Structuring the End of the Data Life Cycle

Daniel Tebernum[1] and Falk Howar[2]
[1]*Data Business, Fraunhofer ISST, Emil-Figge-Strasse 91, 44227 Dortmund, Germany*
[2]*Chair for Software Engineering, TU Dortmund University, Otto-Hahn-Strasse 12, 44227 Dortmund, Germany*

Keywords: Data Engineering, Delete Data, Taxonomy, end of Data Life Cycle.

Abstract: Data is an important asset and managing it effectively and appropriately can give companies a competitive advantage. Therefore, it should be assumed that data engineering considers and improves all phases of the data life cycle. However, data deletion does not seem to be prominent in theory and practice. We believe this is for two reasons. First, the added value in deleting data is not always immediately apparent or has a noticeable effect. Second, to the best of our knowledge, there is a lack of structured elaboration on the topic of data deletion that provides a more holistic perspective on the issue and makes the topic approachable to a greater audience. In this paper, an extensive systematic literature review is conducted to explore the topic of data deletion. Based on this, we present a data deletion taxonomy to organize the subject area and to further professionalize data deletion as part of data engineering. The results are expected to help both researchers and practitioners to address the end of the data life cycle in a more structured way.

1 INTRODUCTION

Data is an important asset that can offer enormous added value if managed, processed, and used in a strategic way. There is a general consensus in the literature and practice that well-managed and high-quality data influences business agility for the better (Otto, 2015; Tallon et al., 2013). It allows for further improved business processes, enhances organizational development, and supports business innovation (Amadori et al., 2020; Azkan et al., 2021). It is therefore not surprising that the phases of the data lifecycle that directly add value, such as the selection, use and transformation of data, are the very ones that are being researched in theory and practice.

We believe that data deletion is also an important part of the data life cycle and should be studied as extensively as the other more prominent phases. In our previous work, we found evidence that deleting data gets little attention in the literature (Tebernum et al., 2021), as well as in practice (Tebernum et al., 2023). One reason for this could be that deleting data rarely provides any visible economic value. We argue that dealing professionally with the end of the data life cycle is an important prerequisite for addressing current and future challenges that affect not only companies but also government institutions and private individuals. To name just a few where data deletion has significant relevance: Legal regulations like the EU Data Protection Regulation (GDPR) (European Commission, 2016) that forces companies to have a complete overview of their data and to be able to delete data in accordance with the law. The energy Crisis 2022 (Council of the EU, 2022) brings topics such as green IT and digital decarbonization into focus, where topics such as the ever-growing amount of data and the associated waste of resources are explored. It is assumed that much of the data generated will be used only once (Trajanov et al., 2018), but will continue to cost energy. Also, when we look into the realm of data sovereignty, we find good reasons to delete data, as e.g., it is of utmost importance that one retains control over one's own data.

These and many other motivations underline the need for further professionalizing and structuring the end of the data life cycle. To the best of our knowledge, there is a lack of structured elaboration on the topic of data deletion that provides a more holistic perspective on the subject. We argue that such an overview of the subject area is important as more and more requirements on data management are imposed, which also affect the deletion. To achieve this goal, this paper addresses the following research questions:

- RQ1: To what extent is data deletion currently reflected in literature?
- RQ2: What are the key building blocks that de-

Tebernum, D. and Howar, F.
Structuring the End of the Data Life Cycle.
DOI: 10.5220/0011999300003541
In *Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023)*, pages 207-218
ISBN: 978-989-758-664-4; ISSN: 2184-285X

scribe the end of the data life cycle?

- RQ3: How can the field of data deletion be structured in general?

To answer these questions, we conducted an extensive systematic literature review, which allowed us to gain an overview of the current state and focus of research. We were able to identify key building blocks of data deletion and associated subtopics. The results led to a general structuring of the subject area in the form of a data deletion taxonomy. The taxonomy is intended to help researchers and practitioners gain an overview of the aspects of data deletion and subsequently integrate them into their own work in a structured manner. Finally, research gaps were identified and presented as possible future work.

The remainder of the paper is organized as follows. In Section 2, we describe our research methodology in great depth. Then, in Section 3, we present our novel result, a data deletion taxonomy. Here we go into detail about all aspects that are a part of the taxonomy and also extend the evaluation. In Section 4, we review and discuss our research findings. We are guided here by the given research questions. Finally, in Section 5, we summarize our results and provide an outlook for future work.

2 RESEARCH METHODOLOGY

This section presents a detailed description of the research methodologies applied in this work. As described previously, we were able to find indicators that data deletion in the field of data engineering has a low level of maturity. However, deleting data was not studied as a main subject in these publications (Tebernum et al., 2021; Tebernum et al., 2023). It is therefore necessary to provide a more general overview of the state of research regarding this topic. For this purpose, a systematic literature review (SLR) was performed. It is intended to identify the extent to which data deletion, and thus the end of the data life cycle, has received attention in literature and to identify the key building blocks. The SLR performed follows the process shown in Figure 1. Additional The process can be divided into four phases.

Phase #1: The first phase deals with the generation of the corpus that will be studied. First, a bibliography suitable for the project had to be identified. Meta bibliographies are particularly useful for this purpose, as they index a number of available bibliographies. Scopus[1] was chosen because of its flexibility in formulating a search string. The search string (see Listing 1) itself has been continually improved over several iterations to find the most comprehensive range of topic-related publications.

```
TITLE-ABS-KEY((destroy* OR destruct* OR
 delet* OR remov* OR eras* OR wip*) W/0
 data) AND (LIMIT-TO(SRCTYPE, "p")) AND
 (LIMIT-TO(PUBSTAGE, "final")) AND (
LIMIT-TO(SUBJAREA, "COMP")) AND (LIMIT-
TO(LANGUAGE, "English"))
```

Listing 1: Scopus search string.

After running the search, 596 conference papers were identified.

