Forensic Science International: Digital Investigation 45 (2023) 301562

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS 2023 USA - Proceedings of the Twenty Third Annual DFRWS Conference

Sharing datasets for digital forensic: A novel taxonomy and legal concerns

Frank Breitinger [a, *], Alexandre Jotterand [b]

[a] School of Criminal Justice, University of Lausanne, 1015, Lausanne, Switzerland
[b] id est avocats, 1002, Lausanne, Switzerland

ARTICLE INFO

Article history:

Keywords:
Datasets
Digital corpora
Taxonomy
Types of data
Legal concerns
Sharing data
GDPR

ABSTRACT

During the last few years, there have been numerous changes concerning datasets for digital forensics like the development of data generation frameworks or the newly released CFReDS website by NIST. In addition, it becomes mandatory (e.g., by funding agencies) to share datasets and publish them in a manner that they can be found and processed. The core of this article is a novel taxonomy that should be used to structure the data commonly used in the domain, complementing the existing methods. Based on the taxonomy, we discuss that it is not always necessary to release the dataset, e.g., in the case of random data. In addition, we address the legal aspects of sharing data. Lastly, as a minor contribution, we provide a separation of the terms structured, semi-structured, and unstructured data where there is currently no consent in the community.

1. Introduction

Digital forensics, like many other domains, relies on data that is available for research purposes and allows the reproducibility/comparison of results (Garfinkel et al., 2009). Besides research, Horsman and Lyle (2021) identified two additional areas where datasets are utilized: training and education as well as tool/process evaluation. While a research study by Grajeda et al. (2017) showed that data is frequently not shared, nowadays it is often mandatory to release the underlying dataset, e.g., funding agencies such as the European Commission, Directorate-General for Research and Innovation (2016), or the Swiss National Science Foundation (2017) require it (there are exceptions for some datasets). Consequently, researchers must develop a data management plan (DMP) that follows the FAIR principles presented by Wilkinson et al. (2016) ensuring that datasets are Findable, Accessible, Interoperable, and Reusable, by humans and machines. In summary, these principles require that datasets can be found in common repositories and have a persistent identifier. It should be possible to download them via standardized protocols, and they should include descriptive information (metadata) that allows reusing it (how the dataset was created/collected, system specifications, the meaning of variables, etc.). Consequently, several repositories listing forensic datasets exist which allow authors to post their own sets and researchers to explore (presented in Sec. 3.1). However, these repositories use different terms and terminology hampering the searches.

In this article, we focus on two research questions that are related to the FAIR principles:

RQ1 How can datasets be described to ensure findability?
RQ2 Under what circumstances can datasets be shared?

We argue that findability goes hand in hand with common terminology, taxonomies, and classification schemes to organize datasets. These should be used by researchers when releasing their datasets, and within repositories to organize data/to enable searches. Existing repositories primarily focus on the content of the data where this article recommends two additional parameters: *organization* and *origin*.[1] Depending on the origin (e.g., real-world data), restrictions may apply which prohibit sharing. Researchers need to know these restrictions. Consequently, this article provides the following contributions:

* Corresponding author.
E-mail addresses: frank.breitinger@unil.ch (F. Breitinger), alexandre.jotterand@idest.pro (A. Jotterand).

[1] Note, in this context the term origin was first used by Grajeda et al. (2017).

https://doi.org/10.1016/j.fsidi.2023.301562

- We summarize terms frequently used to describe datasets to establish a common understanding. We especially describe the differences between structured, semi-structured, and unstructured data in Sec. 2 (RQ1) which we define as the *organization of data.*
- We propose a new taxonomy in Sec. 4 which complements existing taxonomies (e.g., CFReDS). A peculiarity is that it focuses on how the data was created (its origin) which impacts how the data can be shared (RQ1+2).
- We outline existing legal restrictions that one must consider before sharing data where our emphasis is on European law, i.e., we focus on personal data (GDPR) and copyright/licenses (RQ2, Sec. 5).

The remaining sections present the Background and related work, a comprehensive discussion in Sec. 6, and ends with a conclusion.

2. Terminology

The most basic differentiation of data is what we name the organization of data and falls into three categories: structured, unstructured data, and semi-structured. However, there is no clear separation between these terms which we address in Sec. 2.1. Subsequently, we summarize Ground truth data and Metadata which are terms that are also frequently used within digital forensics. The last subsection discusses the sensitivity of data, which is data that must be protected, e.g., due to legal requirements.

2.1. Organization of data

Depending on the source, one differentiates between structured, unstructured, and semi-structured data whereas some sources ignore semi-structured and consider it as structured or unstructured data. Consequently, the differentiation between the three is floating and not defined. As an example, Rizkallah (2017) and Taylor (2021) list Email and video as unstructured data. In contrast, a reviewer of this article argued that Email header fields (from, to, cc, subject, etc.) can be queried and therefore are not unstructured data. The same reviewer disputes that CCTV footage is structured as the files have names indicating format, camera ID, and time of recording intervals. This section discusses the three categories and defines criteria to distinguish between them.

2.1.1. Structured data

This is any data that conforms to a predefined structure, schema, or type (Arasu and Garcia-Molina, 2003), i.e., the data is organized and follows a specification. A common example is data that can be placed in a relational database (Abiteboul, 1997) but also data that has been defined through protocols such as the TCP/IP header information (only the TCP/IP headers but not the content!). Structured data can be separated into fields that can be accessed separately (bigdataframework.org, 2019) and the content already has a context. For instance, querying a database table for employee names will return short strings representing names. Structured data is generated by humans and machines and can be accessed by both entities using defined queries. Depending on the data, queries may allow us to gain new insights, e.g., by combining data or 'joining' it with other information (database perspective). Commonly, a user is only interested in the content (data); the container format (database, tables, etc.) is irrelevant (it only defines how to access the data, e.g., via SQL queries, MS Excel, etc.).

Structured data is often converted into unstructured or semi-structured data, e.g., personal information exported to different formats for easier consumption. Once in its new format, it is hard to track. Note, while the process from structured to unstructured is easy, the inverse is difficult (one-way function).

2.1.2. Unstructured data

This describes data that is *not* organized and does *not* follow predefined models. The data is often unsearchable (besides basic string queries) and difficult to analyze. Note, unstructured data still follows internal structures of the underlying file type/format, but the data (content) cannot be converted into table format (MongoDB.com, 2020). According to Rizkallah (2017) and Taylor (2021), about 80% (some other sources even say 90%) of today's data is unstructured. Many sources provide Email, PDF, audio, video, and social media postings as examples of unstructured data (e.g., Taylor (2021)). While mostly human-generated, it may also be generated by machines, e.g., digital surveillance or satellite imagery[2] (MongoDB.com, 2020). Unstructured data is especially problematic for privacy as filtering, deleting, or anonymizing information is complex, e.g., identifying a photo of a passport in a large gallery. To tackle this challenge, artificial intelligence is used to help process vast amounts of data. Unstructured data is often accompanied by structured data or semi-structured data to ease basic queries. For instance, pictures may include EXIF information, or their filename reflects the date and time the picture was taken allowing us to organize them in a timeline.

2.1.3. Semi-structured data

This type has certain organizational characteristics that facilitate its processing such as tags or other markers to organize data which may be included within the data itself (self-describing data (Buneman, 1997)). Another feature is its flexibility: despite the structure following a defined specification, it has optional fields and no predefined min/max length. Common data structures and formats associated with semi-structured data are trees (Buneman, 1997), JSON, XML, or NoSQL (Taylor, 2021). Compared to the other two, it is fairly new but becoming more important as it is frequently used within the Internet of Things and the Web.

