data responsibly. The checklist aims to empower researchers by providing actionable steps to ensure compliance with ethical and legal standards, reduce risks to data subjects, and maximize the societal value of their work. For institutions, it emphasizes the importance of structured support, including access to legal counsel and data protection officers to assist researchers in navigating these challenges. Journals, too, play a critical role in fostering transparency and accountability. Just as the submission of replication files has become standard practice, journals could require DPIAs for studies involving sensitive or leaked data to ensure rigorous evaluation of risks and safeguards.

Before receiving data, researchers must evaluate and reflect upon the ethical implications of using leaked data, ensuring compliance with institutional and regulatory standards. This involves preparing an ethical statement that explains why and how the leaked data will be used, with a focus on transparency. This statement should be a dynamic document, updated throughout the project to reflect new insights or findings about the data. Researchers must proactively minimize harm during the project, including risks of re-identification or stigmatization, and ensure that research findings are presented in ways that respect the rights and dignity of data subjects.

On the legal side, compliance with regulations like GDPR is essential. Developing a DPIA is a critical step to identify, evaluate, and mitigate risks associated with leaked data processing. The DPIA should document the necessity and proportionality of data processing, outline measures to mitigate risks (e.g., pseudonymization, encryption), and evaluate the classification of data (e.g., anonymous vs. identifiable). Institutions play a crucial role here by providing templates, expert guidance, and oversight to ensure that researchers meet these obligations.

Looking ahead, the responsible use of leaked data must remain adaptive to an evolving geopolitical and institutional context. The framework and best practices outlined in this paper are grounded in our own recent research experience, developed in a period where many governments have become more open to collaboration with researchers and to sharing administrative data. This landscape, however, may be changing. In a time of rising geopolitical tensions, data governance is becoming more politicized, and access to administrative data may be granted more selectively. At the same time, leaks can reflect strategic intent, whether to expose misconduct, influence public narratives, or target political opponents. This is not an argument against the use of leaked data in research, but rather a reminder of the importance of strong routines and safeguards, as with any sensitive data. Researchers, institutions, and journals should remain agile, continuously reassessing both the potential harms and societal value of working with leaked data, and ensuring that ethical and legal frameworks evolve in response to emerging risks and challenges.

Figure 2: Checklist for using leaked data



Ethical considerations	□ Have you evaluated the **ethical implications** of using leaked data and written a statement explaining why and how it is used? □ Have you assessed **potential harms** to individuals during and after data analysis? □ If required (e.g., in the US), have you conducted an **IRB** review?
Legal compliance	□ If required (e.g., for some types of data in the EU), have you conducted a **DPIA** to identify, assess, and mitigate risks, documenting the necessity, proportionality, and safeguards of data processing? □ Are you minimizing the use of **personally identifiable information** and linking datasets only when strictly necessary? □ Have you **anonymized public outputs** to avoid re-identifying individuals? □ Have you ensured compliance with **GDPR** or relevant local regulations, consulting legal counsel where needed?
Data management and security	□ Do you have a clear **data management plan** covering acquisition, cleaning, storage, analysis, and data flow? □ Are your **data storage and flow** mechanisms secure, using encryption, restricted access, and compliant with institutional standards? □ Have you taken measures to protect the **identity of the data source** and ensured your team is aware of and protected against security risks?
Transparency and documentation	□ Are you using institutionally **approved tools and software**, and have you confirmed the authenticity of the data? □ Have you maintained **detailed records** of your processes, decisions, and safeguards? □ Have you ensured **transparency** by articulating the societal benefits of your research and addressing inconsistencies or errors in the data?

Notes: This checklist provides a structured guide to ensure ethical, legal, and methodological rigor when using leaked data in research. It is designed for researchers to navigate the complexities of working responsibly and transparently with sensitive datasets. Note that this should be regarded as a starting point for researchers, academic institutions, and journal editors, keeping in mind that the need for adaption to each specific case.

References

Aanestad, M., Alstadsæter, A., and Ulloa, H. (2025). Hvordan offentlig sektor kan utnytte egne data bedre. Skatteforsk – Centre for Tax Research.