Phase #2: The second phase deals with cleaning the corpus. Despite filtering the search results, there were still publications that were not papers, were not accessible, or were not written in English. These and some duplicates have been removed. The resulting corpus was reduced to 586 papers. To decide which papers were relevant to determine the state of the research, all abstracts were read by the authors. Papers that did not focus on deleting data were removed. This resulted in 121 relevant papers remaining.

Phase #3: The third phase involves the analysis of the papers remaining in the corpus. To be able to determine the state of research, the grounded theory methodology (GTM) is applied according to Glaser (Glaser et al., 1968) and Corbin (Corbin and Strauss, 1990). Due to the number of papers left, open coding was performed on the abstracts. Since the abstracts should already contain the essence of the papers, this is sufficient to generate an overview of the state of the research. Open coding was performed in an iterative process. The codes were written using the in-vivo method. During iterations, codes that meant the same thing were merged. After the last iteration, 141 codes remained.

To further enrich the analysis, on top of the manually extracted codes, additional codes were systematically extracted by an automated mechanism. For this task, algorithms used for keyword extraction, such as TextRank (Mihalcea and Tarau, 2004), RAKE (Rose et al., 2010), and PositionRank (Florescu and Caragea, 2017), can be applied and were considered by the authors. Thushara et al. showed in their work that PositionRank is slightly superior to the other algorithms (Thushara et al., 2019). Therefore, PositionRank was chosen by the authors. As with the manual coding process, the algorithm was applied to the abstracts of the papers. To reduce the number of keywords and boost their quality, only those keywords that had a PositionRank score above 0.15 were considered. The codes were deduplicated and reduced, resulting in 69 additional codes. In total, 210 codes were generated.

[1] https://www.scopus.com/

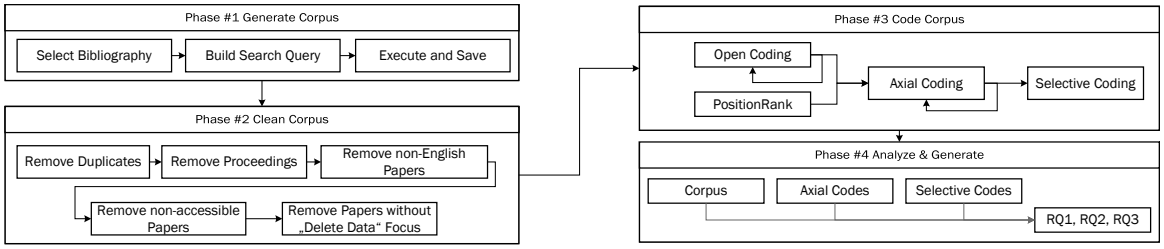

Figure 1: Research Methodology.

Subsequently, an axial coding was carried out. In this process step, open codes were grouped and assigned to higher-level terms. Again, this was an iterative process that was repeated by the authors until a satisfactory knowledge saturation was reached. A total of 44 axial codes could be identified.

Finally, selective coding was performed based on the axial codes found. In this phase, the found categorizations and supergroups are further combined into overarching ideas. A total of 6 codes were identified.

Phase #4: In phase four, the previously generated corpus and GTM codes are further analyzed. For this purpose, several evaluations were performed. The results are discussed in Section 5. First, we looked at whether the topic of deleting data had any anomalies in publication density and quantity. For this, we have aggregated the corpus in buckets per year. For reference, we used the general publications related to data engineering from Scopus. This dataset was cleaned of publications present in data deletion. The trends were calculated using the ordinary least squares method. Next, we examined how frequently the axial and selective codes in our corpus were assigned during coding and over the years. The values were rendered in a heat map to visualize clusters and increase well (see Figures 3 and 4). Finally, with the gained insights, the topic of data deletion was structured by generating a taxonomy.

A listing of all 121 relevant papers and the GTM codes is available online[2].

3 A DATA DELETION TAXONOMY

In this section, we present our novel result; a data deletion taxonomy. The taxonomy can be seen in Figure 2. The creation of the taxonomy is closely oriented to the GTM codes found. In general, the selective codes were used as starting points. The remaining taxonomy was created by the appropriate insertion of the axials or matching open codes. Where appropriate and the corpus seems to have gaps, the content was added by the authors. Not all subtopics can be enumerated exhaustively. Such cases are indicated by "...". The taxonomy answers RQ3 by providing a general structuring of the data deletion topic. In the following, the parts of the taxonomy are explained in detail.

3.1 What

The most important question that needs to be answered in the context of data deletion is the question of what data is under consideration. This means, that the data to be deleted must be identified and addressed.

- **Identify Data:** Identifying/finding the data to be deleted represents an important issue. *Tools*, *methods*, or *languages* can be used. *Tools* primarily include holistic systems that can analyze data and relate them to each other or the environment. In particular, *data catalogs* that store metadata about the data should be mentioned here (Ehrlinger et al., 2021). The metadata can be used as a decision-making tool for deleting data. E.g., the catalog may have stored information about technical representations or data quality metrics that provide a reason for the data to be deleted. *Methods* include everything that creates an indicator for or against the deletion of data. These can be, e.g., solutions that determine data quality, or similarities (Long et al., 2020) to data already marked for deletion. In particular, duplicates (Chen and Chen, 2022; Rashid et al., 2012; Pachpor and Prasad, 2018) and sensitive data (Pecherle et al., 2011) are often mentioned in our corpus for deletion. *Languages* include anything that allows data to be identified by giving a set of instructions according to a given grammar. E.g., *SQL* can be used to identify/find data. One could select data whose age exceeds a certain level.
- **Address Data:** The addressing must be unambiguous to avoid any confusion and to possibly enable automation of the deletion process. By ad-