2.1.4. Distinguishing characteristics

The differentiation between types is often vague, e.g., some say Email is unstructured data while others say it is semi-structured data. Marr (2019) points out that emails can be seen as both. While the content is unstructured, the header includes structured information. Another example the author provides is pictures that include metadata such as GPS coordinates, date/time, or device ID (structured). When stored, a user may also add tags manually, e.g., vacation or birthday party. On the other hand, the content is unstructured. Marr concludes that both are semi-structured as they are a mixture of both types. We argue that

- the category is independent of the file type or format but depends solely on the underlying data. Example: an Excel sheet may contain structured, unstructured, semi-structured, or no data. However, the file type or format may help to decide what data it is.
- there are different perspectives on the data, i.e., why was the dataset created/what may it be used for. Example: CCTV footage may be considered unstructured if one is interested in the content of the footage, but its metadata may be considered structured (or semi-structured) and allows finding all footage within a date range.

To differentiate, we define the following criteria:

[2] Examples include weather data, landforms, and military movements.

1. Is the data under consideration in a (relational) table or can be converted into one?

   If yes, then it is structured data.

2. If no, is the data under consideration in a markup language format (e.g., JSON, XML), or can be converted into one?

   If yes, then it is semi-structured data.

3. If not, it is unstructured data

Consequently, to differentiate between semi-structured and unstructured, it must be possible to have a subset of the data that can be converted into structured data. An example is provided in Appendix A.

2.2. Ground truth data

Depending on the domain, the meaning of ground truth data varies. In digital forensics, the term describes data that is well understood by the community which could be due to three reasons:

1. The data is completely fabricated which means software or a human created it and documented the details. As it has been created, a researcher knows the underlying data and there are no inconsistencies or surprises.
2. The data is gathered from sources and then carefully analyzed. The outcome of the analysis can be either a description, a labeling, or a clustering/sorting of the data. For instance, if the goal is to create a dataset of rhino pictures, a script may be utilized to crawl the web and download these images. A manual assessment is necessary to ensure that there are only rhinos, a description may be added if the dataset also includes albino rhinos or animated rhinos, etc. Depending on the size of the dataset, the assessment can be complex, time-consuming, or even impossible.
3. There is no knowledge about the creation, but the dataset has been thoroughly explored by one or more researchers. An example would be the real-world dataset Enron (Klimt and Yang, 2004) which consists of real-world data but has been used (explored) in various research studies.

As pointed out by Roussev (2011), ground truth data allows controlled studies while "it is infeasible to establish the ground truth on any set of non-trivial size".

2.3. Metadata

Metadata is data about data and has different meanings depending on the context:

Metadata within a digital forensics context summarizes additional information that is available for analysis besides the content data. Buchholz and Spafford (2004) defined it as "all the data in the file system that describes the layout and attributes of the regular files and directories [.] such as timestamps, access control information, file size, but also information on how to locate and assemble a file or directory in the file system". Today, it is more common to separate metadata into two groups: *internal and external* (Berryhill, 2019). Internal metadata is stored within the digital object itself and thus varies by file type. Examples are EXIF information in JPG, or the author attribute in PDFs and Word documents. External metadata, on the other hand, describes the information that is found in other data structures. Examples are timestamps or access rights which are part of the file system.

Metadata within data(set) context is information describing and accompanying data or a dataset. According to the Swiss National Science Foundation (2017), this includes how the data was collected or how it may be reused. With respect to digital forensics, it should also include detailed version numbers and utilized software, e.g., the underlying system was a Windows 10 version 22H2 and the extraction was done using ABCDEFG version 10.8.1a. Desirably, it includes a persistent identifier, creator name, and date.

2.4. Sensitivity of data from a research perspective

For many domains, sharing data means providing access to structured information, i.e., data stored in tables. According to Li et al. (2006), these tables consist of various columns (attributes) and rows where each row corresponds to one individual. Attributes can be separated into *explicit identifiers/direct identifiers* (DI), *quasi-identifiers/indirect identifiers* (II), and *sensitive attributes/data* (SD). When releasing data, SD is the element of interest and should not be linkable to individuals (DI, II). For instance, a medical database contains the name (DI), zip code, gender, birthday (all three are II), diseases, and medication (both SD). For research purposes, SD must be retained, DI is deleted, and II is modified to protect privacy, e.g., shuffling the zip codes, and replacing birthday with an age range such as 60–70 (these 'modification techniques' are summarized in Appendix B). Majeed and Lee (2020) introduced non-sensitive attributes such as eye color. This data is often not/less relevant (and not collected) but if it exists, it can be published as is (as long it cannot be linked to the data subject; see Sec. 5.4).

Concerning digital forensics, the nature of data is broader, and the community often shares **semi-structured** or **unstructured** data. We, therefore, propose the following (based on Garfinkel et al. (2009) and Majeed and Lee (2020)):

Direct identifiers (DI) can directly identify an individual such as name, social security number, email, or phone number. Usually, this data must be removed completely from the dataset.
Indirect identifiers (II) which, when linked with auxiliary information or querying additional sources, may reveal the identity such as age, gender, race, zip code, IP, license plate, patient ID, or credit card number.
Sensitive data (SD) describes the information that someone does not like to share, e.g., personal files, passwords, video footage, or chat messages. It is therefore broader than the concept of sensitive personal data (or special categories of data) in data protection laws (e.g., in Art. 9 GDPR).
Illegal & protected data (IPD) is a different form of sensitive data and describes data that should not (cannot) be shared. This includes intellectual property, child sexual abuse material (or pornography to minors), copyright concerns, or missing license agreements.[3]
Non-sensitive data (NSD) includes all other information and can be published as is, e.g., operating system log files, smartphone model, freeware installed, or the number of received spam emails per day.

DIs are always personal data. In contrast, the others categories will only be deemed personal data if they can be linked to an individual (see Sec. 5.4), e.g., with the help of contextual information available in the environment in which the data is used (Jotterand, 2022).

While external metadata can be mostly seen in NSD, internal

[3] For instance, a bootable Windows 10 image with an individual not possessing a valid license.

metadata (depending on the data) may be considered DI or II and thus have to be used with caution.

In most countries, research involving human subjects (which includes the collection of personal data) requires the approval of an Institutional Review Board (IRB) or an Ethics Commission. As regulations and requirements for obtaining approval can vary between countries, discussing this is beyond the scope of this article.

3. Background and related work

Before discussing the taxonomy, it is important to know how data is currently shared (Sec. 3.1), the idea of data generation frameworks (Sec. 3.2), and frequently used terms to describe the origin of data (Sec. 3.3).

3.1. Data repositories

The three most prominent platforms listing datasets (according to our opinion) are briefly summarized in this section. Other examples could be the digital forensics challenges from DFRWS or the Digital Forensics Tool Testing Images[4] from Brian Carrier.

