Forensic Science International: Digital Investigation 52 (2025) 301882

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

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

Optimising data set creation in the cybersecurity landscape with a special focus on digital forensics: Principles, characteristics, and use cases

Thomas Göbel [a,*], Frank Breitinger [b], Harald Baier [a]

[a] Research Institute CODE, University of the Bundeswehr Munich, 81739 Munich, Germany
[b] Institute of Computer Science, University of Augsburg, 86159 Augsburg, Germany

ARTICLE INFO

Keywords:
Data sets
Digital corpora
Use cases
Digital forensic repository
Data set creation
Synthetic data

ABSTRACT

Data sets (samples) are important for research, training, and tool development. While the FAIR principles, data repositories and archives like Zenodo and NIST's Computer Forensic Reference Data Sets (CFReDS) enhance the accessibility and reusability of data sets, standardised practices for crafting and describing these data sets require further attention. This paper analyses the existing literature to identify the key data set (generation) characteristics, issues, desirable attributes, and use cases. Although our findings are generally applicable, i.e., to the cybersecurity domain, our special focus is on the digital forensics domain. We define principles and properties for cybersecurity-relevant data sets and their implications for the data creation process to maximise their quality, utility and applicability, taking into account specific data set use cases and data origin. We aim to guide data set creators in enhancing their data sets' value for the cybersecurity and digital forensics field.

1. Introduction

Nearly two decades ago, Garfinkel (2007) and Garfinkel et al. (2009) emphasised the critical need for systematic development of reference data in digital forensics to support research and education. Since then, the landscape has shifted, with open science gaining increasing importance among researchers, universities, and funding agencies. The FAIR principles, introduced by Wilkinson et al. (2016), provide a foundational framework for ensuring that data is Findable, Accessible, Interoperable, and Reusable.

Despite this open science push, various studies reveal that data sets are often not shared. Abt and Baier (2014) studied the availability of network security data sets by systematically analysing accepted papers at the top IT security conferences from 2009 to 2013 and found that 70% of researchers manually create their data sets. Still, only 10% of the data sets have been published. Similarly, Grajeda et al. (2017) showed that data sets are often unpublished. The authors examined 715 peer-reviewed forensics research articles from 2010 to 2015. They indicate that only approximately 4% of the authors released their data sets, concluding that the forensic community suffers from limited availability of appropriate data sets. Gonçalves et al. (2022) surveyed the availability of smartphone data sets and found that only 31 publicly available data sets exist (9 older than 5 years, 18 older than 3 years). The authors conclude that most of the 31 data sets contain too few traces to be considered realistic.

Although the amount of available corpora has increased in recent years (Mombelli et al., 2024), technological progress is still hindered by the lack of available data, often also referred to as the *data set gap problem* (Park, 2018; Luciano et al., 2018; Gonçalves et al., 2022).

While releasing the data set is important, a frequently neglected aspect is the *process* of generating data, ensuring that it is a valuable asset to the community. Creating appropriate data sets can be tedious and time-consuming, and there is little formal guidance or best practices to assist with data set creation. As a consequence, data sets suffer from limitations such as privacy issues, intellectual property, unrealistic or insufficient wear and tear, background noise, or unknown ground truth (Grajeda et al., 2017; Park, 2018; Luciano et al., 2018; Göbel et al., 2023; Breitinger and Jotterand, 2023).

This raises the key question of this article:

What principles and properties are essential to produce high-quality data sets for digital forensics?

Answering this question is not trivial as data sets are used for different purposes, such as training, education, testing, or research.

* Corresponding author.
E-mail addresses: thomas.goebel@unibw.de (T. Göbel), frank.breitinger@uni-a.de (F. Breitinger), harald.baier@unibw.de (H. Baier).
URLs: https://www.unibw.de/digfor (T. Göbel), https://www.FBreitinger.de (F. Breitinger), https://www.unibw.de/digfor (H. Baier).

https://doi.org/10.1016/j.fsidi.2025.301882
Received 30 May 2024; Received in revised form 5 December 2024; Accepted 15 January 2025

Available online 29 January 2025

1.1. Scope and contribution

This article identifies aspects that require consideration before/when creating a data set. These factors ought to function as a potential basis for the standardisation of data set creation, serving as a guide for researchers and practitioners to assist them in creating data sets. In summary, this work provides the following contributions:

- A summary of existing terminology, techniques, and mechanisms on how the community can describe data sets based on an extensive literature review.
- A set of principles that provide clear guidelines for creating and processing data sets, ensuring their value to the community. These principles are derived from an earlier discussion of common challenges and expectations associated with forensic data sets.
- A synthesis of these principles with practical usage scenarios, such as tool testing and education, accompanied by a discussion on how effectively the principles meet community expectations in these areas.

