Marlene Saemann*, Daniel Theis, Tobias Urban, and Martin Degeling

Investigating GDPR Fines in the Light of Data Flows

Abstract: While GDPR related fines to big companies like *Amazon* or *Google* have seen widespread media attention, data protection authorities have issued several hundred more penalties since 2018. This work analyzes 856 fines and their summaries provided by the *CMS Law GDPR Enforcement Tracker*. We extend the methodology of previous work that evaluated GDPR fines and, in particular, explore the fines in the light of data flows and we perform a detailed categorization. Our analysis shows that it is a combination of technical and organizational issues that are involved when a fine is imposed. Moreover, data protection authorities more often react to data subjects' complaints when data breaches become public and when health-related data is involved. We further show that the root causes for fined data processing lie in the early data life cycle phases (e.g., data collection). Here, organizational problems are more prevalent (601 fines) than technical issues (314 fines), while technical issues are mentioned more often in later life cycle phases (e.g., retention, access and usage). Especially mistakes in the early phases of the data collection process (e.g., lacking a legal basis) and unauthorized disclosure in later phases are fined. We cluster the most frequent words and analyze relations to understand where data controllers put personal data at risk. The results confirm that access management is a common problem that results in the unintended disclosure of data.

Keywords: data protection, privacy, GDPR fines, personal data life cycle, word frequency analysis, NLP, access management, health related data

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

***Corresponding Author: Marlene Saemann:** Bosch, E-mail: marlenebarth@googlemail.com
Daniel Theis: Institute for Internet Security, E-mail: theis@internet-sicherheit.de
Tobias Urban: Institute for Internet Security & secunet Security Networks AG, E-mail: urban@internet-sicherheit.de
Martin Degeling: Ruhr University Bochum, E-mail: martin.degeling@ruhr-uni-bochum.de

1 Introduction

It has been four years since the *General Data Protection Regulation* (GDPR) came into effect on May 25, 2018. The GDPR set out to standardize data protection rules within the European Union and has impacted businesses and services worldwide. To avoid fines and comply with the new legislation, institutions had to adjust their processes of handling personal data on an organizational and technical level. Research has shown that, for example, on the web, the regulation was quickly adopted by a majority of data controllers [9].

One problem that data protection experts in institutions face is that privacy and security in complex systems concern both technical and organizational processes. Therefore, the responsibility for applying privacy and security safeguards is not just an engineering issue but has to be an interdisciplinary effort [13]. In the light of limited awareness, consulting, and implementation capacities, and with ongoing legal debates regarding the GDPR's interpretation, it is crucial to understand which challenges organizations face.

This work provides respective insights by analyzing GDPR fines ($n = 856$) issued since 2018. To do so, we utilize the *CMS Law GDPR Enforcement Tracker*[1] (from now on, ET) to establish a broader understanding of the regulatory violations that lead to fines. The ET provides an overview of GDPR fines that data protection authorities have imposed since the GDPR took effect. We use the ET to analyze the fines quantitatively, then categorize them according to the underlying problem that caused them, and finally provide insights on how to make data flows compliant with the law. The results show that issues leading up to fines are predominantly organizational rather than technical and that respective DPA actions are primarily based on customer complaints and unwanted disclosure of data. We also find that the cause for fines often lies in the early phases of the personal data life cycle. Moreover, they are primarily related to organizational issues. Issues in later life cycle phases, in turn, are least often the reason for

1 https://www.enforcementtracker.com/

fines but are more frequently attributable to technical causes.

Previous work has already looked at the imposed fines and provided an overview and statistical analysis [3, 20–22, 33]. However, these works do not investigate the data processing details that lead to fines. In this work, we substantially extend previous research by providing an in-depth analysis of exactly these circumstances.
To summarize, the main contributions of our work are:

- We develop a systematization scheme to categorize the GDPR fines of the ET for extended analysis. The scheme covers a detailed categorization of the fined issue, why DPAs investigated it (e.g., data breach or customer complaint), and whether the issue was rooted in organizational or technical problems.
- Based on a data life cycle model, we group the analyzed fines into different processing phases. Knowing when issues that lead to fines occur in a process can help project teams and data protection experts to focus their efforts.
- Finally, we analyze which GDPR principles are commonly infringed in the analyzed cases. We do so by performing a word frequency analysis. We map the identified word stems of all fines to GDPR principles (Art. 4 GDPR). Furthermore, we discuss the issues that lead to fines and two case studies.

2 Background and Related Work

Since the GDPR went into force, many institutions have faced the problem of setting up new IT systems and processes or revising them to meet the requirements and objectives of data protection and privacy. Kutyowski et al. [16] discussed critical challenges in the GDPR's implementation and provided a list of conflicts between the legal concepts of the GDPR and information security technology. Furthermore, Texeira et al. [1] identified GDPR subjectivity and lack of required technology as implementation blockers. However, risk identification, process documentation, or awareness training as enablers.

2.1 GDPR Implementation Support

Many researchers have worked on developer-focused solutions for GDPR implementation and proposed solutions for bridging the gap between the law and the realization process. Sarkar et al. [24] discussed implementation challenges in the context of the right to erasure and presented data flow tracking and managing data and its duplicates as possible solutions. Alshammari and Simpson [2] propose an abstract personal data life cycle model (APDL) for personal data, which is meant to trace and manage personal data. Furthermore, Huth and Matthes [14] presented eight techniques for improving data protection and tested their suitability when integrated into existing systems. In 2018, Senarath et al. [26] analyzed developers' problems when integrating privacy-enhancing technologies. They derived a guide for mitigation of these problems and improvement of privacy. In a separate study, Senarath et al. [25] also investigated developers' perceptions differ from users in terms of privacy expectations. Based on their analysis, they created a guideline to improve the way developers can identify and better understand their users' expectations. Finally, Li et al. [17] analyzed a discussion board for Android developers where the researchers examined implementation issues related to privacy. Based on their results, they created guidelines for other Android developers that cover both development and distribution. Furthermore, they aligned them with the users' expectations. Tahaei and Vaniea [28] examined privacy-related questions developers ask on *StackOverflow*, in a similar approach. Among other things, they performed a quantitative evaluation of 1,733 questions on *StackOverflow* that contained the word "privacy", as well as a qualitative approach of 40 randomly selected questions from the same dataset, which privacy experts analyzed. They show that the term "access control" is the topic with the most questions in total, with 40% of all questions analyzed.

Beyond solving individual aspects, researchers developed tools and platforms meant to monitor and improve the data protection compliance of whole systems. Such solutions are for example reports from the static analysis of software that is customized to the recipients [12] or continuous testing [18]. Furthermore, a platform developed by Piras et al. [19] allows combining various tools to achieve data protection compliance. In 2020, Li et al. [27] operationalized parts of the GDPR and developed a tool that tests for specific privacy requirements. Khaitzin et al. [15] presented a way to make privacy more usable for "Big Data companies" by optimizing performance using intermediate representations, thus providing an alternative or at least a mitigation for privacy/performance trade-offs. Furthermore, Chander

et al. [7] reviewed the compliance costs and evaluated the feasibility of data protection enforcement.

Another aspect related to our paper is the work of the data protection supervisory authorities (henceforth, DPAs). Barrett and other researchers [3, 20, 33], analyzed the fines issued by the authorities in an early stage of the GDPR adoption. In addition, two articles by Ruohonen and Hjerppe [21, 22] predicted the evolution of fines, for instance, in magnitude and frequency. Finally, Félix and Wright [11] presented a conflict between data protection in the broad mass and the authorities' work. They found that interpretations vary between authorities and are inconsistent in the overall picture.

2.2 Databases of GDPR Fines

Multiple projects collect and summarize public information on GDPR fines imposed by DPAs within the EU, European Economic Area (EEA), including Norway, Lichtenstein, Iceland, and the UK. Hence, fines imposed under national or non-European laws, non-data protection laws (e.g., competition laws or electronic communication laws), and pre-GDPR-laws are not listed or only sporadically added.

The private company *CMS.Law* (CMS) provides the *CMS Law GDPR Enforcement Tracker* (ET) that is , based on the information provided by CMS upon request, maintained by professional privacy lawyers as well as law students that have experience in the fields of data protection. For each fine, the ET dataset defines an "Enforcement Tracker ID" *(ETid)* to uniquely identify each case. Each entry additionally contains information about the country and DPA of allocation, date of the decision, amount of the imposed fine, name of the controller or processor that received the fine as well as its business sector, the quoted GDPR article, a generalized type of fine (see Table 1), the link to the data source, and a summary text.