3.1.1. Digital corpora

The digital corpora (https://digitalcorpora.org) is a website first presented by Garfinkel et al. (2009) and is a collection of various datasets for use in computer forensics education research. The website provides freely available datasets as well as real data where the use is possible under a special arrangement. The website separates the data into five major categories:

- Cell phone dumps
- Disk images
- Files
- Network packet dumps
- Scenarios

In addition, there is the possibility to directly search the S3 bucket where the datasets are hosted. However, the bucket organizes the data slightly differently and uses the following directories:

```
$ aws s3 ls s3://digitalcorpora/corpora/
                           PRE bin/
                           PRE drives/
                           PRE drives_bulk_extractor/
                           PRE drives_dfxml/
                           PRE files/
                           PRE hashes/
                           PRE mobile/
                           PRE packets/
                           PRE ram/
                           PRE scenarios/
                           PRE sql/
```

A special characteristic of the digital corpora is that most datasets come with comprehensive descriptions, i.e., metadata. For instance, the 2019-owl scenario comes with a short scenario description as well as Documentation from Creation of Owl Scenario.zip which includes several project logs. Unfortunately, the metadata is not consistent in terms of format or content between the available datasets. It also does not seem to be indexed by the search functionality, e.g., the logs contain "Skype" while a search for Skype fails.

3.1.2. Datasets for cyber forensics

This repository was presented by Grajeda et al. (2017) and is available at datasets.fbreitinger.de. Compared to the digital corpora where the authors release their own datasets, this website is a collection of datasets: the authors reviewed numerous articles and performed online searches. As a result, this website includes datasets from various sources including the digital corpora.

The core of the website is a table with 7 columns (dataset type, available datasets, total size, origin, source, date, and more info) and currently includes 82 entries. The dataset type is a short description of the set and one of the following categories: APK, Apple iPod disk image, chat logs, database, different types of files, computer malware, email datasets, hard disk images, leaked passwords, media (pictures or video), mobile malware, network traffic, ram dumps, SD card images, smartphone images, SIM images, USB images, Video Game Console images, WiFi Traffic, and Text. The origin column describes how the dataset was created and can be user-generated, experiment-generated, or computer-generated (details see Sec. 3.3).

3.1.3. Computer forensic reference DataSet portal (CFReDS)

The CFReDS project by NIST is the third repository and includes data produced by NIST, often from the CFTT (computer forensic tool testing) project, as well as datasets contributed by other organizations and individuals. CFReDS v1.0 was similar to the digital corpora and included a list of various scenarios/datasets each linking to a new website that includes detailed information about the dataset.

In 2021, CFReDS v2.0 was rolled out which included a major update of the website layout as well as the included datasets. To the best of our knowledge, this is the most comprehensive collection and includes listings from the digital corpora (Sec. 3.1.1) and the datasets for cyber forensics (DCF, Sec. 3.1.2) repository. The new version is more like the DCF repo and includes a searchable table with multiple columns, with the 'Title' and 'Tag' columns in particular making it easier for the end-user to search for data. The title can be freely chosen and is similar to the dataset type in DCF. Examples are 'basic memory images' or 'Media samples 3 (pictures)' but may also describe the scenario like the 'Russian Tea Room'.[5] The tag-column allows filtering for datasets based on the CFReDS taxonomy where the first two levels look like the following:

Data/Forensic related: Databases; Date, Time & Place Analysis; Email Search; Evidence Collection & Integrity Management; File Recovery; Internet; Multimedia; Social Media & Messaging; String Searching; File type
IT System Type: File system; Other Devices & Systems; PC & Operating System; Phone, Mobile & Tablet
Simulated Cases/Scenarios: Data Leakage; Hacker Case; M57 … (there are several other cases)

Most entries have additional subcategories, e.g., Internet includes Browser, Cryptocurrencies, Peer To Peer File Sharing, Search History, and Telecommunications. The granularity goes down to the file type, e.g., tags such as MP3, ZIP, or JPEG are possible.

[4] https://dftt.sourceforge.net (last accessed 2023-06-04).

[5] These are often terms that have established themselves in the community as the scenarios have been around for several years. Other examples are the M57, the National Gallery DC Scenario, and Rhino hunt.

3.2. Data generation frameworks

Data generation frameworks, a.k.a. synthesis frameworks, are complex software tools that allow the creation of digital forensic datasets. Depending on the framework or software, some create complex cases (scenarios) while others focus on one particular aspect. One of the first frameworks is Forensig2 by Moch and Freiling (2009, 2011) which creates disk images. It can be configured to simulate user behavior such as copying or deleting files. Over the years, other frameworks have been released such as the computer forensic test image generator (ForGe) by Visti et al. (2012), EviPlant by Scanlon et al. (2017), and TraceGen by Du et al. (2021). The most recent framework is ForTrace by Göbel et al. (2022) which can be seen as a multi-layer framework that can generate disk images, memory dumps, and network traffic. Göbel et al. also provides a good overview of existing frameworks.

Compared with manually created data, these frameworks have two advantages. First, they can be used for mass generation (e.g., creating 30 images at once), and second, they log their activity which represents the ground truth. For instance, Moch and Freiling (2009) state that "file system image that can be analysed by students within exercises on forensic computing. The analysis results of the students can then be compared with the 'truth' encoded in the input script".

3.3. Random, synthetic, experiment generated, and real-world data

Garfinkel et al. (2009) differentiate between categories of data, where they proposed using sensitivity as the deciding factor. Their taxonomy includes the following five categories: (1) Test data, (2) sampled data, (3) realistic data, (4) real and restricted data, and (5) real but unrestricted data. 'Test data' and 'realistic data' are artificially constructed sets where the latter is more complex and similar to what practitioners may encounter. 'Sampled' is a subset extracted from a larger source, e.g., randomly downloaded data. The last two categories are "created by actual human beings during activities that were not performed for the purpose of creating forensic test data". As these sets may contain personal data or protected content, access may or may not be restricted (more details in Sec. 5).

In contrast, Yannikos et al. (2014) differentiate between real-world data, such as the Enron dataset, and synthetic data. In the latter case, the authors separate between 'manually reproducing real-world actions' and 'tool-supported synthetic data corpus generation'.

Grajeda et al. (2017) researched the availability of datasets for digital forensics and categorized available data into the following three categories: (1) *Experiment-generated* which describes data created by humans through scenarios and experiments; (2) *user-generated* data which refers to real-world data; and (3) *computer-generated* data where the data has been created through algorithms or/dev/urandom. However, the authors remain vague when separating between (1) and (2). For instance, the authors list mobile device images that were created over 3 months as 'user-generated data' which could also be experiment-generated data.

4. Forensic dataset taxonomy

This section presents a novel taxonomy that can be used to classify data. Instead of focusing on the content/type of the data (e.g., HDD images or network traffic), it focuses on how the data was created and complements taxonomies such as the one proposed by CFReDS.

4.1. Methodology

To create an appropriate taxonomy, we followed a three-step procedure:

1. We first read related literature to identify terms that have been frequently used in literature. The references have been discussed in Sec. 3.3.
2. Next, we reviewed the three existing repositories described in Sec. 3.1 to see what data has been released and how it was created where we again focused on the used terminology.
3. Lastly, we consulted dictionaries and blogs to find terms used in other domains.

Findings were then combined with the aim to stick as close as possible to the existing terminology but also provide a clear separation between the various datasets.

4.2. Terms and definitions

Within our taxonomy, we will use certain terms which we think require a discussion beforehand as they are used within digital forensics as well as other disciplines with slightly different meanings. We aim to utilize terms that are already established within the community but are also widely used. After consulting numerous blogs, websites, and articles, we made the following choices:

*Artificial data vs. Synthetic data.*Artificial data (AD) is any data that is created by software and can be seen as the most general term. Synthetic data, on the other hand, is often seen as a subset of AD that aims to mimic real data. Some sources differentiate between artificial data and synthetic data while other do not. In digital forensics, the term synthetic data seems more common and thus will be used by us. A more thorough discussion and examples are in Appendix C.