[2]https://doi.org/10.5281/zenodo.7867278

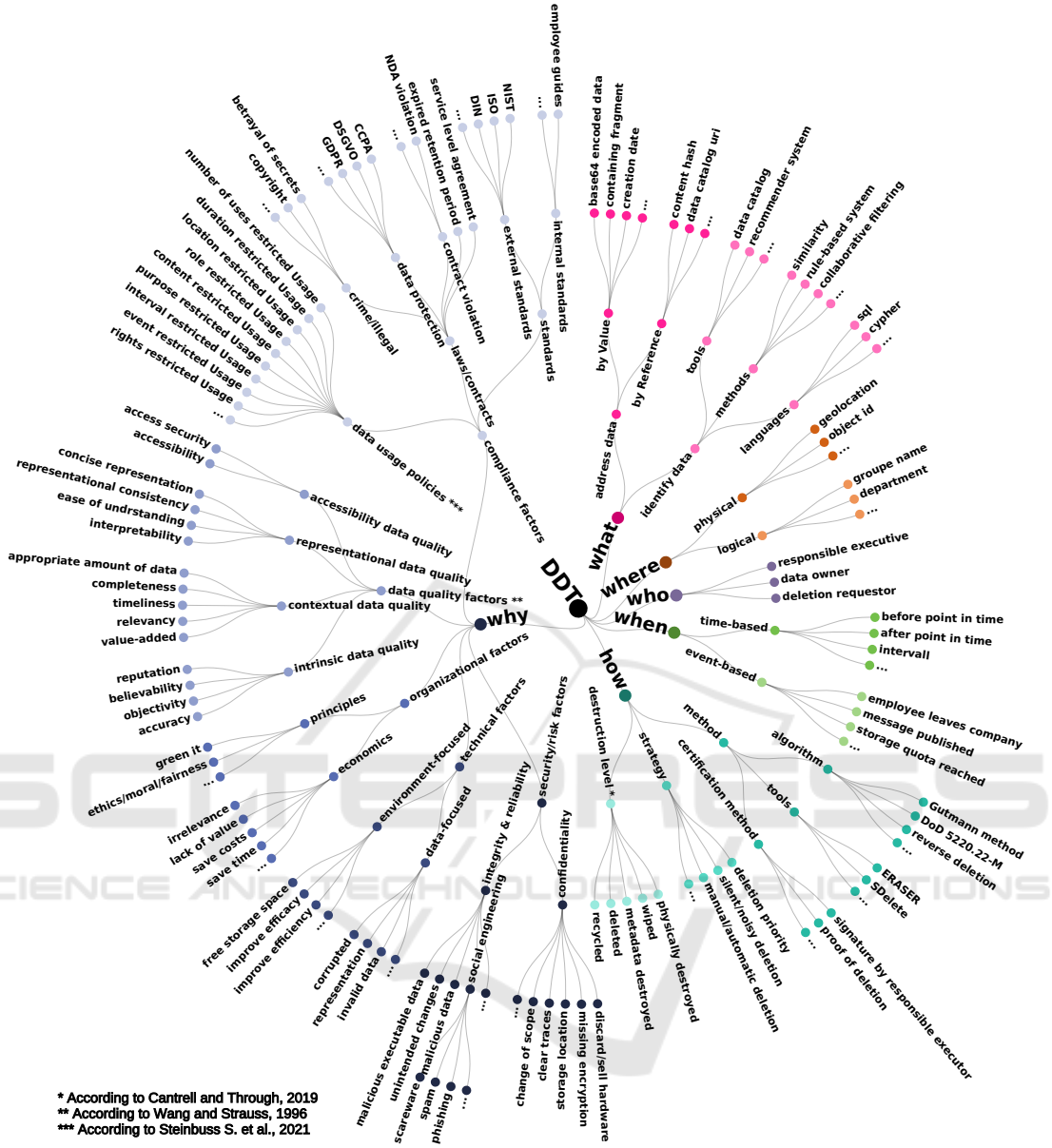

Figure 2: Data Deletion Taxonomy (DDT).

dressing we do not mean any kind of data locator, but a one-to-one identification of the data to be deleted. E.g., if we consider files, a unique address could be a hash of the contents. When we look at data in a database, we need, e.g., the connection endpoint, the specific database, and, e.g., in relational databases, an SQL statement that selects exactly the data to be deleted.

Describing the data accurately and unambiguously is a difficult undertaking. The heterogeneity of the data sources complicates this. For this reason, it is recommended to reference the object to be deleted. This, e.g., can be done by using a *data catalog uri*. The data catalog must be able to map individual data assets and also just parts of them via URIs. These URIs can then be used to look up exactly what data is involved within the data catalog itself. If we follow the previous examples, the data catalog would provide the file's content hash or the database endpoint and SQL statement. This does not solve the problem of heterogeneity and almost infinite description and addressing possibilities. But in this case, we leave the description to the technology that knows best how to address data.

3.2 Why

Deleting data is not always easy to undo. For this reason, it is crucial to be confident about what you are doing. This confidence is supported in part by the existence of good reasons for deletion. In other words, the *why* plays an important role here. If there were no reasons for deletion, data would hardly ever be deleted, because there is no such thing as the natural selection here. To establish a common ground of why certain data should or must be deleted, we need a shared understanding of the reasoning. During the SLR, serveral reasons have been identified that may necessitate the deletion of data. In practice, these standardized reasons can be used to achieve some effects. E.g., depending on the reason for deletion, other processes can be triggered. In addition, a common understanding of the reason can increase the acceptance of deletion requests. The following subchapters go into detail about the reasons for deleting data.

Security/Risk Factors

This category includes all security reasons for which data should be deleted, i.e., all aspects where attackers could cause undesired effects, in this case through data.