Aloosh, A. and Li, J. (2024). Direct evidence of bitcoin wash trading. *Management Science*.

Alstadsæter, A., Collin, M., Planterose, B., Zucman, G., and Økland, A. (2024). Foreign investment in the Dubai housing market, 2020-2024. EU Tax Observatory Note.

Alstadsæter, A., Johannesen, N., and Zucman, G. (2019). Tax evasion and inequality. *American Economic Review*, 109(6):2073–2103.

Alstadsæter, A., Planterose, B., Zucman, G., and Økland, A. (2022). Who owns offshore real estate? evidence from Dubai. EU Tax Observatory Working Paper No. 1.

Andersen, J. J., Johannesen, N., and Rijkers, B. (2022). Elite capture of foreign aid: Evidence from offshore bank accounts. *Journal of Political Economy*, 130(2):388–425.

Bachas, P., Collin, M., Flores, T., Scot, T., and Lyu, H. (2024). Offshore data leaks and tax enforcement in developing countries. *Equitable Growth, Finance and Institutions Notes*.

Barake, M., Le Pouhaër, E., and Økland, A. (2024). Who owns cryptocurrency? Technical report, Skatteforsk Working Paper #10.

Bomare, J. and Collin, M. (2025). When bankers become informants: Behavioral effects of automatic exchange of information. Memo.

Bomare, J. and Le Guern Herry, S. (2024). Avoiding transparency through offshore real estate: Evidence from the UK. Technical report, Working Paper.

Boustead, A. E. and Herr, T. (2020). Analyzing the ethical implications of research using leaked data. *PS: Political Science and Politics*, 53(3):505–509.

Braguinsky, S. and Mityakov, S. (2015). Foreign corporations and the culture of transparency: Evidence from Russian administrative data. *Journal of Financial Economics*, 117(1):139–164.

Caruana-Galizia, P. and Caruana-Galizia, M. (2016). Offshore financial activity and tax policy: evidence from a leaked data set. *Journal of Public Policy*, 36(3):457–488.

Chang, H.-C. H., Harrington, B., Fu, F., and Rockmore, D. N. (2023). Complex systems of secrecy: the offshore networks of oligarchs. *PNAS nexus*, 2(3):pgad051.

Chernykh, L. and Mityakov, S. (2017). Offshore schemes and tax evasion: The role of banks. *Journal of Financial Economics*, 126(3):516–542.

Collin, M. (2021). What lies beneath: evidence from leaked account data on how elites use offshore banking. *Brookings Global Working Paper Series*.

Collin, M., Hollenbach, F. M., and Szakonyi, D. (2023). The end of Londongrad? The impact of beneficial ownership transparency on offshore investment in UK property. Technical Report 2023/11, WIDER Working Paper.

Cong, L., Grauer, K., Rabetti, D., and Updegrave, H. (2023a). *Blockchain forensics and crypto-related cybercrimes*. Digitally published handbook.

Cong, L. W., Harvey, C. R., Rabetti, D., and Wu, Z.-Y. (2023b). An anatomy of crypto-enabled cybercrimes. Technical report, National Bureau of Economic Research.

Darnton, C. (2022). The provenance problem: Research methods and ethics in the age of WikiLeaks. *American Political Science Review*, 116(3):1110–1125.

Ederer, F., Goldsmith-Pinkham, P., and Jensen, K. (2024). Anonymity and identity online. Working Paper.

Fernando, G. A. and Antoine, M. (2022). The network structure of global tax evasion evidence from the Panama Papers. *Journal of Economic Behavior & Organization*, 197:660–684.

Figueroa, M. B., Fisman, R. J., Knill, A. M., and Mityakov, S. (2024). Deterrence and displacement in offshore trade: Evidence from the Panama Papers leak. *Available at SSRN 4698314*.

Griffin, J. M. and Kruger, S. (2024). What is forensic finance? *Foundations and Trends in Finance*, 14(3).