*Subcategories of synthetic data.*Many terms are used in literature to describe subcategories of synthetic data: fully synthetic, partially synthetic (or hybrid), simulated, generated, rule-based data generation, and several others. While some of these terms can be used interchangeably, some differ in nuances. Out of the terms we chose, especially two require an additional discussion:

> **Rule-based data generation** (a.k.a. generation) describes the process of generating data following strict rules and thus the data-generation process is often deterministic (i.e., outputs will have identical hashes). One may think of it as a software (function) accepting one or more arguments and returning the data sample. An example would be test data for a file carver. Input arguments are an empty disk image formatted in FAT32 and several JPGs. The software cuts the JPGs into pieces, places them somewhere in the image (it may manipulate the FAT table or bitmap), and returns the disk image.
> **Simulation**, in contrast, is often non-deterministic, and thus the ground truth is vaguer. Instead of following rules, these simulators rely on other functionality that they call in sequence. In the case of the file carver data, a simulator mounts the disk image, uses the copy command to place the data on the disk, and then deletes (`rm`) files. The process may be repeated several times to ensure fragmentation.

In digital forensics, simulators seem to be more common. However, making this differentiation is important as it impacts the ground truth. For the former, the ground truth is a very detailed list where the pieces of each image can be found (down to the offsets). We know the exact content of the FAT/Bitmap. This granularity may be required to test string matching, carvers, or parsers. On the other

hand, *simulation* provides a less detailed ground truth. One knows that X images should be found but we do not know if they are fragmented and where the fragments reside. The process is also often not deterministic, i.e., creating two images will likely be different.

Deterministic.When deciding if the data generation is deterministic or not, it is important to first decide what the relevant data is. As an example, let us assume zip file carver. A function accepts files and a disk image, zips the files, splits them, and places them inside the image. Given that zipping is non-deterministic (hashing the same file twice produces different hashes), the resulting images will be different. However, from the carver's perspective, they will be identical and thus should be seen as ruled-based data generation. A similar scenario is network traffic where identical requests are sent but timestamps differ.

4.3. Proposed taxonomy

Given the previous classifications, we propose the taxonomy depicted in Fig. 1 where data is either synthetically (artificially) generated or created by humans. More detail may be added in the future, e.g., when we better understand what AI can contribute. Detailed descriptions are provided in the following:

Synthetic data describes any data that is created by software with a certain degree of autonomy, i.e., a user may adjust settings or may define a playbook, but the heavy lifting is done by software. Ideally, the software documents the data (ground truth) in some machine-readable format which can then be used for evaluation purposes and is developed so that it supports pseudo-random behavior. Sharing the dataset may not be necessary to reproduce results if the software and settings are released.

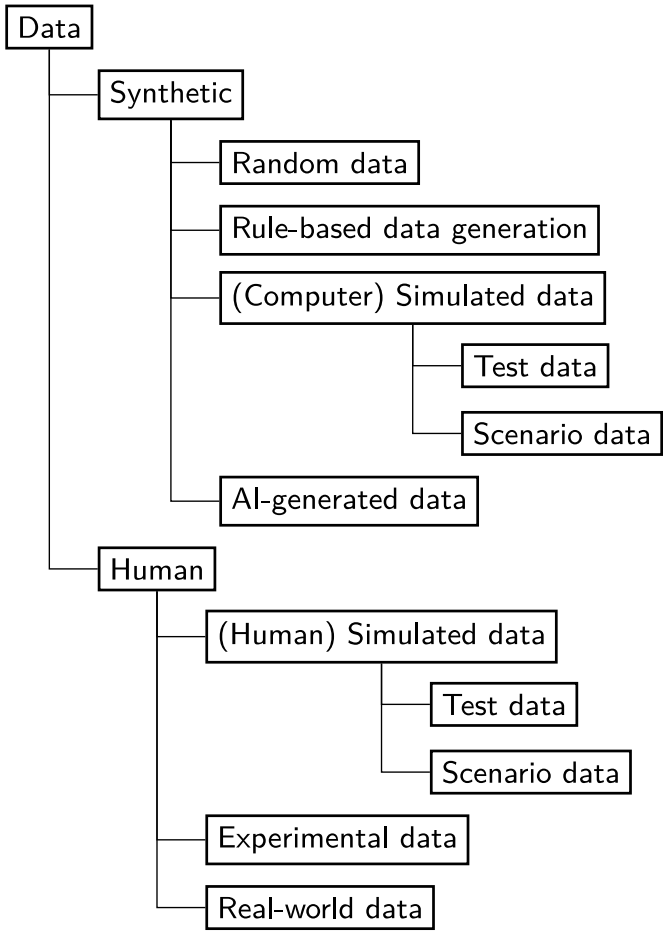

Fig. 1. Visual representation of the proposed taxonomy.

Random data is a special case of synthetic data originating from a (pseudo-)random source[6] where the generation may or may not be deterministic (e.g., through the utilization of a seed). The actual data is often not relevant but more the fact that it is unique. For instance, one may place a random sequence in a memory dump to validate a parser. Random sequences are also used to mark the beginning and end of something or to do some initial assessment. Researchers are more interested in finding the offset/location wherefore the data has to be unique but the content is irrelevant for reproducibility. It is not necessary to keep/provide the data but describe the methods of how it was created or provide the algorithm and size/length. Consequently, it is usually not shared.

Rule-based data generation, as described in the previous section, is mostly deterministic and allows a very detailed description of the ground truth which makes it ideal for forensic tool testing. On the other hand, it is impractical for complex datasets such as full scenarios. It must be decided whether it is worth the development or whether the dataset should be created manually (human).

(Computer) Simulated data follows a process and relies on existing functionality which makes it practical for any complexity of data generation. It is often not deterministic due to some (pseudo-)random behavior of the utilized tools, e.g., the allocation mechanism of the file system, varying timestamps, or security features (e.g., random initialization vectors). Ground truth (metadata) is more general but sufficient depending on the use case. An example would be event reconstruction where the ground truth does not need to contain the events down to the nanosecond.

> Test vs. scenario: Simulators may only produce specific data (e.g., a network capture) or complex scenarios comprising several different artifacts (e.g., network capture, memory dump, and disk images). Consequently, if only one type (format) is returned, it is test data and otherwise scenario.

AI-generated data can be deterministic or non-deterministic, depending on how they are designed and trained. As of now, they are not common in the digital forensics community, but we likely see them soon. Examples would be AI systems that return text (e.g., ChatGPT OpenAI (b)) or images (e.g., DALL·E 2 OpenAI (a)).

Human driven dataset creation describes data that is the result of one or more humans interacting with a system. This data is unique, and datasets should be shared (if there are no legal concerns) to ensure the reproducibility of results and to allow comparisons. Metadata is created manually, and the granularity depends on the dataset author(s).

(Human) Simulated data is the equivalent of synthetic simulated data and summarizes datasets that are created by one or more researchers to test or validate the functionality of the software. Currently, this is the most common category found. *Test data examples* range from network captures of one HTTP connection (PCAP file) to complex sets such as the corpora of 77 SQLite files where "every single database file was specifically crafted and has at least one peculiarity" (Nemetz et al., 2018). *Scenario data,* on the other hand, has a higher complexity, is generated over a longer time frame, or is based on real scenarios (as it tries to mimic them). The result is not a single category of files but often disk images (containing all sorts of data) or full scenarios.