Another database, that has a similar purpose, is the *GDPRhub* of *noyb.eu*.[2] It clusters the fines in "decisions by article", "DPA divisions", "court decisions", and further provides additional GDPR knowledge. GDPRhub contains a higher amount of metadata for each fine than the ET, e.g. the national case number. Everyone can contribute by editing pages of the GDPRhub, even without a user account. This circumstance makes the GDPRhub a less suitable source for our analysis.

Privacy Affairs by *Zisk Web*[3] is a database provided by international cybersecurity professionals, tech journalists, and privacy advocates. The "GDPR Fines Tracker and Statistics" is one of their focus areas. Privacy Affairs provides the same information and metadata as the ET but is updated less frequently.

While all databases provide a similar number of DPA decisions, we chose the *ET* for our analysis because we observed that the entries were of consistent quality and categorization. Upon request, CMS confirmed that the ET is "*operated and updated by professional privacy lawyers and law students that also have experience in the fields of data protection.*"

CMS [8] has published a report on the ET data. It provides a high-level overview of the distribution of fines in countries and over time. We provide a similar overview in Section 3.2. Besides that, the report focuses on informing businesses by highlighting procedural information like the fact that many fines have been reduced after being challenged in court. It also summarizes essential fines and common themes based on industry sectors like "hospitality," "health care," and "industry and commerce." This industry perspective is helpful for businesses to learn about common pitfalls in their industry and inform legal departments.

Our analysis goes beyond previous work and the CMS report by focusing on data flows and processes to provide privacy experts in software engineering and other fields information about common problems. Our additional annotations and reproducible methodology allow for an analysis of specific technical and organizational difficulties that lead to a fine. Based on the dataset, we can also contextualize previous research results and confirm that the topics raised during development (which manifest on *StackOverflow* [28]) are similar to issues that led to data protection fines. We build bridges between the ET dataset and existing methods, such as the APDL, and further extend the dataset with NLP. Compared to the CMS that focuses on different industry sectors, we identify common themes and problems across industries but focus on the (technological) problems that lead to a fine. While the CMS report is an excellent resource for understanding the breadth of data protection issues that lead to fines, our study provides a more in-depth analysis with a more detailed dataset and additional evaluations concerning data flows and technical issues.

2 https://gdprhub.eu

3 https://www.privacyaffairs.com/gdpr-fines

3 Overview and Categories of Violations

This section provides a brief overview of GDPR fines documented in the ET. We describe how we extend the ET dataset by adding (1) the technical or organizational origin of a fine, (2) the cause of investigation that led to a fine, and (3) a more detailed categorization of fines and report on findings according to this new categorization.

3.1 Dataset and Fine Overview

As mentioned before, each fine in the ET maps to one "category of fine" (e.g., "Insufficient legal basis for data processing") that is assigned by CMS professionals based on the violated GDPR articles. Table 1 provides the mapping criteria to GDPR articles, which CMS shared with us upon request. While one fine can only be assigned to one type, complex fines that result from multiple causes are assigned to the category that gave the greatest contribution to the GDPR non-compliance. These happen because e.g., Art. 5 and Art. 12 GDPR set out more general processing requirements and are often cited together with other articles. Art. 5 and 6 GDPR are often cited similarly, e.g., for processing without legal basis (*ETid-387*) or the illegal automatic forwarding of e-mails from former employees (*ETid-540*). Art. 12 and 13 GDPR are jointly assigned to fines that lack information obligations. For example, the installation of surveillance cameras without proper notice *(ETid-347)*.

The first fine listed in the ET was imposed on 07/17/2018, while the latest fine—included in our analysis—was assigned on 09/29/2021. In total, the fines sum up to around 1,839 million EUR (max: 746 million EUR; min: 0; mean: 2,149,000 EUR; SD: 27,045,034 EUR). The highest fine was imposed to *Amazon Europe Core S.à.r.l.* in July 2021 for non-compliance with general data processing principles by the data protection agency of Luxembourg (*National Commission for Data Protection* (CNPD); *ETid-778)*[4]. Taking a closer look at the countries with the highest number of fines, named in the ET, the Spanish DPA imposed the largest number of fines (296; 35%), followed by Italy (92; 11%), Romania (62; 7%) and Hungary (44; 5%).

Type of Fine	GDPR Articles
Insufficient legal basis for data processing	5, 6, 7, 8, 9
Insufficient technical and organizational measures to ensure information security	5, 32
Non-compliance with general data processing principles	5, 25, 30, 35
Insufficient fulfillment of data subjects rights	12, 15, 16, 17, 18, 19, 20, 21
Insufficient fulfillment of information obligations	12, 13, 14
Insufficient fulfillment of data breach notification obligations	33, 34
Lack of appointment of data protection officer	37
Insufficient cooperation with supervisory authority	31, 36, 58
Insufficient data processing agreement	28, 29

Table 1. Mapping types of fines to the corresponding GDPR article(s).

3.1.1 Comparing GDPR Fines Over Time

In the first year after the GDPR went into effect, the number ($n = 12$) and the sum of imposed fines (458,688 EUR) was comparatively low (June – December 2018). Barrett [3] explains this with a conservative approach of the DPAs. In 2018, authorities tended to define remedial measures along with the fines to encourage positive behavior in the handling of personal data. A year later, in 2019, 104 fines were imposed over 12 months, representing an increase of 867%. The sum of GDPR fines in this period was 21 million EUR, an increase of 4,579%[5]. The rise in the number and amount of fines continued in 2020 when 357 (an increase of 119% compared to 2019) fines were issued with a summed amount rising from 73 million EUR (2019) to 172 million EUR (2020), an increase of 136%. Comparing the numbers in the same period (January–September) from 2020 to 2021, 225 fines resulted in an amount of 59 million EUR fines in 2020. In 2021, the number of fines further rose to 307 (an increase of 36%), while the summed amount increased to 1,035 million EUR (an increase of 1,654%), compared to the previous year. The latter increase is heavily impacted by the large fines of 746 million EUR to *Amazon Europe Core S.à.r.l.* in July (*ETid-778*) and 225 million EUR to *WhatsApp Ireland Ltd.* in August *(ETid-820)*.

4 At the time of writing, Amazon is challenging the fine in court.

5 21 fines in 2019 are not dedicated to a concrete date and therefore are not included.

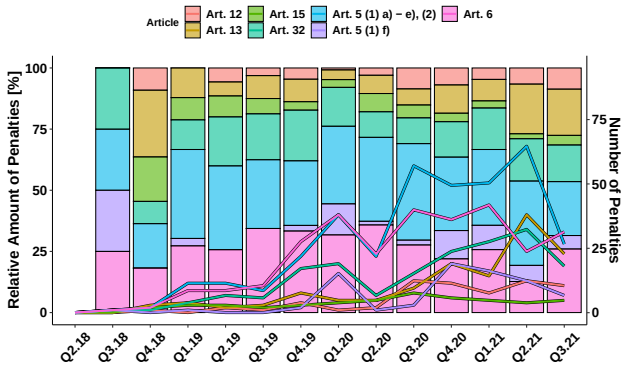

Fig. 1. Total Number and Relative Amount of Fines by GDPR Articles over time.

3.1.2 GDPR Articles Over Time

Figure 1 displays the development of the imposed fines from Q2/2018 until Q3/2021. We only list the top seven articles that received the most fines to increase readability. All values are non-accumulated, and a double calculation of fines is possible if the penalty was assigned to more than one article. As mentioned above, there were few fines in the first months of the GDPR, but their number increased steadily over time, besides a small decrease between Q2 and Q3 of 2020. However, the fines tend to increase over time over all three years. The decrease at the end of 2021 is an artifact of the dataset. While we included all fines listed as of the 1st of October 2021, fines are often added with a delay depending on when they become public.

3.2 Categorization of Fines

The categories provided by the ET are defined along with legal constraints and provide limited insights into data processes that violate the law. For example, the ET's category "Insufficient technical and organizational measures to ensure information security" describes a more general issue lacking information about the specific measure that was regarded insufficient by a DPA. To better understand the causes that led DPAs and further investigate the respective data processing, we categorize each fine on three additional levels. First, we identify the general cause (technical or organizational problem). Second, we label who initiated the investigation into the data processing. Third, we categorize the incident that led to the fine in one of seven high-level categories listed in Table 2 in and a number of sub-categories listed in the Appendix 5.s, used in the *ET*, isand does not allow to understand the underlying problems per se.what caused to be subject to investigation by DPAsdiededd

3.2.1 Technical or Organizational Origin

To understand what types of problems led to fines, we analyzed whether a fine was caused by *organizational* or *technical* issue. Starting with the latter, by *technical problem* we mean issues that occur during the design and implementation of a system, e.g., a developer that did not correctly anonymize a data set or a service administrator that did not implement a security measure correctly, which lead in consequence to data leakage. In contrast, *organizational problems* are more process-related issues, e.g., when an institution does not have or does not follow a policy on handling privacy-related tasks. This can lead to one or more violations of data privacy regulations, e.g., not reporting a known data breach to the DPA and the data subject in time.