- **Confidentiality:** This refers to all situations where the unwanted disclosure of the content of the data can be prevented by deleting the data itself. This includes reasons such as the hardware, on which the data is stored, being discarded or sold. Data on *discarded hardware*, be it hard drives, optical media, or flash storage, can be a major risk. The research literature includes a plethora of papers that cover the topic of secure data erasure on various storage media. In this case, it makes sense to delete the data beforehand. However, a security risk can also exist if the data is *missing encryption* or is located in storage areas where it can be accessed by unauthorized persons. By deleting the data, we may still be able to intervene in time here. Likewise, data should be deleted if it is unintentionally possible to use it to trace events or persons. Furthermore, if people still have access although they are no longer authorized due to changes in access rights, the data should be deleted.
- **Integrity & Reliability**: This category includes everything where data itself would lead to improper alterations or malicious behavior of our systems, data, or services. One subcategory would be *malicious data*. This is data that is intended to achieve undesirable effects when applied in specific applications. This data is not executable but develops its effect only in the context of its application. Depending on the characteristics, it can represent a servere security risk. This data, if known to exist, should be destroyed so that it is not accidentally deployed or applied. Poisoning attacks are one example where we want to delete malicious data (Chiba et al., 2020). Here, corrupted data is introduced into the machine learning training process to weaken the resulting model. Another class is represented by the *malicious executable data*. This refers to data that can act as an executable program and cause damage. This includes, e.g., viruses, trojans, or worms. Also, many situations from the area of *social engineering* can be a reason to delete data. There, data can only unleash its dangerous potential in combination with a human actor. Examples are phishing, baiting, or scareware that is sent as email or over other communication channels and that should be deleted. Also covered by the point of integrity are *unintended changes* to data that, e.g., trigger unwanted and possibly safety-relevant effects in the environment.

Technical Factors

Here, all reasons to delete data, that have a primarily technical nature, are included. This can be divided into reasons that lie in the data itself and reasons where the deletion of data affects the environment.

- **Data-Focused:** This category includes reasons where the data itself is the reason why technically something is not working and the data should be deleted. This includes, e.g., data that is syntactically correct but in a *representation* that the system cannot process. Another reason would be if we have *corrupted data* that is simply broken, e.g., blocks were lost during transmission, Also *invalid data*, whose content negatively affects the behavior of the system, could be the target of deletion (Othon et al., 2019). This could be, e.g., test data that needs to be removed from a system.
- **Environment-Focused:** This category includes deletion reasons that affect the environment. This includes, e.g., to *free storage space*. Even though memory is cheap, in certain areas such as embedded systems it may be necessary to delete data to provide memory. Sometimes deletion can also *improve efficiency* (Lin et al., 2009; Reardon et al., 2012; Gao et al., 2019; Pachpor and Prasad, 2018) and *improve efficacy* of the system, e.g. because unimportant data is no longer included in calculations.

Organizational Factors

This includes reasons for deletion that are primarily shaped by human actors in the context of companies. It can be subdivided into economic reasons, in which the deletion of data offers an economic advantage, and principles or values, which can be represented and according to which one acts.

- **Economics:** One reason to delete data can be to *save costs*. This can be the case if, e.g., the costs for data storage are significant. Another reason can be to *save time*. E.g., processes such as backups can be accelerated. Sometimes it turns out that data has a *lack of value* for one's own project or has completely gotten *irrelevant*. This can also be a reason for deletion.
- **Principles:** This category deals with principles that one can follow. *Ethics* represents a sub-item. If not already regulated by law, it may be necessary to delete data because it does not generate fair models during machine learning training. The topic of *green IT* can also play an important role (Van Bussel and Smit, 2014). E.g., deleting data on a large scale can benefit the environment by reducing the amount of hardware needed for storage, or by reducing the amount of power needed for network transfers. This is a very important issue as we look at the EU energy crisis in 2022 (Council of the EU, 2022).

Data Quality Factors

This category includes all reasons for which data should be deleted due to quality aspects. This category includes metrics that describe dimensions of data quality. The listing for data quality used in this work is guided by the work of Wang and Strong who created a conceptual framework for data quality (Wang and Strong, 1996). The data quality metrics identified there as part of the "Conceptual Framework of Data Quality" can be used one-to-one as reasons for why one could want to delete certain data. E.g., one could encourage the deletion of data if the accuracy and correctness of a data set is not good enough (Othon et al., 2019). It can be important when the data produces a bad result in processes such as machine learning training and one wants to prevent further accidental use of the data set. A detailed explanation of the metrics can be taken from the work of Wang and Strong.

Compliance Factors

Compliance can be an important reason for deleting data. There are several types of compliance factors that can be distinguished. The subpoints are described in more detail below.

- **Data Usage Policies:** Cross-company data exchange will play an increasingly important role in the future. Maintaining sovereignty over one's own data is an important goal in order to motivate the participants to disclose their data. This is a niche for projects such as Gaia-X[3] and IDSA[4], which aim to create an environment in which trustworthy data exchange is possible. One aspect of this is *data usage policies*, which regulate the use of data. From a data deletion perspective, the invalidity of a usage policy may be a reason to not only stop using data, but to delete it. E.g., the *role restricted usage* policy allows data to be used if a user has a specific role. If a user is stripped of this role, it would be logical to also delete the data from the users system. The *data usage policies* shown here are taken from the work of Steinbuss et al., who managed to collect a set of important policies in regard to data usage (Steinbuss et al., 2021).
- **Laws/Contracts:** First, there are mandatory regulations issued by the legislature within the framework of *laws*. The applicable laws differ depending on the location and must be adapted appropriately. Deleting data may become necessary if the data violates a law and constitutes a *criminal* offense or is otherwise *illegal* in some way. This may be the case, e.g., in the event of a *copyright* infringement or the *disclosure of government secrets*. In this case, the data may be deleted. Under many personal *data protection* laws, like the EU *GDPR* (European Commission, 2016), deletion may also be necessary upon request of the owner of the data (Sarkar et al., 2018; Kuperberg, 2020). In some cases, data must be retained for a certain period of time. When this time is over, the data may be deleted accordingly.

  Second, *contracts* also establish a binding nature in this way between business partnerns, often backed up by laws. It may be part of the contract to delete data in certain situations, e.g., at the end of the contract period. However, the breach or end of a contract may also result in data having to be deleted.
- **Standards:** *Standards* are another type of regulation that, unlike laws, are not always mandatory. *External Standards* are often used to maintain a certain specified quality and are developed and

[3]https://gaia-x.eu/

[4]https://internationaldataspaces.org/

published by organizations or groups like *DIN*[5], *ISO*[6], or *NIST*[7]. E.g., there is the ISO 27000 series, which deals with information security. In the sub-specifications, e.g., the deletion of data can be defined as a standard in response to certain events. Another type of standards are *Internal Standards*, which are only used within a company or a certain area. Here, too, situations may arise that require the deletion of data.