[6] Examples to generate this data include a random bitstream from/dev/urandom or a program that creates text files with random text, i.e., a text of random letters given an alphabet.

Experimental data requires a group of actors to produce the data and is orchestrated by one individual/a small group. Participants are informed upfront about the purpose of their activity and the dataset generation. It is recommended that they sign an agreement. An example is the dataset provided by Guido et al. (2016) who collected smartphone user data from 34 participants over the duration of three months. In the case of a larger experiment, all participants should provide documentation that may be comprehensive.

Real-world data follows Garfinkel et al. (2009) description and refers to data created by humans with no intention to create a forensic dataset. Examples are the Enron dataset, malware samples, or Fraud detection dataset[7] but also data sold on the dark web. This data may include sensitive or even illegal content and should be used with caution. Garfinkel et al. discuss challenges with real-world data in Sec. 2.4 and 3.2. Metadata cannot be provided as the set is unknown. The more researchers use the corpora, the better it will be understood.

4.4. Application

Our proposed taxonomy complements existing ones such as CFReDS and should be used within the metadata to describe the underlying dataset. An example in JSON format could be as follows:

```
{ "description": {
    "organization" : "unstructured",
    "origin" : "Real-world data",
    "content-tags" : ["memory", "windows xp", "image"]
   }
  ....
}
```

While sometimes information may be redundant, there are cases where it is not and it will allow researchers to find the most appropriate dataset: Email could be real-world such as Enron, as well as human or synthetically generated. The proposed terminology also allows researchers to describe their experiment, e.g., our picture classification algorithm was tested based on 10′000 AI-generated images.

5. Legal barriers to sharing data

Data sharing in the field of digital forensics can be hindered by legal barriers. This section identifies the main legal concepts that may restrict data sharing, focusing on privacy and copyright.

5.1. Data ownership, control, and sharing restrictions

Before sharing data, it is important to understand if and how it is protected and who owns (or can exercise rights over) it. However, there is no harmonized concept of data ownership worldwide or in Europe (European Commission and Directorate General for Communications Networks, Content and Technology, 2016). While at the semantic level, the information to which data relates can be protected by a special law, such as intellectual property law or trade secrets, no such protection is universally granted to the data itself. As a result, data can generally be freely used and shared by the person who has effective control over it. For example, data which is generated by a researcher as part of an experiment is factually controlled by the researcher, who can in principle freely decide if and how to use it. There are, however, three main categories of legal barriers that may impede this freedom: (1) as mentioned above, the data at its semantic level may be protected by copyright or another special law (such as patent, databases *sui generis* rights, or trade secrets); (2) even if the data is not protected by a special law, contractual provisions may restrict how the data can be shared; (3) lastly, privacy or data protection laws may impose rules and restrictions on the sharing of personal data.

5.2. Special law: copyright and its limitations

Copyright and other intellectual property rights grant their holders certain prerogatvies. While it is not the only special law to do so (e.g., trade secrets or the EU database directive also grant some rights), it is the most relevant legal concept in the present context. We will therefore focus on copyright hereafter.

Copyright law grants the copyright holder moral and exclusive property rights for a relatively long period (generally at least 50 years after the death of the author), such as the right to decide if and how their work may be communicated to the public, reproduced, adapted, or modified (Berne Convention, 1979). As a result, copyrighted content can generally only be shared with the permission of the copyright holder, unless an exception applies. However, copyright does not apply to all data but only to 'original works of authorship', i.e., any human-made production in the literary, scientific, and artistic domain, whatever the mode or form of its expression (Art. 2(1) of the Berne Convention). This may include for example images, videos, and software code both in object and source code. Although each national law defines the threshold of originality required to benefit from copyright protection, it is generally understood to be low. In addition, national laws may protect content that has no individual character or originality at all, as is the case in Switzerland for photographs (Art. 3bis Swiss Federal Act on Copyright and Related Rights by the Federal Assembly of the Swiss Confederation (1992)). Since protection is automatically granted when the work is published, without the need for registration, it is not possible to check a public registry to ascertain if a dataset contains copyrighted elements.

Copyright is subject to a fundamental limitation (for the time being at least) that is highly relevant in the context of digital forensic: under most current national laws, a work created without sufficient human intervention will not be protected by copyright (United States Patent And Trademark Office (USPTO), 2020; Court of Justice of the European Union, 2009). Accordingly, works created by a machine without sufficient human intervention, such as AI-generated content, are generally deprived of copyright protection. In the context of digital forensics, this means that synthetic data will generally be deprived of copyright protection (and, conversely, free of copyright claims) but may still be subject to other restrictions such as licenses. Copyright will remain a potential barrier in the presence of human-driven datasets, and in particular real-world data (although human simulated data may potentially be subject to copyright protection, we assume that the copyright will be owned by the same researchers who wish to share the dataset, and who will therefore consent to share; we further assume that copyright will often not be relevant in the context of experimental data, or that consent will be easy to obtain). It follows from the above that while copyright law may apply to the sharing of forensic data, it will mainly be the case in presence of real-world data. As we will see later (Sec. 6.1) legal limitations to copyright, such as the U.S. 'fair use' doctrine, may in such cases allow researchers to share copyrighted content without having to obtain the consent of the copyright holder(s).

[7] This is a set containing anonymized credit card transactions labeled as fraudulent or genuine and can be found on Kaggle: https://www.kaggle.com/datasets/whenamancodes/fraud-detection.

5.3. Contractual restrictions

In presence of data that is not protected by copyright or another special law, the question arises whether the researcher may impose licensing terms restricting its use. In RyanAir vs PR Aviation, the Court of Justice of the European Union (2014) ruled that one may contractually impose restrictions on how data, which is not protected by a special law, may be used. In that case, however, the existence of a valid contract was not disputed (one had to click to accept the legal terms before accessing the content). To the extent the data is made accessible on a repository, which is publicly available, the question of the valid acceptance of a license or contract may arise, making it potentially difficult to impose contractual restrictions.

5.4. Personal data

Depending on the regulations and the domain, various terms have established themselves. In the EU the term personal data is used which is described in the following. Note, Personal Identifiable Information (PII) as described by (McCallister et al., 2010, NIST) is a US concept and similar to the European concept of personal data.

Within the privacy domain, a finer granularity is used such as direct (explicit) identifiers and indirect (quasi) identifiers are widely used (explained in Sec. 2.4).

Privacy laws/GDPR: Privacy laws grant rights to individuals which may restrict how information that relates to them may be used and shared. In the EU the GDPR (Council of the European Union and European Parliament, 2016), for instance, requires that any processing of personal data (including its use and sharing) is justified by a lawful basis such as consent, a legal obligation, or an overriding legitimate interest (Art. 6 GDPR).

Personal data, anonymous data, and pseudonymization: Personal data is information relating to an identified or identifiable natural person (Art. 4(1) GDPR). Whether data can be linked to an identified or identifiable natural person depends on the data itself and the environment in which it is shared (Elliot et al., 2020). This means that the same data may be anonymous information for one entity, but personal data for another if it holds additional information enabling it to identify the concerned individuals (Jotterand, 2022).