For the classification of "technical" fines, we have oriented ourselves to the category "Insufficient Security Measures". This category and a citation of Art. 32 GDPR indicates a technical or organizational deficiencies. To identify the lack of technical measures, we examined the fine summaries for references to technical deficiencies. Accordingly, as long as a deficiency in technical terms was not ruled out, we indicated the origin as "technical". We sorted edge cases into both categories (technical and organizational) if we had no precise information about a fine's origin. If the summary of the fine gave any indication of an organizational failure or missing management of private data, we marked the origin of the fine as "organizational". In rare cases, we have not marked cases as either "technical" or "organizational" because the summary does not provide insight into the origin of the fine, or there is doubt as to whether it is a case of incorrect management or a legal conflict (such as with national law).

3.2.2 Causes of Investigations

In addition to the origin of a fine, we categorize the causes of the DPA's investigation leading to it. To create these categories, we analyzed the summary texts of each fine in the ET. Overall, we differentiate between four entities that cause an investigation: (1) the data controller, (2) a third party handling data on behalf of a partner, (3) data subjects, and (4) the respective DPA. For example, a data breach may be investigated

Category of Fine	Occur.
1. Insufficient Security Measures	386
2. Unauthorized Data Processing	751
3. Data breach information/ DPA cooperation	95
4. Data Subject Rights	182
5. General Obligations	91
6. Information Obligations	169
7. Violation of Basic Data Protection Principles	125

Table 2. Overview of the high level categories that we used in our analysis. A detailed review of all uses categories can be found in Appendix A.2.

by the supervisory authority because of a self-report by the *data controller*, by a *third party* (e.g., a business partner that had to report the institution to stay compliant with privacy regulations), an individual *data subject* that was affected by the breach or the respective DPA may detect an *unauthorized data publication* itself. This could mean that they observed information being disclosed to the broad public that can be traced back to an institution (e.g., as the data is visible on a domain owned by the data controller)

We base our evaluation solely on the textual summaries of the ET and not on assumptions that we could deduce from the text, even though they were very likely. If, for example, a video surveillance system is installed to monitor employees (*ETid-136*), an employee likely submitted a complaint to the DPA. However, since the text does not name the cause of the investigation by name, we did not consider these fines as the description lacks evidence. This is the main reason why we could only identify the cause of an investigation in roughly a third of all cases. Overall, we were able to identify the cause of the investigation for 295 out of the 856 cases (34%).

3.2.3 Detailed Classification of Fines

To categorize fines beyond the legal-oriented listing provided by the ET, we developed the following categorization schema by creating a codebook based on similarities of the violated GDPR articles in the fines (see Table 2):

1. *Insufficient Security Measures* to concretize the technical measures with regards to Art. 5 (f) and 32 GDPR.
2. *Unauthorized Data Processing* defines weaknesses that lead to unauthorized data processing as described in Art. 6 GDPR.
3. Categories in accordance with Art. 31, 33, and 34 GDPR are described in segment *Data breach information / cooperation with the DPA*.
4. *Data Subject Rights* describes the data subject rights according to Art. 15–22 GDPR.
5. *General Obligations* provides detailed requirements related to the appointment of a data protection officer (Art. 37 GDPR), the establishment of a data protection management system, the documentation of any processing of personal data in a register of processing activities (Art. 30 GDPR), the performance of data protection impact assessments (Art. 35 GDPR) or the implementation of "Privacy-by-Design and -Default" (Art. 25 GDPR).
6. *Information Obligations* concretizes the existing information obligations according to Art. 12–14 GDPR.
7. *Violation of Basic Data Protection Principles* for the processing of special categories of personal data as defined in Art. 9 GDPR.

3.3 Codebook Development and Annotation Process

One author developed the category schema by creating a codebook based on similarities of the violated GDPR articles in the fines. The categories that appear frequently are divided into further subcategories as listed in Appendix A.2, providing a more detailed view of the cause of the violation and thus the fine. For example, the category "Insufficient Security Measures" is divided into further subcategories such as "Missing role management" or "Unauthorized access". Hence, we can provide a more detailed analysis of issues that lead to the fines. The resulting codebook was discussed and enhanced with a second author to eliminate redundant or ambiguous categories.

Two authors then performed the categorization based on the defined codebook. The two researchers independently annotated the 100 latest fines in the dataset in the first step. Without revealing their results, they then discussed differences in the assigned categories to ensure that they had the same understanding of the mapping. Afterward, they updated their annotations of the entries, if necessary. For the resulting annotated datasets, we compare the inter-rater reliability of ratings using "Cohen's Kappa." Across all categories, the inter-rater reliability yielded an "almost perfect agreement" ($\kappa = 0.91$). The remaining entries in the database were then split equally among the two researchers and, again, independently annotated. Table 5 in Appendix A.2 pro-

vides a detailed list of all categories and subcategories that we used in our experiment.

4 Understanding Causes of GDPR Fines

The following section discusses the results of the three additional classifications of GDPR fines we conducted.

4.1 Types of Issues and Cause of investigation

We first differentiate the issues that lead to a fine between organizational and technical. We also classify who or what caused the investigation of an issue that led to a fine.

4.1.1 Quantifying Technical and Organizational Fines

The data shows that only 65 (8%) of all fines can be attributed to technical issues, whereas 405 (47%) result from organizational issues. 369 (43%) of fines are rooted in organizational as well as technical failures, as one fine can be based on multiple noncompliances. For 15 fines (2%), no information is given to indicate the origin of it (e.g., The controller did not take adequate security measures when processing personal data, thereby breaching the obligation to protect the processed personal data; *ETid-95*). This shows that GDPR fines are rarely based on purely technical problems but that most of the time, organizational problems are at least a part of the issue.

4.1.2 Understanding Causes of Investigations

For 247 (29%) of all fines, it was possible to identify causes of investigations that lead to a fine (see Table 3). Overall, data breaches that are either reported by the data processors or investigated by supervisory authorities when the data become public are the most common causes for investigations. Complaints by data subjects are also common, whereas these are mostly customers but in some instances also initiated by (former) employees.

Of the 247 fines, 29 (12%) are initiated by a controller's data breach notification, e.g., when reporting a security breach after six unencrypted USB sticks containing personal data were lost (e.g., *ETid-113*). Furthermore, 45 fines (18%) are caused by complaints of other companies. We further find 61 fines (25%) caused by data subject complaints; 9 of them (15%) reported by (former) employees and 52 (85%) by customers. Lastly, as the DPA itself can also initiate an investigation (e.g., when personal data is unlawfully published on a website or social media), we find such data publication in 111 cases (45%).

Cause of Investigation	Total	Technic.	Organiz.
data breach notification by the controller	29	23	24
(company) complaints	45	24	41
data subject complaint (former employee)	9	8	9
data subject complaint (customer)	52	25	49
data publication	111	73	94

Table 3. Causes of investigations that lead to a fine.

4.2 Classification of the Reported Fines

The following categories of fines are based on the annotations described in Table 2 and detailed in Table 5.

4.2.1 Insufficient Security Measures

In our data set, 186 fines (22%) are related to insufficient measures to combat attacks or misuse effectively (e.g., an attack on a company's website or server; *ETid-71*). Only five fines (1%) result from a digital loss of data, and 16 fines (2%) are based on the physical loss of data, meaning a loss of availability due to lack of backups. While this emphasizes the risks of data breaches, the result may also be skewed as data breaches have to be reported and are more likely to result in a fine. However, DPAs may refrain from issuing a fine if the data processor follows best practices in protecting the data. A high number of fines (107; 12%) is related to a lack of enforcement of access control. Thereof, missing role management was the cause for 52 fines (6%) and resulted in either a customer accessing personal data of other customers (e.g., *ETid-132*, *ETid-210*) or employees having unauthorized access to personal data of customers (e.g., *ETid-1250*, *ETid-1540*). Another 55 fines (6%) are related to cases of unauthorized access, including those where a user account was showing personal data of another user (e.g., *ETid-464*). Furthermore, 56 fines (6%) are attributed to an insufficient data life cycle management, which includes a missing process for

handling data at its end of life (e.g., *ETid-98*, *ETid-130*, *ETid-391*), meaning it is never deleted.