Henriksen, A. M., Gulbrandsen, H. P., and Ulloa, H. (2024). The difference between personal data and anonymous data. Skatteforsk – Centre for Tax Research.

How Choon, T., Marcolongo, G., and Pinotti, P. (2024). Money talks to autocrats, bullets whistle to democrats: Political influence under different regimes.

Ienca, M. and Vayena, E. (2021). Is it ethical to use hacked data in scientific research? Unpublished.

Johannesen, N., Miethe, J., and Weishaar, D. (2022). Homes incorporated: Offshore ownership of real estate in the U.K. *CESifo Working Paper*, (10159).

Kavakli, K. C., Marcolongo, G., and Zambiasi, D. (2023). Sanction evasion through tax havens. Technical Report 212, BAFFI CAREFIN Centre.

Lafitte, S. (2024). The market for tax havens. EU Tax Observatory Working Paper No. 22.

Londoño-Vélez, J. and Ávila-Mahecha, J. (2021). Enforcing wealth taxes in the developing world: Quasi-experimental evidence from Colombia. *American Economic Review: Insights*, 3(2):131–148.

Londoño-Vélez, J. and Avila-Mahecha, J. (2025). Behavioral responses to wealth taxation: Evidence from Colombia. *Review of Economic Studies*.

Marcolongo, G. and Zambiasi, D. (2024). Offshore finance and corruption in oil licensing. *Energy Economics*, 137:107787.

McGoey, S. (2021). Panama Papers revenue recovery reaches $1.36 billion as investigations continue. https://www.icij.org/investigations/panama-papers/panama-papers-revenue-recovery-reaches-1-36-billion-as-investigations-continue/. Accessed: 12/13/2024.

Mironov, M. (2013). Taxes, theft, and firm performance. *The Journal of Finance*, 68(4):1441–1472.

Mironov, M. and Zhuravskaya, E. (2016). Corruption in procurement and the political cycle in tunneling: Evidence from financial transactions data. *American Economic Journal: Economic Policy*, 8(2):287–321.

Nershi, K. and Grossman, S. (2023). Assessing the political motivations behind ransomware attacks. Technical report.

Obermaier, F., Obermayer, B., Wormer, V., and Jaschensky, W. (2017). About the Panama Papers. https://panamapapers.sueddeutsche.de/articles/56febff0a1bb8d3c3495adf4/. Accessed: 12/16/2024.

Omartian, J. D. (2017). Do banks aid and abet asset concealment: Evidence from the Panama Papers. Technical report.

Ortiz, M. M. and Imbet, J. F. (2023). Private firms and offshore finance. Technical report.

O'Donovan, J., Wagner, H. F., and Zeume, S. (2019). The value of offshore secrets: Evidence from the Panama Papers. *The Review of Financial Studies*, 32(11):4117–4155.

Page, M. T. (2020). Dubai property: an oasis for Nigeria's corrupt political elites. Technical report, Carnegie Endowment for International Peace.

Saggese, P., Belmonte, A., Dimitri, N., Facchini, A., and Böhme, R. (2023). Arbitrageurs in the bitcoin ecosystem: Evidence from user-level trading patterns in the Mt. Gox exchange platform. *Journal of Economic Behavior & Organization*, 213:251–270.

Schmal, F., Schulte Sasse, K., and Watrin, C. (2023). Trouble in paradise? disclosure after tax haven leaks. *Journal of Accounting, Auditing & Finance*, 38(3):706–727.

Szakonyi, D. (2023). Corruption and co-optation in autocracy: Evidence from Russia. *American Political Science Review*, pages 1–18.

Szakonyi, D. (2024). Opposition rule under autocracy: Evidence from Russia.

Thomas, K., Moscicki, A., Margolis, D., Paxson, V., Bursztein, E., Li, F., Zand, A., Barrett, J., Ranieri, J., Invernizzi, L., Markov, Y., Comanescu, O., and Eranti, V. (2017). Data breaches, phishing, or malware?: Understanding the risks of stolen credentials. In *CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Securit*, pages 1421–1434.