3.3 Who

When it comes to the topic of *who*, both a natural person or a legal entity are possible. In addition to this general establishment, the question of *who* has three aspects to it that must be answered depending on the context. The first aspect deals with the description of persons that should delete the data described in *what*. This must be described if the data is to be deleted only for certain persons. An example would be if a license for use has been purchased, but the duration was different for various users. The second aspect deals with *who* is responsible for deleting the data in the first place. It is recommended to model this aspect of *who*, so that the executing instances have a contact person for inquiries. Specifying such a person can also increase confidence that the deletion of data is reasonable and should be carried out. Last, it may be important to indicate the person responsible for the data itself if different from the previous person. This person can also serve as a contact to answer questions and to strengthen trust. Cross-role topics should also be addressed, such as the understanding of data deletion (Murillo et al., 2018) or the formal modelling of the user (Del Tedesco and Sands, 2009). User-friendliness can also be an important topic here when it comes to human actors. Among other things, this involves the perception of the users (Diesburg et al., 2016).

3.4 When

The questions about *when* to delete can be divided into two subcategories, *time-based* and *event-based*. It must only be considered whether it is not already known by convention or implicitly.

- **Time-Based:** When(ever) time is the focus of consideration, we speak of a *time-based* reasoning. Here, a point in time can be meant. E.g., data must be deleted from or up to a certain point in time. However, time intervals or recurring events, such as "every Monday", can nevertheless be modeled if appropriate. When modeling times, it must also always be clearly defined which time is meant. The time zone for the deleting person or tool can be different than the time zone in which the physical copy of the data is located.
- **Event-Based:** Sometimes data should be deleted when certain conditions have been met. Here, a specific event is the trigger and time is not the primary focus. These events can be defined very broadly. E.g., one could model that data should be deleted when a certain percentage of memory on the hosting system is occupied or if a new version of the data is made available.

3.5 Where

The answer to this question has two different aspects in it, *physical* and *logical*. These two modeling possibilities, *physical* and *logical*, can also be combined so that more complex requirements can be addressed. If the *where* is not already implicitly determined by the context, e.g. because the data is only located in one place or every copy of the data should be deleted, the question of *where* must be answered.

- **Physical:** First, the question of *where* can be answered with regards to something *physical*. E.g., it is possible to specify that data from a specific region should be deleted. If data is located in a certain cloud region, e.g., outside of Europe, it is possible to specify that data should be deleted. It is also possible to model areas where data is to be deleted by accurately specifying the geo-coordinates in using points, polygons, or other shapes. Sometimes addressing a physically existing *object* like a smartphone is also sufficient if the data contained on or in it is to be deleted.
- **Logical:** Second, the question of *where* can be answered with a *logical* response. E.g., data should be deleted from computers that belong to a specific *group* or a *department*. There can also be a selection based on *device classes*, *storage classes*, and many others.

3.6 How

This question is primarily about modeling specific deletion methods, overarching deletion strategies, and the level of destruction that will be applied for this particular data. The *how* does not describe how the data can be accessed to technically perform the data deletion in practice. In the following, we will take a

[5]https://www.din.de/
[6]https://www.iso.org/home.html
[7]https://www.nist.gov/

closer look at which aspects of the question play an important role when deleting data.

- **Methods:** *Methods* are standardized procedures that are intended to achieve a certain result when deleting data. This includes, e.g., the application of special deletion *algorithms* to be used (Wang and Zhao, 2008; Subha, 2009; Xu et al., 2014). One could think of different methods to wipe data on hard drives (Wei et al., 2011; Chen et al., 2019). In the last couple of years, methods for deleting data from machine learning models have also become of interest (Nguyen et al., 2022; Graves et al., 2021; Nguyen et al., 2020; Ginart et al., 2019). Sometimes deleting data requires the use of special *tools* (Riduan et al., 2021; Martin and Jones, 2011; Žulj et al., 2020; Sahri et al., 2018). This can be the case, e.g., when users are asked to delete data manually and are provided with software that removes data according to certain criteria. In certain situations, it is necessary for the deletion of data to be documented by a *certification* (Guo et al., 2019). In this case, the certification method to be used should be specified. This is necessary if you need to guarantee or prove (Klonowski et al., 2019) that data has really been deleted. Another point that falls under *methods* is the secure deletion of data. Here, it should be modeled how it is ensured that the deletion happens comprehensively and correctly.
- **Deletion Strategies:** This includes aspects that have no direct technical relation, but can implicitly influence them. The aspects are viewed from a higher-level perspective. The *deletion priority* indicates how urgently a deletion job must be executed. A lower priority can, e.g, save costs in the cloud area if only free, cost-effective capacities of the provider are used. Another strategy is to cut costs as much as possible. This can be achieved through various solution approaches mentioned before like by reducing the priority to favorable computing contingents or, e.g., by trade-offs between different algorithms. Sometimes it may happen that you can choose between *manual* and *automatic* execution. Depending on the selection, this may result in further effects. If *manual* deletion is desired, the persons involved should be notified. In the case of *automatic* deletion, the specific automatism may still need to be specified. Another possible strategy deals with how much deletion needs to be communicated. There are situations where deletion can be done *silently* through automated processes. E.g., if the data is stored on the computers of employees, the user can be informed that the data is being deleted in the interest of trustworthy work. Users should always have the option of preventing (semi-)automated deletion on their end devices. This could otherwise be perceived as an infringement of one's own sovereignty.
- **Destruction Level:** These levels are taken from the work of Cantrell and Through and originally describes the level of recoverability of data (Cantrell and Through, 2019). In the context of this taxonomy, the meaning is adapted to the question of the degree to which the data should be deleted. Generally speaking, the lower the level of destruction, the easier it is to recover the data. *Recycled* means, e.g., that the data is only moved to the recycle bin for the time being. This can be easily undone by any user. In the next step, the data is marked as *deleted* by the operating system. The data and metadata is still mostly left in the file system. Data can be recovered using special tools. When we reach the level of *metadata destroyed*, there is no more indication that data may have existed. One needs to check the whole storage using specific tools. By *wiped*, the data in the storage was overwritten using specific methods. This makes it almost impossible to recover data. The most effective way to delete data is when the storage is *physically destroyed*. Done right, no data can be recovered. While the reliability of data deletion continues to increase with each level, it also has an increasing impact on cost, time, and effort.