4.2.2 Unauthorized Data Processing

As displayed in Table 5, the processing of personal data without a legal basis was the most frequent reason for a fine (409 cases; 48%). The list of fines resulting from unauthorized disclosure (225 fines; 26%) includes cases where personal data was accidentally published due to inadequate internal control mechanisms *(ETid-101)*, sending data or login credentials to the wrong person without verifying the identity of the receiver (*ETid-119*, *ETid-134*) or publishing personal data or photos without consent *(ETid-614, ETid-616)*. We also find 37 cases (4%) in which a fine was imposed with relation to consent, mentioning that it was not "specific", "unambiguous" (e.g., *ETid-23*, *ETid-426*) or obtained "voluntarily" (e.g., *ETid-224*, *ETid-516*). In other cases, the obtained consent does not meet the requirements for withdrawal (e.g., *ETid-47*), ignores "opt-outs" (e.g., *ETid-82*) or the controller is simply not able to prove the existence of an individual's choice (e.g., *ETid-177*, *ETid-214*).

4.2.3 Data Breach Notification & Cooperation with Authorities

In contrast to the processing of data, fewer fines are issued regarding communication around data breaches. We found 49 cases (6%), in which no sufficient cooperation with the authorities took place, e.g., by ignoring warnings (*ETid-39*, *ETid-42*) disregarding concrete orders (*ETid-148*, *ETid-175*), or by not responding the authority at all (*ETid-94*, *ETid-583*).

29 fines (3%) are imposed for not reporting a data breach to the supervisory authority on time or at all. Additionally, 17 fines (2%) are caused by not reporting a data breach to the data subject promptly or at all.

4.2.4 Data Subject Rights

Data subject rights have been the topic of several studies, which have shown that many institutions do not (correctly) comply with requests by data subjects [5, 6, 31]. The analysis of fines underscores that such issues are widespread. In our analysis, the rights of data subjects are the basis for 182 fines (21%). From these fines, 62 times (34%) the "right of access" was not sufficiently granted, e.g., by failure to respond to the data subject at all or in a timely manner (*ETid-315*, *ETid-499*). The "right to erasure" further is the cause for a fine in 56 cases (31%) e.g., customers not being able to have their personal data deleted (*ETid-316*, *ETid-480*). One notable observation is that the Spanish DPA (*AEPD*) fined telephone companies that, instead of deleting customers' data on request, continued their processing of data extensively (*ETid-93*, *ETid-240*). Lastly, 42 cases (23%) of an insufficient "right to object" are caused by controllers ignoring the data subjects withdrawing of consent (*ETid-111*, *ETid-196*) or in particular ignoring data subjects to unsubscribe from newsletters or marketing calls (*ETid-264*, *ETid-336*).

4.2.5 General Obligations of Organizations

By general obligations, we mean organizational structures are required by the GDPR (e.g., appointing a data protection officer or implementing a management system for data protection). This is the smallest category in our data set (in terms of the number of fines and the aggregated amount of fines). The majority of cases are related to privacy-unfriendly design (57 cases; 7%) that is caused by e.g., a lack of implementation of GDPR related security measures (e.g., *ETid-39*) or using collected personal data instead of dummy data for testing purposes e.g., (*ETid-494*). In 16 cases (2%), a missing data protection impact assessment was the cause of a fine. In contrast, only eight fines (1%) are filed due to failure of appointing a data protection officer or a missing management system for data protection.

4.2.6 Information Obligations & Violation of Basic Data Protection Principles

Regarding the GDPR's obligations for providing information to data subjects, 106 fines (12%) are caused by insufficient transparency about the processing, e.g., through a privacy policy. Our analysis reveals such notices as either incomplete (*ETid-539*), inconsistent (e.g., *ETid-522*), or are missing completely (*ETid-588*). Analyzing cookie policies on websites, also an important research topic, ten fines (1%) are assigned for not providing users the option to refuse their cookies (*ETid-86*), for insufficient information about the purpose, properties, and activation time (*ETid-364*), or for cookie banners missing completely (*ETid-220*). This is in line with previous works that identified that cookie

banners do not work as intended [23], are misleading, or unusable in general [9, 32].

4.2.7 Special Categories of Personal Data

Furthermore, the analysis shows 125 cases (15%) in which sensitive personal data according to Art. 9 GDPR is being processed. Such sensitive personal data processings reach from processing biometric data (e.g., fingerprints) for granting access to certain rooms (*ETid-185*), recording attendance (*ETid-274*) or health-related data being accessed without unauthorization or by a too wide range of people (*ETid-539*, *ETid-555*).

5 GDPR Fines in the Data Life Cycle

The categorization that we applied helps us better understand issues that lead to fines, but it does not directly allow us to assess at which step in the data processing an issue occurs. We mapped our categories to different phases of a data life cycle model to answer this question. This analysis helps institutions and data protection professionals better understand in which development steps additional care is needed to comply with legislation and protect personal data.

5.1 Abstract Personal Data Life Cycle Model

As a basis for this analysis, we use the *abstract personal data life cycle model (APDL)* proposed by Alshammari and Simpson [2]. The APDL describes a variety of processing phases that personal data pass during their lifetime (e.g., data collection, usage, and destruction). These phases can be matched with our categorization scheme. The APDL consists of eight phases: (1) *Initiation*, (2) *Collection*, (3) *Retention*, (4) *Access*, (5) *Disclosure*, (6) *Usage*, (7) *Review*, and (8) *Destruction*. However, personal data might not go through all phases of the APDL. For example, following the principle of data minimization, collected data might be deleted right after usage and, thus, never be accessed or reviewed afterward. Figure 2 provides an overview of all phases and the transitions between them.

5.2 Mapping Fines to the ADPL

We assign each fine category (see 5) to a life cycle phase (e.g., data collection or data retention) in which the data protection noncompliance occurred. Furthermore, we quantify the number of issues in each stage.

One of the authors used the descriptions of Alshammari and Simpson [2] to map categories to an ADPL phase. The following categories are not matched to the APDL since they do not fit in any of the phases: “Not reporting a data breach to the DPA”, “Not reporting a data breach to the data subject”, “Insufficient cooperation with DPA”, and “Injust processing of article 9 data.” Except for the privacy policy category that we allocate to both the *initiation* and the *collection* phase, the mapping of categories to phases is evident in all cases. The “privacy policy” categories were mapped to two phases because a controller is required to develop a processing policy in advance (*initiation* phase) and disclose the policy at the time of collection (*collection* phase).

5.2.1 Mapping Fines to Life Cycle Phases

The *initiation* phase describes the development of a processing plan, which includes the definition of the purpose and how personal data is collected and processed. This is the phase where Privacy-by-Design and -Default are applied, which is why we attributed the categories privacy-unfriendly design and the privacy policy to this phase.

The initiation phase is then followed by the *collection* phase, in which data is recorded, collected, or obtained by the data subject itself or by another institution (e.g., in the context of data processing under commission). Here, a sufficient legal basis is required, e.g., by obtaining the explicit consent of the data subject or by negotiating a data processing agreement between organizations. The collection of personal data from the data subject also includes the communication of transparency obligations by e.g., providing information regarding camera surveillance or privacy and cookie policies and data minimization.

The collection phase precedes *retention*, in which personal data is organized, structured, and stored for a specific time. During this phase, special measures must be taken to prevent digital and physical theft, loss of, and unauthorized access to the data. While retaining data, it may be necessary to re-evaluate parts of the collection phase (e.g., when processing for a different purpose is planned or the legal basis changes).

The retention is followed by the *access* phase, in which personal data is specified and retrieved. In this stage, effective access control is essential, as it ensures that the correct information is accessed by the right person at the right time. One of two phases that follows the access phase is the *disclosure* phase, which describes how data is made available for internal or external use. Here it is crucial to ensure that the data is only disclosed to authorized parties. The right to data portability, whereby data is transferred from controller to controller or to the data subject, also falls under the disclosure phase.

The second phase that follows the access of data is the *review* phase, in which the data subject rights to access, rectification, restriction, and object are ensured. During the usage phase, the data must be protected by activities against manipulation, attacks, and use contrary to the specified purpose.

The last life cycle phase is the *destruction* phase referring to the deletion of data based on the defined retention periods and deletion policies, which includes the right to erasure.

5.2.2 Quantifying Technical and Organizational Issues

To assess how the type of a fine is related to a phase, we sum up the number of fines for each life cycle state, based on the previous categorization described in Section 3.2, distinguishing between *technical* and *organizational* fines. We chose to analyze the number of fines instead of the total sum to understand when authorities tend to impose fines in the data life cycle. Analyzing the amount of the fines creates a bias towards higher fines (e.g., to *Amazon*).

5.3 Results

Figure 2 shows the APDL complemented by the type and number of fines. Each hexagonal box describes one phase of the model, the number-letter combinations in each box refer to the categories of GDPR fines from Table 5, which we have assigned to the respective phase. The numbers on the left represent the organizational fines, while the numbers on the right display technical fines. Besides the total number of fines per phase (organizational 683 fines; technical 434 fines), we list the relative amount of organizational and technical fines that were imposed in each phase of the APDL, in relation to the total number of fines.