The concepts of data anonymization and data pseudonymization both refer to the action of processing personal data in such a manner that it can no longer be attributed to a specific data subject without the use of additional information. In the case of anonymization, the reference to the person is irreversibly removed, while in the case of pseudonymization, it is only reversibly removed (a key, or assignment rule, being retained to enable re-identification) (Castelluccia et al., 2022).

GDPR will only apply to personal data, i.e., it will not apply to anonymous information. This includes information that is technical by nature (e.g., weather data), information that is synthetically generated (Castelluccia et al., 2022), or personal data that is appropriately anonymized. Because the link to the individual is only reversibly removed, pseudonymized data will be deemed personal data and the GDPR will fully apply to it (Recital 26 of GDPR).

6. Discussion

The introduction raised the research question *"Under what circumstances can research data be shared?"* which is discussed below based on our taxonomy.

In general, data can be shared unless it contains personal data, is protected by a special law, or is subject to contractual restrictions. Consequently, any *synthetic data and human simulated data* generated by research can generally be shared without legal barriers, since it is unlikely to contain any such problematic content. In contrast, experimental data and real-world data are more likely to present risks of containing personal data or copyrighted content. Note, synthetic data can be shared but it may not always be necessary. If the data can be reproduced, sharing the software and settings to generate the data may be sufficient and efficient (e.g., less space).

6.1. Sharing restricted content (copyright and contractual protection)

Copyright protection: As stated above, copyrighted content (such as images or software code created by humans) may generally only be shared with the consent of the copyright owner(s), unless an exception applies. This consent may be obtained in a variety of ways, including through specific waivers granted by the copyright holder on a case-by-case basis, or through licenses granted by the copyright holder. Common examples of such licenses include the Creative Commons family of licenses[8] or the licenses developed by the Open Knowledge Foundation.[9] Researchers should review the terms of these licenses to determine whether they are permitted to reuse and share specific data.

Copyright limitations: The use and sharing of copyrighted materials may, however, be legally permitted in some circumstances without the need to obtain the consent of the right holder(s). Indeed, national laws on copyright place various limitations on the owner's exclusive right to determine how their work may be used. One such limitation is the U.S. 'fair use' doctrine under which a copyright holder may not prevent another person from making a 'fair use' of a copyrighted work, such as using the copyrighted work for research purposes (17 U.S. Copyright Act §107 U.S. Copyright Office (2022)).

The 'fair use' doctrine is interpreted based on four guiding factors outlined in the U.S. Copyright Act's fair use provision: the purpose and character of the use; the nature of the copyrighted work; the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and the effect of the use upon the potential market for or value of the copyrighted work. In a recent landmark case, the U.S. Supreme Court (2020) confirmed that the fair use doctrine applies to computer programs. In Europe, Directive 2001/29/EC of the European Parliament and the Council of the European Union (2001), in conjunction with Directive (EU) 2019/790 of the European Parliament and the Council of the European Union (2019), lists several optional or mandatory exceptions or limitations to some of the author's exclusive economic rights, including the use of the work for teaching, research, or private study. Determining whether such a limitation of the exclusive rights of the copyright owner may be asserted requires a case-by-case analysis. Accordingly, it may be permissible to share datasets that contain copyrighted material without having to obtain the consent of the right holder. This may require legal expertise, given the complexity of the issue, and researchers should be advised to evaluate such statutory limitations carefully before relying on them to share data.

Contractual restrictions: As we have seen, contractual restrictions that limit the use and sharing of datasets may apply even when the dataset or the data itself is not protected by a special law. Those contractual restrictions may stem from general terms and conditions that must be accepted prior to being allowed to access

[8] https://creativecommons.org/.
[9] https://opendatacommons.org/licenses/.

the dataset. Researchers should thus be cognizant of these contractual restrictions prior to sharing the datasets. They are advised to use repositories that comply with the FAIR principles and thus offer clarity of the licensing status (such as by relying on an open-source license).

6.2. Sharing data under GDPR

Researchers who wish to share *experimental data* or *real-world data* should ensure that the data is fully anonymized (e.g., by aggregating the data or removing any direct and indirect identifiers; see Sec. 2.4). However, it will often not be possible to fully and irreversibly anonymize a dataset. In such cases, researchers will need to comply with the restrictions of GDPR, including ensuring that they have a lawful basis for processing the data, such as the valid consent of the concerned individuals.

6.2.1. Sharing experimental data

Although the data is experiment generated, it cannot be guaranteed that it does not contain personal data. For instance, Woods et al. (2011) mention that during the scenario creation, actors accidently logged into a personal account. As the data can be shared if consent/authorization from third parties is given, it is wise to always do so and note it in the accompanying metadata.

6.2.2. Sharing real-world data

As for experimental data, if consent can be obtained, then the data can be shared. Alternatively, it has to be ensured that the data is anonymized. Consequently, this boils down to the question *can we anonymize real-world data* where the short answer is no.

Let us consider structured and unstructured data (including semi-structured data) as well as anonymization through modification and adding noise (differential privacy). Currently, the only combination providing guaranteed privacy is structured data plus differential privacy (Dwork, 2008; Kurakin, 2022).

Modifying structured data: Despite all research, several cases are known where wrongfully conducted anonymization resulted in the re-identification of individuals (Agencia Espanola Proteccion Dats, 2021). For instance, Sweeney (2000) linked hospital data with voter registration using ZIP, birth date, and gender. Another example is the Netflix prize dataset where Narayanan and Shmatikov (2008) de-anonymization the subscribers by linking records with IMDb. There are many more examples. The difficulty is to predict what new datasets will be available in the future allowing linkage attacks.

Modifying unstructured data: Anonymization techniques are difficult to apply to unstructured data apart from a few cases, e.g., one can develop regular expressions to identify identifiers in text such as email addresses or credit card numbers. However, identifying names or sensitive information in formats such as images or voice is not trivial. Depending on the amount of data, the anonymization is either done using artificial intelligence or manual work where both have limitations. For instance, let us consider the anonymization of legal documents which has been discussed by Csányi et al. (2021). The authors highlight several challenges including linking attacks for these documents. A more specific example is provided by Vokinger and Muehlematter (2019) who successfully conducted an attack against anonymized legal cases in Switzerland. McPherson et al. (2016), on the other hand, showed that obfuscation techniques such as pixelation or blurring can be reverted by modern image recognition methods. Lastly, AI does not (yet) have 100% accuracy.

Unstructured data and differential privacy: This is an area receiving more attention recently and has been summarized by Zhao and Chen (2022). As of now, the major challenge is that there is no general DP approach for unstructured data and each format must be processed separately, i.e., DP for images, DP for texts, etc. Operations are often complex and time-consuming. While this might be a viable option in the future, more research should be done especially with the question: how does it impact forensic research?

6.3. Utilizing real-world data

While it is challenging to share real-world data, there are instances where real-world data is released or can be collected and then can be utilized. Garfinkel et al. (2009) define those datasets as real but unrestricted and provide examples of "photos that can be downloaded from the Flickr photo sharing website and user profiles on Facebook." Another example is data that has intentionally been decommodified, e.g., through a court order such as the Enron dataset.