4 DISCUSSION

In this section, the previously gained results will be discussed. The discussion is structured around the research questions established in Section 1.

RQ1: First, we wanted to identify to what extent data deletion is reflected in the literature. We compared the publication quantities from data deletion with those from the field of data engineering from 2000 to 2022. The publications of the *International Conference On Data Engineering* were used as a comparison dataset. When performing the interpretation, it should be noted that the papers from the data deletion corpus come from various conferences. Thus, it cannot be ruled out that the popularity of the conference influences the trend more than the interest in data engineering itself. Also, growth may be limited by the conference itself. We could see that both subject areas have experienced growth. A general growth can easily be explained by the fact that overall, the

volume of publications has increased over the years. Although data deletion is showing slightly stronger growth than data engineering publications, the topic is not currently experiencing a real trend. The still very small number of publications, ranging from 0 to 15 per year, indicates a general weariness towards the subject. The small number of 121 publications in 23 years that have data deletion as their core theme, compared to about 3000 publications at just one single data engineering conference, cannot do justice to the subject area. It can be assumed that an increasing number of publications is necessary for the professionalization of data deletion.

The next step is to examine how the topics are distributed in the publications found. For this purpose, we used the axial codes found. As can be seen from Figure 3, there are several frequently discussed topic areas, while others fall far behind. The leading codes are *method*. *security*, and *hardware*. We believe that these topics are addressed particularly frequently because they are often addressed together and offer the greatest value when implemented. *Methods* contains e.g. work on algorithms and processes. These works are necessary as a basis in the first place to delete data. When it comes to *security*, a lot of work deals with deleting it in such a way that it cannot be recovered by third parties. The *hardware* plays a decisive role here since appropriate methods must be developed based on the hardware in use. Another important reason for deleting data is *data quality*. The main aim is to identify data with insufficient quality and then delete it. When it comes to *laws*, there is a great interest in complying with them to avoid fines. The introduction of GDPR in 2018 has certainly contributed to the increase in the volume of publications. This assumption is supported by the increasing number of *laws* publications starting around 2017. Perhaps the reason that the topic of *efficiency* comes up so often is that people are always trying to do things more economically. Another trend can be found in the *machine learning* area. The deletion of data from already trained models has experienced increased interest for the last 3 years. Regarding the other codes, we see a possible need for action, particularly in the fields that are poorly represented or even not represented at all. These are often topics that do not directly concern the deletion process but topics of structuring, standardization, and management of deletion. Overall, it can be said that many topics are severely underrepresented in this research area.

RQ2: Second, we wanted to identify the key building blocks of data deletion. By analyzing the literature,

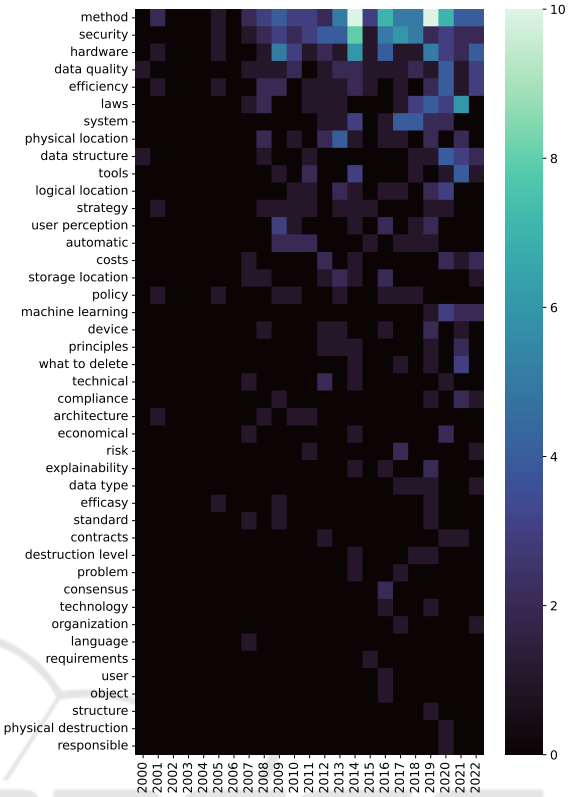

Figure 3: Axial code distributions over years.

these could be identified during the selective coding process. In order to describe the deletion of data extensively, the questions of *what*, *why*, *who*, *when*, *where*, and *how* must be answered. When it comes to describing the deletion of data in particular, it is not always necessary to answer all questions. Often, questions about, e.g., the method or the location where data is to be deleted arise from the context. Also, we looked at the distribution of the selective codes. The distribution can be seen in Figure 4. The evaluation clearly shows that the deletion of data is not only considered as an end in itself, but that there is always a reason (*why*) that makes the deletion of data necessary in the first place. In 96 papers, the abstract already motivated why data should be deleted at all. The *how* is also highly prominent. This is primarily due to the fact that the how is composed of, among other things, *methods*. These have already been assigned most frequently in the axial codes. Surprisingly, the *where* is also prominent to such an extent. We believe this is mainly due to the fact that many papers write about deleting data on hard drives or in the cloud. Much less often the papers deal with the more organizational topics of *what*, *who*, and *when*. In particular, the fact that very few papers report on the influence of or impact on

human actors (*who*) (e.g. (Diesburg et al., 2016) and (Murillo et al., 2018)) is a major research gap. Deleting data is an experience that every person who operates a computer has. The fact that more and more people are working with data that will have to be deleted at some point should also be reason enough to investigate the human factor more closely. A further area that almost no work has to its core is the *when*. We believe this is due to the fact that the time at which one wants to delete data is very individual and there are few generalizable situations that would justify a research contribution. Nevertheless, it would be interesting to see what generalizations are possible here. As can also be seen, the *what* shows increased publications in the last 4 years. This growth can be attributed in particular to work where parts from trained models are to be deleted within the machine learning domain. In addition to the deletion methods, these works also describe exactly the data that is to be deleted. Unfortunately, with the available data, it is not possible to predict how trends might develop further.