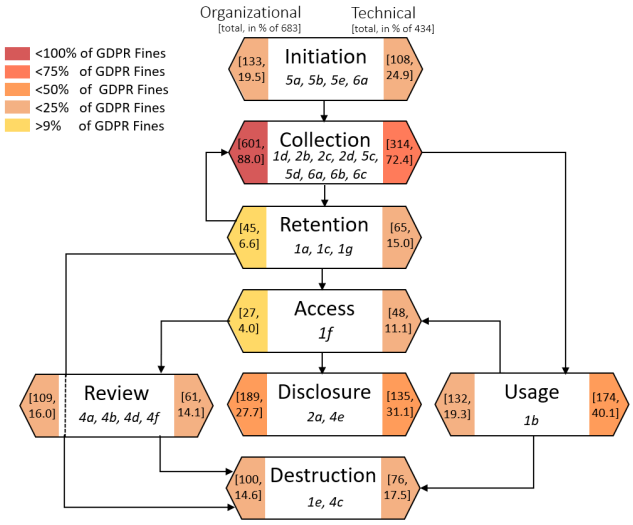

Fig. 2. Mapping of GDPR fines to the personal data life cycle model of Alshammari and Simpson [2] according to our criteria. Since fines can be in multiple categories, the summed up numbers exceed 100%.

A high-level view shows that technical and organizational problems are similarly distributed along all phases. For both, technical and organizational fines, the *collection* phase holds the largest number of fines (organizational: 601 fines, 88%; technical: 314 fines, 72%). While this is a result of the number of categories (9 fines; 36%) belonging to this phase, this is based on the fact that most of the requirements must be addressed during the collection process of personal data.

Some phases have a similar amount of fines in both categories differing less than 4 %. These are the *disclosure* phase (organizational: 189 fines, 28%; technical: 135 fines, 31%), the *review* phase (organizational: 109 fines, 16%; technical: 61 fines, 14%) and the *destruction* phase (organizational: 100 fines, 15%; technical: 76 fines, 18%). While the *disclosure* phase mostly consists of fines for "unauthorized disclosure of the data," the *review* phase contains problems that institutions face when handling data subject rights. Furthermore, the *destruction* phase addresses problems of an "insufficient data life cycle managament" and the "insufficient right to erasure".

The *usage*, *access*, and *retention* phases are the only phases that hold significantly more technical than organizational fines, roughly twice to three times as many. While the *usage* phase only consists of fines attributed to the category "insufficient measures to effectively combat attacks/misuse", the *access* phase holds fines related to "missing role management". Further, the *retention* phase consists of fines connected to "digital loss of data", "physical loss of data" and "unauthorized access".

5.3.1 Life Cycle Phases with Predominantly Organizational Fines

In the *collection* phase, the amount of organizational fines is twice as high as technical fines. This imbalance is present in all categories assigned to this life cycle phase (see details in Table 5). The largest category within the *collection* phase is "insufficient legal basis" with 356 organizational and 135 technical fines. Such violations include institutions not sufficiently defining a legal basis for the processing of personal data, either when collecting data from the data subject itself (*(ETiD-55, ETiD-92*, *ETiD-276*), or when acting as data processor on their behalf (*ETiD-97*, *ETiD-423*, *ETiD-617*). We further found cases in which a valid legal basis existed. However, the processing was unlawfully extended in a manner that the legal basis no longer covered the processing, e.g., by publishing personal data (*ETiD-44*). In addition, other fines are related to processors ignoring data subject requests to restrict the data processing, e.g., by using a "Robinson list" (*ETiD-155*, *ETiD-660*), which is an opt-out mechanism not to receive marketing material. Some violations of the legal basis are directly related to technical issues, such as the failure to obtain consent when using cookies (*ETiD-135*).

The *collection* phase is the only phase where we found more fines being rooted in organizational (601 fines; 88%) than technical issues (314 fines; 72%). All fines in these categories are related to data processing that was carried out without previously ensuring that the data could be legally used. While this result may be an effect of a bias in the data set because data subjects more often notice these issues, it still underlines the importance of data protection assessments before processing personal data.

5.3.2 Life Cycle Phases with Predominantly Technical Fines

In the *retention*, *access* and *usage* phases, we identified a higher number of technical issues compared to organizational violations. In the *retention* period, this is largely influenced by the category "unauthorized access" (organizational: 31 fines; technical: 51 fines). In the *access* phase, the high amount of technical fines is solely related to "missing role management" (organizational: 27 fines; technical: 48 fines). Finally, the high proportion of technical fines in the *usage* phase is related to "insufficient measures to effectively combat attacks / misuse" (organizational: 132 fines; technical: 174 fines). In all of these phases, institutions are fined that fail to implement technical mechanisms that restrict access and usage of personal data. For example, a school implemented a tool that allowed teachers and parents to communicate via an app. However, it had not been appropriately hardened, and consequently, a data breach occurred and was reported to the authorities (*ETid-292*). In another example, an adversary could tamper with an insufficiently secured chatbot hosted by a third party. This resulted in unauthorized access to customers' financial information *(ETid-440)*.

This highlights the importance of technical measures to control intentional and prevent unintentional access to personal data. While technical solutions exist, they are often not properly implemented.

5.3.3 Life Cycle Phases with Balanced Fine Types

The *review*, *disclosure* and *destruction* phases contain a similar (relative) amount of organizational and technical fines. Although the fulfillment of data subject rights at first seems like an organizational problem, the numbers in the *review* phase show that they include technical problems as well. This is in line with previous works, which show that organizations lack sufficient processes to verify the identity of the data subject, which, for example, allows unauthorized access to data [4, 6, 10]. Other work has revealed that organizations have issues identifying the data related to a subject and, therefore, fail to implement the rights on a technical level [5, 30, 31].

In the *disclosure* phase, personal data is made accessible to unauthorized persons. This happens due to technical or organizational errors. technical errors include, e.g., insufficient password management (*ETid-64*), storing personal data on a public storage server, or making data available over the internet (*ETid-715*, *ETid-716*, *ETid-719*). Example for organizational errors, in turn, are employees sending emails, SMS, or documents containing personal data to the wrong receiver (*ETid-110*, *ETid-171*, *ETid-243*) or employees sharing video recordings or photos in social media (*ETid-566*, *ETid-616*)). Furthermore, there are several issues related to both technical and organizational problems. For example, a data controller did not fulfill its data breach notification obligations when a flash memory with personal data was lost *(ETid-74)*, or patients were able to access not only their medical reports but also the personal health data of other patients due to a human error in the IT systems integration *(ETid-433)*.

Finally, the *destruction* phase holds fines connected to the categories "missing data life cycle management", as well as an "insufficient right to erasure". Both categories assigned to this phase have a slightly higher total share of organizational fines. Similar to issues implementing other data subject rights, many fines in the *review* phase are connected to the right to erasure. Given that research has shown that many organizations fail to identify data subjects [5, 31], they will have trouble deleting data related to them.

5.3.4 Organizational Fines Caused by Technical Failure

The *initiation* phase has many "general obligations" categories assigned to it, which appear primarily organizational (e.g., appointing a data protection officer or missing a management system for data protection). However, concerning the types of issues, we see that the relative share of technical fines (25%) is higher than the share of organizational fines (20%). We explain this by a high number of fines in the category "privacy-unfriendly design", addressing the general obligations for "Privacy-by-Design" and "Privacy-by-Default," which contain more technical than organizational fines. This category includes, for example, fines for the wrong implementation of "opt-in vs. opt-out" *(ETiD-182)* or controlling associates' work performance using their smartphones *(ETiD-790)*. In addition, the *initiation* phase holds a high number of fines in the category "missing / incorrect privacy policy", which is also an organizational issue at first sight. However, we find that while an adequate privacy policy is an organizational task in terms of content, the correct implementation in the IT systems and code is a technical issue (e.g., website, app, or server) is a technical one.

Concluding, we explain the high number of GDPR fines in the *initiation* phase by many organizational issues whose implementation is of a technical nature and therefore has technical dependencies.

6 Technical Issues that Lead to Fines

The data life cycle analysis highlights the data processing phases in which issues occur that lead to GDPR fines. However, such an analysis does not answer questions regarding the violated principles in the GDPR. In this section, we analyze the identified *technical* fines in the ET to understand their relation to the definitions (Article 4 GDPR) made in the general provisions of the GDPR. We achieve this by analyzing the word frequency of each fine's summary to map the mentioned issues to GDPR definitions. This allows us to connect the technical problems of the fined institutions with the legal perspective defined in the regulation.