7. Conclusion

Generating and sharing data is essential to progress and to allow the comparison of results. Fortunately, we see that on the one hand, more data generation frameworks are developed and that researchers start to share their data – voluntarily or because they are required to by funding agencies. With more and more data being released, it is essential to understand when data can (cannot) be shared and to develop taxonomies and classifications allowing to search for needed datasets. In this article, we proposed a novel taxonomy (origin) that complements existing ones that primarily focus on the content of the dataset. We suggested that metadata should include the organization of the data (i.e., structured, semi-structured, or unstructured), the origin (our taxonomy), and content tags like those used by CFReDS. Based on the origin, we highlighted legal considerations and conclude that the dataset creator should obtain consent (including providing their own consent) as well as be careful in terms of special laws protecting data such as copyright or licensing.

Author contribution statement

Frank Breitinger: Conceptualization, Methodology, Validation, Investigation, Writing - Original Draft. Alexandre Jotterand: Writing - Original Draft, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Sofya for her bachelor thesis which provided the starting point for this work.

Appendix A

A Examples of organization of data

Let us consider Email as an example. Given the description, an email can be structured, semi-structured, or unstructured data, depending on which of the data is under consideration.

- If one only considers a fixed set of fields in the header (e.g., From, To, Subject, Message-ID, and Content-Type) this can be converted into a table and thus is considered *structured data.*
- If we consider the complete headers of a set of Emails, it is semi-structured data as the header will differ between emails, e.g., some may have X-MS-Fields others do not, some may have SPF information while some do not, etc. Given the inconsistency in fields and the impossibility to predict all fields, it is best described as *semi-structured.*
- If we consider the Email body, i.e., its content, we argue that this is *unstructured data* as it contains text and may include other elements such as images.

B Anonymization techniques

Anonymization techniques are used to generate non-identifiable datasets which then can be released. This section briefly[10] reviews anonymization techniques. Generally, the idea is to remove all direct identifiers (DI) and at least obfuscate indirect identifiers (II) so that identification is not possible anymore. According to Eyupoglu et al. (2018) and Majeed and Lee (2020), the following obfuscation operations exist:

Generalization replaces II with more general values, e.g., instead of providing the full email address, only the domain is given (abc@vwx.yz is replaced by vwx.yz), or age is replaced by an age range (25 is replaced by 20–30).
Suppression deletes (hides) II records, values, cells, or parts thereof. For instance, the last twelve digests of the credit card number are replaced by asterisks.
Permutation means that groups of attributes are shuffled which makes the association of II and SA attributes impossible. For instance, for each record, the columns zip, age, and gender are shuffled.
Anatomization divides the II and the SD into two tables and the data is released separately.
Perturbation means replacing the data with synthetically generated values that have identical (or at least similar) statistical information.

A common model describing the level of anonymity is *k*-anonymity originally presented by Sweeney (2002). By applying generalization and suppression, data is made less specific to a point that *k* entries in the database look alike and one cannot differentiate between them. However, the model is susceptible to attacks such as the homogeneity attack or the background knowledge attack allowing deanonymization.

Another technique to project privacy is differential privacy (DP) which was presented by Dwork (2008). DP makes general statistical information available, but at the same time protects the privacy of individual participants. To accomplish this, DP adds random noise to the data that has only a minor impact on the outcome. While secure, it only works on large datasets and the complexity of implementing it is higher than the aforementioned techniques. Nevertheless, many public and private organizations including the US Census Bureau (Jarmin, 2019) utilize DP to release their data.

C Artificial data vs. synthetic data

When looking into both terms, we found that some sources use them as synonyms while others see synthetic data as a subset of artificial data as synthetic. For instance, when exploring the term 'artificial data' using a search engine, most are about synthetic data and not artificial data. On the other hand, ChatGPT argues that they are similar but not the same as synthetic data refers to data that is artificially created and mimics real-world data (artificial data does not necessarily mimic real-world data, e.g., random data). Among several other sources, the European Data Protection supervisor agrees with the fact that synthetic data relates to real-world data and defines it as "artificial data that is generated from original data" (Riemann, 2022).

In digital forensics, the term artificial data is less common: searching on google scholar for "artificial data"+"digital forensics" lists 98 results whereas "synthetic data"+"digital forensics" returns 423 results. This 1:4 ratio is similar when removing "digital forensics", i.e., 83′100 vs. 398′000 results.

[10] A significant amount of work has been done in this domain and summarizing all is beyond the scope of this article. We outline some common approaches which we deem important for our discussion.

References

Abiteboul, S., 1997. Querying semi-structured data. In: International Conference on Database Theory. Springer, pp. 1–18.

Agencia Espanola Proteccion Dats, 2021. 10 Misunderstandings Related to Anonymisation. https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf, 2022-10-08.

Arasu, A., Garcia-Molina, H., 2003. Extracting structured data from web pages. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 337–348.

Berryhill, J., 2019. What is metadata? https://www.computerforensics.com/news/what-is-metadata, 2022-10-06.

bigdataframework.org, 2019. Three different data structures. https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/.

Buchholz, F., Spafford, E., 2004. On the role of file system metadata in digital forensics. Digit. Invest. 1, 298–309. https://doi.org/10.1016/j.diin.2004.10.002. URL: https://www.sciencedirect.com/science/article/pii/S1742287604000829.

Buneman, P., 1997. Semistructured data. In: Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 117–121.

Castelluccia, C., D'Acquisto, G., Hansen, M., Lauradoux, C., Jensen, M., Orzeł, J., 2022. Data Protection Engineering: from Theory to Practice. Technical Report. URL:. European Union Agency for Cybersecurity https://data.europa.eu/doi/10.2824/09079, 10.2824/09079.

Berne Convention, 1979. Berne Convention for the Protection of Literary and Artistic Works. https://www.sakpatenti.gov.ge/media/page_files/trt_berne_001en.pdf.

Council of the European Union and European Parliament, 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). URL: http://data.europa.eu/eli/reg/2016/679/oj.

Court of Justice of the European Union, 2009. Infopaq international A/S v Danske Dagblades Forening (C-5/08). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62008CJ0005.

Court of Justice of the European Union, 2014. Case C30/14. https://curia.europa.eu/juris/document/document.jsf;jsessionid=9ea7d2dc30ddb0cd971480fb446392f4aa5ac48d8cb3.e34KaxiLc3qMb40Rch0SaxuPaxj0?text=&docid=161388&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=253682.

Csányi, G.M., Nagy, D., Vági, R., Vadász, J.P., Orosz, T., 2021. Challenges and open problems of legal document anonymization. Symmetry 13, 1490.

Du, X., Hargreaves, C., Sheppard, J., Scanlon, M., 2021. TraceGen: user activity emulation for digital forensic test image generation. URL: Forensic Sci. Int.: Digit. Invest. 38, 301133. https://doi.org/10.1016/j.fsidi.2021.301133 https://www.sciencedirect.com/science/article/pii/S2666281721000317.

Dwork, C., 2008. Differential privacy: a survey of results. In: International Conference on Theory and Applications of Models of Computation. Springer, pp. 1–19.