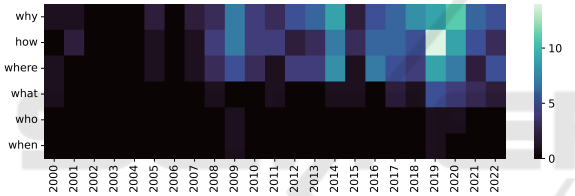

Figure 4: Selective code distributions over years.

RQ3: Third, we wanted to structure the field of data deletion in general. We provided an answer to this question with a novel data deletion taxonomy presented in Section 3. The taxonomy structures the subject areas found in the SLR in a hierarchical order. The taxonomy itself shows how extensive the subject area of data deletion is (see Figure 2). Also, the taxonomy serves as a foundation on which the data engineering community can build. It shows research directions and helps to create a uniform understanding of data deletion. We believe that the taxonomy will support both researchers and practitioners. For researchers it serves as a research map, describing the topic areas and indicating the direction of efforts to further improve data deletion. Also, by structuring the subject area, future research projects can be better planned and constructed. For practitioners, the taxonomy can be used as a blueprint for developing real-world applications so that important aspects are not neglected. At this point it should be noted that the taxonomy was spit from the codes derived from the current literature. Therefore, it is possible that this taxonomy will change over further iterations.

5 CONCLUSION AND OUTLOOK

We believe that data deletion represents an important part of the data life cycle and data engineering field. Despite the increasing importance of this area through topics like GDPR, green IT, or data sovereignty, previous works indicate that data deletion is underrepresented in the literature (Tebernum et al., 2021) and in practice (Tebernum et al., 2023). To get a better overview of the subject area, data deletion was explored through a SLR. We discovered that the topic of data deletion is indeed poorly represented in literature. In fact, with only 121 publications in 23 years that deal with data deletion at their very core, the topic has been woefully neglected. The few publications in existence were primarily about the actual methods of deletion or security considerations. Work that takes an overarching view of the deletion process, structures it, standardizes it, or takes human actors into account is rare or even non-existent.

The analysis of the literature was also used to fundamentally structure the subject area for the first time. Such fundamental structuring is required by every domain in order for the community to describe and understand the research object in a fundamental way in the first place. Only with this foundation, further work can be motivated. To the best of our knowledge such work does not exists and this paper addresses the issue by providing a data deletion taxonomy. We believe that this novel contribution can help researchers and practitioners to look at the end of the data life cycle from a more global perspective. It allows to identify aspects of data deletion that may not have been considered before. In addition, the taxonomy represents a contribution to the data engineering community to improve communication among each other. A limitation that should not go unmentioned here is that the taxonomy to a large extent reflects the current state of the literature. In the future, we will need to evolve the taxonomy so that it evidentially reflects the very nature of data deletion.

Finally, we want to give an outlook on possible future research topics. As the evaluations of the axial and selective codes show, there are subject areas that are rarely addressed or not addressed at all. This often concerns topics that do not represent the deletion method or process itself, but address the surrounding area. E.g., there is little work on data deletion responsible (*who*). Future work should explore how to consider the human factor in deleting data. Also, many other questions arise regarding the identification and addressing of the data to be deleted (*what*). There should be a larger contribution on how to identify data to be deleted in an automated and secure way.

ACKNOWLEDGMENTS

The authors acknowledge the financial support by the Federal Ministry for Economic Affairs and Climate Action of Germany in the project Smart Design and Construction (project number 01MK20016F).

REFERENCES

Amadori, A., Altendeitering, M., and Otto, B. (2020). Challenges of data management in industry 4.0: A single case study of the material retrieval process. In *International Conference on Business Information Systems*, pages 379–390. Springer.

Azkan, C., Iggena, L., Möller, F., and Otto, B. (2021). Towards design principles for data-driven services in industrial environments.

Cantrell, G. and Through, J. R. (2019). The five levels of data destruction: A paradigm for introducing data recovery in a computer science course. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 133–138. IEEE.

Chen, N. and Chen, B. (2022). Duplicates also matter! towards secure deletion on flash-based storage media by removing duplicates. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, pages 54–66.

Chen, S.-H., Yang, M.-C., Chang, Y.-H., and Wu, C.-F. (2019). Enabling file-oriented fast secure deletion on shingled magnetic recording drives. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE.

Chiba, T., Sei, Y., Tahara, Y., and Ohsuga, A. (2020). A defense method against poisoning attacks on iot machine learning using poisonous data. In *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 100–107. IEEE.

Corbin, J. M. and Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21.

Council of the EU, G. S. o. t. C. (2022). Energy crisis: Three eu-coordinated measures to cut down bills.

Del Tedesco, F. and Sands, D. (2009). A user model for information erasure. *arXiv preprint arXiv:0910.4056*.

Diesburg, S., Feldhaus, C. A., Fardan, M. A., Schlicht, J., and Ploof, N. (2016). Is your data gone? measuring user perceptions of deletion. In *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*, pages 47–59.

Ehrlinger, L., Schrott, J., Melichar, M., Kirchmayr, N., and Wöß, W. (2021). Data catalogs: A systematic literature review and guidelines to implementation. In *International Conference on Database and Expert Systems Applications*, pages 148–158. Springer.

European Commission (2016). 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 (General Data Protection Regulation) (Text with EEA relevance).

Florescu, C. and Caragea, C. (2017). PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Vancouver, Canada. Association for Computational Linguistics.

Gao, B., Chen, B., Jia, S., and Xia, L. (2019). ehifs: An efficient history independent file system. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, pages 573–585.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. (2019). Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Glaser, B. G., Strauss, A. L., and Strutzel, E. (1968). The discovery of grounded theory; strategies for qualitative research. *Nursing research*, 17(4):364.