6.1 Methodical Approach

This section limits the analysis to technical fines only to provide clear and applicable hints for technical data protection professionals, e.g., developers or architects. This approach helps understanding the issues that lead to a fine and reflecting the handling of personal data accordingly.

6.1.1 Word Frequency Analysis

To understand the technical problems that lead to a fine, we take a closer look at the fine summaries of the ET. The summaries contain detailed information on how and why a fine was imposed. Thus, analyzing the words mentioned in all summaries estimates technical problems leading to a fine. This analysis aims to shed light on common areas in which fined problems occur. Such a statistical approach comes with the downside that we probably include counting artifacts. However, since we only focus on the most frequent words for the in-depth analysis, we argue that this effect is minimal.

We first combined all ET summary texts of technical fines into a single text. Afterward, we remove filler words (e.g., they, have, how) from the text using a *Natural Language Toolkit (NLTK)*. The NLTK comes with a pre-defined list of words to be eliminated since they do not describe any content. We extend this list with data protection related terms (e.g., "data", "personal", "dpa", "GDPR", or "fine") so that frequent but unspecific words are eliminated (stopwords). The complete list of our introduced stopwords can be found in Appendix A.1. Next, we clean the text by removing all punctuation, transforming all remaining text into lower case, and finally, by lemmatizing and stemming the output. The goal of lemmatization and stemming is to reduce a word's various forms to a common base form. Lemmatization means to group inflected forms of a word into a single word (e.g., the lemmatization for "studies"

and "studying" would give "study"). Stemming means the trimming of the word's ends so that derivational affixes are removed and more realistic word frequencies are obtained. In our analysis, we list a word's absolute number of occurrences, meaning if a fine's summary text contains a word multiple times, it is counted multiple times. Thus, the list serves as an indicator of fined issues that e.g., developers face.

6.1.2 Mapping Fine Reasons to GDPR Definitions

The absolute frequency with which a word occurs in the analyzed technical fines provides insights into penalized data protection issues. However, it is unclear how and which legislative issues are encountered. Towards understanding the relation between a fine and the principles depicted in the GDPR, we map them to the general definition of terms made in the legislation. Within the general provisions, Article 4 GDPR (1) – (12) provides a rich set of definitions that apply uniformly throughout the regulation and concern the processing of personal data. For example, Art. 4 GDPR (1) defines the terms "personal data" and "data subject". To establish the connection between the words and the terms, we first cluster the terms to combine similar definitions into one category. For example, we cluster the terms "controller", "processor", and "recipient" into one group as they all process personal data. From the twelve considered articles, we derived seven categories: (1) "Processing", (2) "Restriction", (3) "Filing System", (4) "Data Subject", (5) "Personal Data", (6) "Data Breach", and (7) "Controller/Processor/Recipient". We dropped two terms ("profiling" and "pseudonymization") as they are a specific processing type and do not generally affect the process.

We use defining categories to map the words from our word frequency analysis. To assign the words to the corresponding category, three of the authors independently match all words that occurred at least 20 times in our analysis. In total, the authors mapped 195 words in this process, but for further analysis, we only kept the words all authors assigned to the same category. This applies to 157 (81%) of the analyzed words. We further dropped words that could not be assigned to any cluster (e.g., "region"). Table 4 shows the most frequent words of technical GDPR fines. If a category contains a large number of words, we limit the maximum number of words per category to eight.

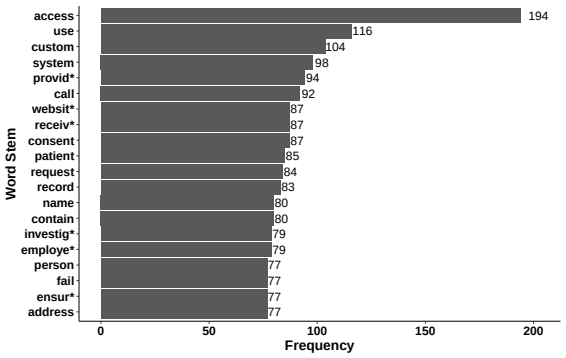

Fig. 3. Word frequency of word stems (e.g., "ensur" for ensure) of technical fines, based on summary texts of the ET.

6.2 Results

In the following section, we provide an overview of the results from the word frequency analysis and our mapping of them to GDPR definitions.

6.2.1 Common Words in GDPR Fines

As described, we performed a word frequency analysis of the summaries of technical fines. In our corpus, the summary of a technical fine has 35 words on average after pre-processing. Figure 3 shows the resulting list of the most frequent keywords in the analyzed GDPR fines. In our data set, the word stem "access" is mentioned 194 times in 135 distinct technical fines. Given that our data set only contains violations that went public, this is not surprising since an unauthorized disclosure will (probably) lead to a fine and is always related to improper data access management. Thus, an over-representation of such cases is likely. However, it is still interesting to have a closer look at the causes of the fines to better understand what created the issues (e.g., stolen databases or lost USB devices).

Case Study: Access Control Issues

To gain deeper insights into the causes of "access" related fines ($n = 135$), we performed an in-depth analysis of all related summary texts. One author clustered these fines according to their type of violation. This process resulted in two general cluster types: (1) institutions collected or disclosed too much information (e.g., when sharing data with partners) and (2) institutions provided too little information (e.g., upon data subject request). Within each of these clusters, different categories of violations exist, see Figure 4.

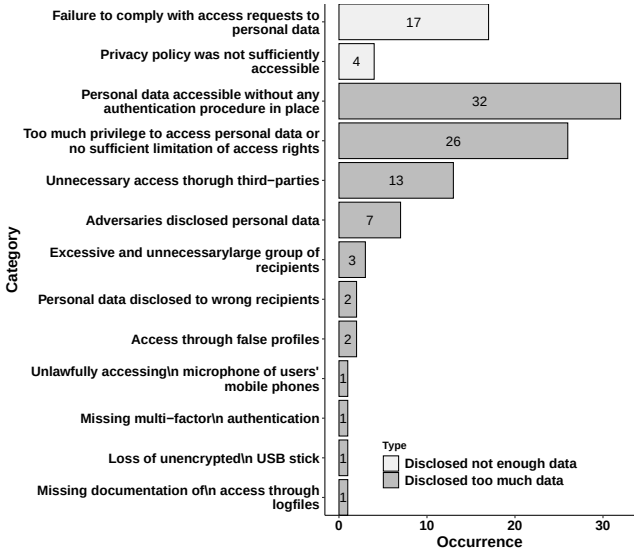

Fig. 4. In-depth analysis of access control issues of technical GDPR fines, due to the frequency of the term "access" in the word frequency analysis

The main reason (32 or 24% of cases) for fines related to institutions collecting or disclosing too much data was that data was accessible without any authentication. Furthermore, in 26 (19%) cases, fines were assigned due to insufficient limitation of access rights of persons processing data. In such cases, persons received access to a specific set of information, but they could and did access data they were not allowed to see ("privilege escalation"). In addition, in 13 (10%) cases, third parties unjustifiably received access to personal data. Only in seven (5%) cases do adversaries get unauthorized access to or disclosed personal data. A previous study of Tahaei et al. [28] found, in a developer-centered analysis, that developers are aware of vulnerabilities in access management systems and with regards to privacy, which is consistent with our data that shows when these vulnerabilities persist and are misused.

In contrast, in cases where institutions provide too little information, they were mostly (17 cases; 13% cases) unable to comply with access requests of data subjects. These findings are in line with the observations made by previous work [5, 31] Four (3%) cases of fines mentioning "access" were related to privacy policies that were not sufficiently accessible.

6.2.2 Understanding Relation to GDPR Principles

Based on the general principles for processing personal data as listed in Art. 5 GDPR, we want to understand the general topic of issues that lead to a fine based on our word frequency analysis.