Elliot, M., Mackey, E., O'Hara, K., 2020. The Anonymisation Decision Making

Framework: European Practitioners' Guide.
European Commission, Directorate-General for Research and Innovation, 2016. H2020 programme - guidelines on FAIR data management in horizon. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. Accessed: 2022-10-10, doi = 10.25607/OBP-774.
European Commission and Directorate General for Communications Networks, Content and Technology, 2016. Legal Study on Ownership and Access to Data : Final Report. Publications Office. https://doi.org/10.2759/299944.
European Parliament and the Council of the European Union, 2001. Directive 2001/29/EC on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32001L0029.
European Parliament and the Council of the European Union, 2019. Directive (EU) 2019/790 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC. https://eur-lex.europa.eu/eli/dir/2019/790/oj.
Eyupoglu, C., Aydin, M.A., Zaim, A.H., Sertbas, A., 2018. An efficient big data anonymization algorithm based on chaos and perturbation techniques. Entropy 20, 373.
Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. Digit. Invest. 6, S2–S11.
Göbel, T., Maltan, S., Türr, J., Baier, H., Mann, F., 2022. ForTrace - a holistic forensic data set synthesis framework. URL: Forensic Sci. Int.: Digit. Invest. 40, 301344. https://doi.org/10.1016/j.fsidi.2022.301344. selected Papers of the https://www.sciencedirect.com/science/article/pii/S2666281722000130 (Ninth Annual DFRWS Europe Conference).
Grajeda, C., Breitinger, F., Baggili, I., 2017. Availability of datasets for digital forensics – and what is missing. URL: Digit. Invest. 22, S94–S105. https://doi.org/10.1016/j.diin.2017.06.004 https://www.sciencedirect.com/science/article/pii/S1742287617301913.
Guido, M., Brooks, M., Grover, J., Katz, E., Ondricek, J., Rogers, M., Sharpe, L., 2016. Generating a corpus of mobile forensic images for masquerading user experimentation. J. Forensic Sci. 61, 1467–1472.
Horsman, G., Lyle, J.R., 2021. Dataset construction challenges for digital forensics. URL: Forensic Sci. Int.: Digit. Invest. 38, 301264. https://doi.org/10.1016/j.fsidi.2021.301264 https://www.sciencedirect.com/science/article/pii/S2666281721001815.
Jarmin, R., 2019. Census Bureau Adopts Cutting edge Privacy Protections for 2020 Census. https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census_bureau_adopts.html, 2022-10-09.
Jotterand, A., 2022. Personal data or anonymous data: where to draw the lines (and why)? Jusletter. https://doi.org/10.38023/f5daab01-6d80-472f-a0be-72a226aafe70.
Klimt, B., Yang, Y., 2004. The enron corpus: a new dataset for email classification research. In: European Conference on Machine Learning. Springer, pp. 217–226.
Kurakin, A., 2022. Applying differential privacy to large scale image classification. https://ai.googleblog.com/2022/02/applying-differential-privacy-to-large.html, 2022-10-12.
Li, N., Li, T., Venkatasubramanian, S., 2006. t-closeness: privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering. IEEE, pp. 106–115.
Majeed, A., Lee, S., 2020. Anonymization techniques for privacy preserving data publishing: a comprehensive survey. IEEE Access 9, 8512–8545.
Marr, B., 2019. What's the difference between structured, semi-structured and unstructured data? https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/?sh=6fedbb7d2b4d, 2022-10-05.
McCallister, E., Grance, T., Scarfone, K.A., 2010. Sp 800-122. Guide to Protecting the Confidentiality of Personally Identifiable Information (Pii).
McPherson, R., Shokri, R., Shmatikov, V., 2016. Defeating Image Obfuscation with Deep Learning arXiv preprint arXiv:1609.00408.
Moch, C., Freiling, F.C., 2009. The forensic image generator generator (forensig2). In: 2009 Fifth International Conference on IT Security Incident Management and IT Forensics. IEEE, pp. 78–93.
Moch, C., Freiling, F.C., 2011. Evaluating the forensic image generator generator. In: International Conference on Digital Forensics and Cyber Crime. Springer, pp. 238–252.
MongoDB.com, 2020. Unstructured data. https://www.mongodb.com/unstructured-data, 2022-10-05.
Narayanan, A., Shmatikov, V., 2008. Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (Sp 2008). IEEE, pp. 111–125.
Nemetz, S., Schmitt, S., Freiling, F., 2018. A standardized corpus for SQLite database forensics. URL: Digit. Invest. 24, S121–S130. https://doi.org/10.1016/j.diin.2018.01.015 https://www.sciencedirect.com/science/article/pii/S1742287618300471.
OpenAI, a.. DALL·E 2. https://openai.com/product/dall-e-2, 2023-04-06.
OpenAI, b.. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2023-04-06.
Riemann, R., 2022. Synthetic data. URL: https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en.
Rizkallah, J., 2017. The Big (Unstructured) Data Problem. https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/?sh=3fb6f8cd493a, 2022-10-05.
Roussev, V., 2011. An evaluation of forensic similarity hashes. URL: Digit. Invest. 8, S34–S41. https://doi.org/10.1016/j.diin.2011.05.005 https://www.sciencedirect.com/science/article/pii/S1742287611000296 (the Proceedings of the Eleventh Annual DFRWS Conference).
Scanlon, M., Du, X., Lillis, D., 2017. EviPlant: an efficient digital forensic challenge creation, manipulation and distribution solution. URL: Digit. Invest. 20, S29–S36. https://doi.org/10.1016/j.diin.2017.01.010. dFRWS 2017 Europe https://www.sciencedirect.com/science/article/pii/S1742287617300397.
Sweeney, L., 2000. Simple demographics often identify people uniquely. Health 671, 1–34.
Sweeney, L., 2002. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowledge-Based Syst. 10, 557–570.
Swiss National Science Foundation, 2017. Data Management Plan (DMP) - Guidelines for Researchers. https://www.snf.ch/SiteCollectionDocuments/ORD_Research_Council_3May2017_E.pdf.
Taylor, C., 2021. Structured vs. Unstructured Data. https://www.datamation.com/big-data/structured-vs-unstructured-data/, 2022-10-05.
Federal Assembly of the Swiss Confederation, 1992. Federal Act of 9 october 1992 on copyright and related rights (copyright Act, CopA). https://www.fedlex.admin.ch/eli/cc/1993/1798_1798_1798/en (Status as of 1 January 2022).
United States Patent And Trademark Office (USPTO), 2020. Public views on artificial intelligence and intellectual property policy. Technical Report. URL: https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf.
U.S. Copyright Office, 2022. Copyright Law of the United States (Title 17). https://www.copyright.gov/title17/title17.pdf.
U.S. Supreme Court, 2020. Google LLC V. Oracle America, INC. Certiorari to the United States Court of Appeals for the Federal Circuit No. 18–956. https://www.supremecourt.gov/opinions/20pdf/18-956_d18f.pdf.
Visti, H., Tohill, S., Douglas, P., 2012. Automatic creation of computer forensic test images. In: Computational Forensics. Springer, pp. 163–175.
Vokinger, K.N., Muehlematter, U.J., 2019. Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten (Banken). Eine Empirische Analyse anhand von Bundesgerichtsbeschwerden Gegen (Preisfestsetzungs-) Verfügungen von Arzneimitteln. Jusletter (online).
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 1–9.
Woods, K., Lee, C.A., Garfinkel, S., Dittrich, D., Russell, A., Kearton, K., 2011. Creating Realistic Corpora for Security and Forensic education. Technical Repor. Naval Postgraduate School Monterey Dept of Computer Science t.
Yannikos, Y., Graner, L., Steinebach, M., Winter, C., 2014. Data corpora for digital forensics education and research. In: IFIP International Conference on Digital Forensics. Springer, pp. 309–325.
Zhao, Y., Chen, J., 2022. A survey on differential privacy for unstructured data content. ACM Comput. Surv. 54, 1–28.