Graves, L., Nagisetty, V., and Ganesh, V. (2021). Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. (2019). Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*.

Klonowski, M., Struminski, T., and Sulkowska, M. (2019). Universal encoding for provably irreversible data erasing. In *ICETE (2)*, pages 137–148.

Kuperberg, M. (2020). Towards enabling deletion in append-only blockchains to support data growth management and gdpr compliance. In *2020 IEEE International Conference on Blockchain (Blockchain)*, pages 393–400. IEEE.

Lin, C.-W., Hong, T.-P., and Lu, W.-H. (2009). An efficient fusp-tree update algorithm for deleted data in customer sequences. In *2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC)*, pages 1491–1494. IEEE.

Long, S., Li, Z., Liu, Z., Deng, Q., Oh, S., and Komuro, N. (2020). A similarity clustering-based deduplication strategy in cloud storage systems. In *2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 35–43. IEEE.

Martin, T. and Jones, A. (2011). An evaluation of data erasing tools.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Murillo, A., Kramm, A., Schnorf, S., and De Luca, A. (2018). "if i press delete, it's gone"-user understanding of online data deletion and expiration. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 329–339.

Nguyen, Q. P., Low, B. K. H., and Jaillet, P. (2020). Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036.

Nguyen, Q. P., Oikawa, R., Divakaran, D. M., Chan, M. C., and Low, B. K. H. (2022). Markov chain monte carlo-based machine unlearning: Unlearning what needs to be forgotten. *arXiv preprint arXiv:2202.13585*.

Othon, M., Mile, S., de Melo, A., Junior, D. A., and Arruda, A. (2019). Evaluation of the removal of anomalies in data collected by sensors.

Otto, B. (2015). Quality and value of the data resource in large enterprises. *Information Systems Management*, 32(3):234–251.

Pachpor, N. N. and Prasad, P. S. (2018). Improving the performance of system in cloud by using selective deduplication. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 314–318. IEEE.

Pecherle, G., Győrödi, C., Győrödi, R., Andronic, B., and Ignat, I. (2011). New method of detection and wiping of sensitive information. In *2011 IEEE 7th International Conference on Intelligent Computer Communication and Processing*, pages 145–148. IEEE.

Rashid, F., Miri, A., and Woungang, I. (2012). A secure data deduplication framework for cloud environments. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*, pages 81–87. IEEE.

Reardon, J., Capkun, S., and Basin, D. (2012). Data node encrypted file system: Efficient secure deletion for flash memory. In *21st USENIX Security Symposium (USENIX Security 12)*, pages 333–348.

Riduan, N. H. A., Foozy, C. F. M., Hamid, I. R. A., Shamala, P., and Othman, N. F. (2021). Data wiping tool: Byteeditor technique. In *2021 3rd International Cyber Resilience Conference (CRC)*, pages 1–6. IEEE.

Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1(1-20):10–1002.

Sahri, M., Abdulah, S. N. H. S., Senan, M. F. E. M., Yusof, N. A., Abidin, N. Z. B. Z., Azam, N. S. B. S., and Ariffin, T. J. B. T. (2018). The efficiency of wiping tools in media sanitization. In *2018 Cyber Resilience Conference (CRC)*, pages 1–4. IEEE.

Sarkar, S., Banatre, J.-P., Rilling, L., and Morin, C. (2018). Towards enforcement of the eu gdpr: enabling data erasure. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 222–229. IEEE.

Steinbuss, S., Eitel, A., Jung, C., Brandstädter, R., Hosseinzadeh, A., Bader, S., Kühnle, C., Birnstill, P., Brost, G., Gall, Bruckner, F., Weißenberg, N., and Korth, B. (2021). Usage control in the international data spaces.

Subha, S. (2009). An algorithm for secure deletion in flash memories. In *2009 2nd IEEE International Conference on Computer Science and Information Technology*, pages 260–262. IEEE.

Tallon, P. P., Ramirez, R. V., and Short, J. E. (2013). The information artifact in it governance: toward a theory of information governance. *Journal of Management Information Systems*, 30(3):141–178.

Tebernum, D., Altendeitering, M., and Howar, F. (2021). DERM: A reference model for data engineering. In Quix, C., Hammoudi, S., and van der Aalst, W. M. P., editors, *Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021, Online Streaming, July 6-8, 2021*, pages 165–175. SCITEPRESS.

Tebernum, D., Altendeitering, M., and Howar, F. (2023). A survey-based evaluation of the data engineering maturity in practice. In press: INSTICC Springer https://www.researchgate.net/publication/367309981_A_Survey-based_Evaluation_of_the_Data_Engineering_Maturity_in_Practice.

Thushara, M., Mownika, T., and Mangamuru, R. (2019). A comparative study on different keyword extraction algorithms. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 969–973. IEEE.

Trajanov, D., Zdraveski, V., Stojanov, R., and Kocarev, L. (2018). Dark data in internet of things (iot): challenges and opportunities. In *7th Small Systems Simulation Symposium*, pages 1–8.

Van Bussel, G.-J. and Smit, N. (2014). Building a green archiving model: archival retention levels, information value chain and green computing. In *Proceedings of the 8th European Conference on IS Management and Evaluation. ECIME*, pages 271–277.

Wang, G. and Zhao, Y. (2008). A fast algorithm for data erasure. In *2008 IEEE International Conference on Intelligence and Security Informatics*, pages 254–256. IEEE.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33.

Wei, M., Grupp, L., Spada, F. E., and Swanson, S. (2011). Reliably erasing data from {Flash-Based} solid state drives. In *9th USENIX Conference on File and Storage Technologies (FAST 11)*.

Xu, X., Gong, P., and Xu, J. (2014). Data folding: A new data soft destruction algorithm. In *2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6. IEEE.

Žulj, S., Delija, D., and Sirovatka, G. (2020). Analysis of secure data deletion and recovery with common digital forensic tools and procedures. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1607–1610. IEEE.