Word Stem	Count	Fines	Word Stem	Count	Fines
Personal data (Art. 4 (1) GDPR)					
name	80	71 (16%)	address	77	59 (14%)
health	68	48 (11%)	medic(al)	30	29 (7%)
credit	27	20 (7%)	code	20	16 (4%)
Data subject (Art. 4 (1) GDPR)					
custom(er)	104	67 (15%)	patient	85	35 (8%)
employe(e)	79	53 (12%)	person	77	25 (6%)
user	59	41 (10%)	individu(al)	54	48 (11%)
peopl(e)	34	29 (7%)			
Processing (Art. 4 (2) GDPR)					
access	194	135 (31%)	use	116	86 (20%)
provid(e)	94	74 (17%)	call	92	39 (9%)
receiv(e)	87	71 (16%)	request	84	59 (14%)
record	83	43 (10%)	contain	80	70 (16%)
Restriction (Art. 4 (3) GDPR)					
consent	87	54 (12%)	adequ(ate)	74	65 (15%)
time	60	58 (13%)	suffici(ent)	49	44(10%)
appropri(ate)	46	40 (9%)	oblig(ation)	44	36 (8%)
period	34	29 (7%)			
Filing System (Art. 4 (6) GDPR)					
system	98	67 (15%)	websit(e)	87	64 (15%)
document	71	49 (11%)	telephone	52	42 (10%)
camera	40	23 (5%)	email	40	25 (6%)
card	39	28 (7%)	order	35	31 (7%)
Controller, Processor, Recipient (Art. 4 (7), (8), (9) GDPR)					
bank	59	35 (8%)	oper(ator)	59	19 (4%)
vodafone	41	22 (5%)	hospital	36	23 (5%)
employ(er)	27	9 (2%)	school	24	13 (3%)
Data breach (Art. 4 (12) GDPR)					
complaint	53	47 (11%)	incid(dent)	37	32 (7%)
attack	37	9 (2%)	error	31	30 (7%)
notif(y)	23	13 (3%)	failur(e)	22	22 (5%)

Table 4. Mapping of the words that appeared frequently in the fines to the relevant GDPR principles.

Table 4 shows the mapping of the words mentioned in technical fines to the GDPR principle. Words with the highest occurrence relate to *processing* activities. This includes "access" (194), analyzed above, but also "use" (116), "provide" (94) and "call" (92).

The analysis also shows what types of *personal data* are often referenced in descriptions of technical fines. The list includes categories of personal data like "name" (80) and "address" (77) which are likely frequently collected by any organization. We further find a high frequency of health related personal data (e.g., "health" (data) (68), and "medical" (records) (30)). Similarly, "patient" (85) is a *data subject* often mentioned, besides "customer" and "employee"

Specific *filing systems* often named in connection with technical GDPR fines are "websites" (87) and "telephones" (52). We find that websites often miss privacy statements *(ETid-142)* or are used to disclose personal data to a broad (unauthorized) audience *(ETid-41)*.

Telephones often refer to unauthorized data processing in cases of data subjects did not consent *(ETid-661)*, objected the processing activity *(ETid-676)*, or even requested to delete their personal (telephone) data *(ETid-33)*.

When looking at *controller, processor and recipients* as actors involved in technical GDPR fines, we can identify "banks" (59), "hospitals" (22) and "vodafone" as the only company identified specifically (41). This confirms our previous observations regarding health-related data and telephone data. In addition, our analysis reveals banks and, therefore, payment data as additionally often fined by the DPAs. Such GDPR violations include publishing bank-related data on the internet *(ETid-40)* or sending such data to the wrong recipient *(ETid-326, ETid-350)*.

Case Study: Health Data Issues

Based on the observation that "health" (data) and "medical" (records), together with data subjects "patient", and "hospitals" as organizations are frequently mentioned in descriptions of fines related to technical issues, we further analyzed this domain. It highlights that authorities emphasize investigating potential violations that involve sensitive data (i.e., health records in this case) as laid out in Art. 9 GDPR.

In our data set, 115 GDPR fines (13% of all fines) involve health data. While the cause of the DPA investigations that lead to fines is not always clear, a report of the *Future of Privacy Forum* [29] lists focus areas of the national DPAs and reports that in France, Italy, Sweden, and Belgium health has a high investigation priority. If we look at health-related fines of these countries (46 GDPR fines), we see that compared to the total number of fines in these countries (154 GDPR fines), the health-related focus is disproportionately represented (30%).

For data protection professionals in this area, this implies an increased risk of DPA investigations. It can be explained, to some extent, with the public attention (e.g., via media coverage) of GDPR violations that involve sensitive data. In addition, this type of data is associated with higher risks and negative consequences for individuals so that they might be more willing to take action and send complaints in case of incidents.

7 Limitations

We profit from the rich *CMS Law GDPR Enforcement Tracker* dataset in this work. However, using this and other similar datasets is limited because only those GDPR violations that were made public and ultimately fined are analyzed. This may create a bias in the dataset that is hard to assess. First, small fines and fines with less public attention may remain unnoticed. Second, fined issues are not necessarily distributed similarly to data protection issues that exist but go unnoticed. Still, we can show that the results overlap with other work (e.g., on problems with access control). Furthermore, we rely on the summaries of cases provided by *CMS* employees that create all the entries manually. Misinterpretations may occur (e.g., assigning the wrong GDPR articles). However, our comparison with other data (see Section 2.2) did not reveal any systematic errors. Finally, some ET summary texts do not provide an appropriate reason (e.g., *ETid-40*, *ETid-134*, *ETid-226*) for the fine. Hence, we had to exclude such cases from some parts of our analysis. However, we are confident that the dataset still serves to study the origins of GDPR fines.

Manual evaluation methods, as used in our paper for categorization of the ET, leave room for misinterpretations. Although we tried to minimize them by ensuring cross-checking through several authors, they cannot be avoided entirely. We have made all our manual evaluation documents transparent to provide as much transparency as possible regarding our manual evaluation methods.

8 Conclusion

This work presents the analysis of 856 fines issued by DPAs for violations of the GDPR. The dataset originates from a list published by *CMS* and has been updated regularly since May 2018. It provides a basic categorization of the fines based on the articles violated by the respective institution. We enhance this category scheme by adding several subcategories that include details about the root cause, whether it was a mainly technical or organizational issue, and why the DPA initiated the investigation. The fines' primary drivers are customer complaints and data becoming public (e.g., through breaches).

We further provide a deeper analysis of GDPR-related fines than in previous work. While especially security research has focused on the technical aspects

of data protection and focused on, e.g. developers, we show that the DPAs issue most fines for reasons of an organizational origin. Our results show that many fines are related to problems that are well known and studied by the PETs community (e.g., measures to combat attacks, prevention of unauthorized disclosures, or privacy-unfriendly design). Given that the fines are all based on issues within the past four years, it raises the question of where and why the state of the art has not reached practice yet. Our analysis can provide information to researchers as well as practitioners. For example, researchers in privacy-enhancing technologies can use our results as a starting point for future studies. The categorization of fines (see Table 5) reveals areas where research could help better understand what causes the problems or how they can be addressed. For example, 409 fines are related to processing data without a legal basis. Legal and security researchers need to work together to, for example, ensure that the legal basis is known and data processing is compliant with it. Another area for research that has not seen widespread attention is data breach detection and notification (95 fines) or what motivates customers to complain about a company's data practices (52 fines).

Our results support the need for rigorous data protection management and privacy-by-design, which are necessary for practitioners and institutions processing personal data to avoid fines. In all phases of the data life cycle, violations have been identified. A high number of fines is attributed to the early phases of data processing, which indicates widespread data processing issues. In addition, the literature suggests conducting a data protection analysis at the beginning to ensure that there is a legal basis to start the processing. The fact that many violations are also related to issues in access control management emphasizes that organizational processes need to be established (and continuously monitored) to avoid unwanted data disclosures.

Our analysis also shows that customer data protection is becoming an essential issue as customer complaints are the second most frequent causes of investigations by DPAs. In addition, the high frequency of health-related terms indicates the focus of DPAs on sensitive data processing.

Availability of Data & Code Artifacts

To foster future research, we release our code, measurement data, and other supplementary information at: https://github.com/RUB-SysSec/GDPR-fines.

Acknowledgments

The authors would like to thank their shepherd, Dominik Herrmann, and the anonymous reviewers for their helpful comments. Furthermore, we want to thank Frank Pallas for his support on this work. This work was partially supported by the German Federal Ministry for Economic Affairs and Energy (grant 01MN21002H "IDunion") and the research training group "Human Centered Systems Security" sponsored by the state of North Rhine-Westphalia.

References

[1] Gonçalo Almeida Teixeira, Miguel Mira da Silva, and Ruben Pereira. The Critical Success Factors of GDPR Implementation: a Systematic Literature Review. *Digital Policy, Regulation and Governance*, 21(4), 2019.

[2] Majed Alshammari and Andrew Simpson. Personal Data Management: An Abstract Personal Data Lifecycle Model. In *Business Process Management Workshops*, BPM, 2018.

[3] Catherine Barrett. Emerging Trends from the First Year of EU GDPR Enforcement. *Scitech Lawyer*, 16(3), 2020.

[4] Marlene Barth. A Case Study on Data Portability. *Datenschutz und Datensicherheit*, 45(3), 2021.

[5] Coline Boniface, Imane Fouad, Nataliia Bielova, Cédric Lauradoux, and Cristiana Santos. Security Analysis of Subject Access Request Procedures. In *Annual Privacy Forum*, APF, 2019.

[6] Matteo Cagnazzo, Thorsten Holz, and Norbert Pohlmann. GDPiRated—Stealing Personal Information On- and Offine. In *European Symposium on Research in Computer Security*, ESORICS, 2019.

[7] Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, and Isabel Yu. Achieving Privacy: Costs of Compliance and Enforcement of Data Protection Regulation. https://openknowledge.worldbank.org/handle/10986/35306, 2021.

[8] CMS. GDPR Enforcement Tracker Report – Executive Summary. https://cms.law/en/media/local/cms-hs/files/publications/publications/gdpr-enforcement-tracker-report-2021-executive-summary?v=1, 2021.

[9] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your

Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Symposium on Network and Distributed System Security*, NDSS, 2019.

[10] Mariano Di Martino, Pieter Robyns, Winnie Weyts, Peter Quax, Wim Lamotte Lamotte, and Ken Andries. Personal Information Leakage by Abusing the GDPR "Right of Access". In *Symposium on Usable Privacy and Security*, SOUPS, 2019.

[11] Dubhe Sarmiento Félix and Steve Wright. Data Privacy Progress, Enforcement and Brexit. *Journal of Data Protection & Privacy*, 3(4), 2020.

[12] Pietro Ferrara and Fausto Spoto. Static Analysis for GDPR Compliance. In *Italian Conference on Cybersecurity*, ITASEC, 2018.

[13] S. Gürses and J. M. del Alamo. Privacy Engineering: Shaping an Emerging Field of Research and Practice. *IEEE Security Privacy*, 14(2), 2016.

[14] Dominik Huth and Florian Matthes. "Appropriate Technical and Organizational Measures": Identifying Privacy Engineering Approaches to Meet GDPR Requirements. In *America's Conference on Information Systems*, AMCIS, 2019.

[15] Ety Khaitzin, Roee Shlomo, and Maya Anderson. Privacy Enforcement at a Large Scale for GDPR Compliance. In *Systems and Storage Conference*, SYSTOR, 2018.

[16] Mirosław Kutyłowski, Anna Lauks-Dutka, and Moti Yung. GDPR – Challenges for Reconciling Legal Rules with Technical Reality. In *European Symposium on Research in Computer Security*, ESORICS, 2020.

[17] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I. Hong. How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of a Developer Forum on Reddit. *Human-Computer Interaction*, 4(CSCW3), 2021.

[18] Harshvardhan Pandit, Declan O'Sullivan, and David Lewis. Test-Driven Approach Towards GDPR Compliance. In *Conference on Semantic Systems*, SEMANTiCS, 2019.

[19] Luca Piras, Mohammed Ghazi Al-Obeidallah, Andrea Praitano, Aggeliki Tsohou, Haralambos Mouratidis, Beatriz Gallego-Nicasio Crespo, Jean Baptiste Bernard, Marco Fiorani, Emmanouil Magkos, Andrès Castillo Sanz, Michalis Pavlidis, Roberto D'Addario, and Giuseppe Giovanni Zorzino. DEFeND Architecture: A Privacy by Design Platform for GDPR Compliance. In *Conference on Trust and Privacy in Digital Business*, TrustBus, 2019.

[20] Wanda Presthus and Kaja Felix Sønslien. An Analysis of Violations and Sanctions Following the GDPR. *Journal of Information Systems and Project Management*, 9(1), 2021.

[21] Jukka Ruohonen and Kalle Hjerppe. Predicting the Amount of GDPR Fines. https://arxiv.org/abs/2003.05151, 2020. arXiv eprint 2003.05151.

[22] Jukka Ruohonen and Kalle Hjerppe. The GDPR Enforcement Fines at Glance. https://arxiv.org/abs/2011.00946, 2021. arXiv eprint 2011.00946.

[23] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control. In *Proceedings of the 14th ACM Asia Conference on Computer and Communications Security*, AsiaCCS '19, pages 340–351, New York, New York, USA, 2019. ACM Press.

[24] Subhadeep Sarkar, Jean-Pierre Banatre, Louis Rilling, and Christine Morin. Towards Enforcement of the EU GDPR: Enabling Data Erasure. In *IEEE International Conference of Internet of Things*, iThings, 2018.

[25] Awanthika Senarath and Nalin Arachchilage. Understanding User Privacy Expectations : A Software Developer's Perspective. *Telematics and Informatics*, 35(7), 2018.

[26] Awanthika Senarath and Nalin A. G. Arachchilage. Why Developers Cannot Embed Privacy into Software Systems? An Empirical Investigation. In *Conference on Evaluation and Assessment in Software Engineering*, EASE, 2018.

[27] Ze Shi Li, Colin Werner, Neil Ernst, and Daniela Damian. GDPR Compliance in the Context of Continuous Integration. https://arxiv.org/abs/2002.06830, 2020. arXiv eprint 2002.06830.

[28] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. Understanding Privacy-Related Questions on Stack Overflow. In *Conference on Human Factors in Computing Systems*, CHI, 2020.

[29] The Future of Privacy Forum. New Decade, New Priorities: A Summary of Twelve European Data Protection Authorities' Strategic and Operational Plans for 2020 and Beyond. https://fpf.org/wp-content/uploads/2020/05/FPF_DPAStrategiesReport_05122020.pdf, 2020.

[30] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. "Your Hashed IP Address: Ubuntu.": Perspectives on Transparency Tools for Online Advertising. In *Anual Computer Security Applications Conference*, ACSAC, 2019.

[31] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. A Study on Subject Data Access in Online Advertising after the GDPR. In *International Workshop on Data Privacy Management*, DPM, 2019.

[32] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *ACM Conference on Computer and Communications Security*, CCS, 2019.

[33] Josephine Wolff and Nicole Atallah. Early GDPR Penalties: Analysis of Implementation and Fines Through May 2020. *Journal of Information Policyl*, 11(1), 2020.

A Appendix

A.1 Stopwords

For our word frequency analysis, we manually added the following data protection terminology stopwords: *'data', 'personal', 'fine', 'company', 'purpose', 'fined', 'breach', 'violation', 'technical', 'subject', 'subjects', 'measures', 'organizational', 'number', 'authority', 'security', 'dpa', 'eur', 'information', 'gdpr', 'subject', 'processing', 'controller', 'art', 'due', 'also', 'without', 'aepd', 'found', 'addition', 'processed', 'spanish', 'protection', 'municipality', 'however', 'could', 'italian', 'party', 'imposed'*.

No.	Category of Fine	Total Occur.	Technical	Organizational
1	**Insufficient Security Measures**	386	340	269
1a	Digital loss of data	5	5	1
1b	Insufficient measures to effectively combat attacks/ misuse	186	174	132
1c	Physical loss of data	16	9	13
1d	Missing anonymization or pseudonymization	16	9	14
1e	Insufficient data life cycle	56	44	51
1f	Missing role management	52	48	27
1g	Unauthorized access	55	51	31
2	**Unauthorized Data Processing**	751	350	627
2a	Unauthorized disclosure of data	225	134	187
2b	Processing personal data without legal basis	409	174	356
2c	Insufficient consent	37	16	26
2d	Illegal camera surveillance scope	80	26	58
3	**Data breach information/ DPA cooperation**	95	47	81
3a	Not reporting a data breach to the DPA	29	22	27
3b	Not reporting a data breach to the data subject	17	13	15
3c	Insufficient cooperation with DPA	49	12	39
4	**Data Subject Rights**	182	94	160
4a	Insufficient 'right of access'	62	28	53
4b	Insufficient 'right to rectification'	10	6	9
4c	Insufficient 'right to erasure'	56	32	49
4d	Insufficient 'right to restriction of processing'	10	4	10
4e	Insufficient 'right to data portability'	2	1	2
4f	Insufficient 'right to object'	42	23	37
5	**General Obligations**	91	64	65
5a	Failing to appoint a data protection officer	8	2	6
5b	Missing management system for data protection	5	0	2
5c	Missing register of processing activities	5	4	4
5d	Missing data protection impact assessment	16	10	14
5e	Privacy unfriendly design	57	48	39
6	**Information Obligations**	169	75	129
6a	Missing/ incorrect privacy policy	106	58	86
6b	Camera surveillance without notice	53	11	34
6c	Missing/ incorrect cookie policy	10	6	9
7	**Violation of Basic Data Protection Principles**	125	76	92
7a	Injust processing of article 9 data	125	76	92

Table 5. The extended list of fine categories based on the *ET*. Categories are extended by the total number of occurrences of them and the share of technical and organizational fines. Based on our manual classification, fines could be assigned to either one origin or to both origins.

A.2 Detailed Listing of GDPR Fine Categories

Table 5 shows all categories that we used in our analysis, as discussed in Section 3.