Consequently, this article complements existing efforts which have stressed the importance of data sets (Sec. 2.1), discussed methods to classify works, e.g., through taxonomies (Sec. 2.2), or developed standardised corpora and centralised repositories (Sec. 2.3 to Sec. 2.5). In addition, there is more general literature such as the FAIR principles presented by Wilkinson et al. (2016), which ensures that data sets are *Findable*, *Accessible*, *Interoperable*, and *Reusable*, by humans and machines. These principles require data sets to be easily found, publicly available in standard repositories, and to have persistent identifiers.

1.2. Terminology

Many terms have established themselves when describing data sets which we summarise in Sec. 2. For us, a *data set* is a collection of digital data, which can be provided as files or images (e.g., volumes, discs, main memory dumps). We use the term *corpus* as a synonym of data set. For this work, we decided on the following terms: As this work targets the **data set creation process**, the term **principle** is best suited to describe recommendations that should be considered during data set creation. Ultimately, principles of the creation process impact the resulting data sets. On the other hand, every principle induces different aspects of the respective data set, which we call **properties** or **characteristics** of the particular principle.

1.3. Paper outline

Overall this article splits into two key parts. The first part gathers information about the usage of data sets in the community so far, while the second part presents our principles and the related discussion.

The first part starts with Sec. 2, where we present relevant background information and related work discussing the importance of data sets in cybersecurity, data set classifications and taxonomies, standardisation efforts, and specific examples of data sets in digital forensics. In Sec. 3, we describe our methodology in more detail, i.e., how exactly we conducted the extensive literature review to identify relevant principles for creating a qualitative data set. Sec. 4 outlines community expectations and requirements for valuable data sets, while Sec. 5 examines common issues encountered with data sets.

We begin Sec. 6 by outlining common use cases, introducing data set principles that complement FAIR guidelines, and urging researchers to adopt them to enhance usability and value. In Sec. 7, we evaluate our proposed data set principles by relating them to the desired properties, common issues, and typical use cases for data sets, thereby discussing the main findings and limitations of our work. Finally, in Sec. 8, we discuss future work and conclude, advocating for improved practices in data set creation and sharing.

2. Background and related work

This section highlights essential related work that discusses current problems with data sets and serves us to identify our principles and properties.

2.1. On the importance of appropriate data sets

Garfinkel (2007) states that without appropriate data sets, research in the various fields (e.g., disc forensics, network forensics, memory forensics, mobile forensics, etc.) is limited by the inability of experimenters to obtain large data sets that are realistic, varied, and representative of the data in the field. In other words, data sets are crucial in research as they facilitate experimentation and ensure the comparability and reproducibility of results (Garfinkel et al., 2009).

The position of Garfinkel (2007); Garfinkel et al. (2009) is supported by many other publications that point to the general lack of available data sets due to significant challenges (Abt and Baier, 2014; Baggili and Breitinger, 2015; Woods et al., 2011). A major challenge is the lack of data sources, especially real-world data, as law enforcement keeps the data secure and private or wipes disc images after a case is completed (Baggili and Breitinger, 2015). Furthermore, data from real cases contain personally identifiable information (PII) and cannot be shared for copyright, privacy, and data protection reasons (Abt and Baier, 2014; Grajeda et al., 2017; Breitinger and Jotterand, 2023). It is self-explanatory that real-world data are often unsuitable for education, as privacy-sensitive or illegal digital materials are confidential and cannot be shared (Woods et al., 2011).

Carrier (2010) points out that testing in the public view is essential to increasing confidence in software and hardware tools. However, suitable data sets are required for testing tools. Yannikos et al. (2014) state that well-known data corpora provide a basis for comparing methodologies and tools to identify the advantages and shortcomings. Similarly, Baggili and Breitinger (2015) refer to sufficient forensic tool validation to gain insights into the error rates for commonly used forensic tools (e.g., law enforcement agencies rely on properly functioning algorithms and tools in a court of law). According to Garfinkel et al. (2009), researchers and developers solve this problem by creating specific scenarios with synthetic data to conduct experiments, better understand new technologies, and test and verify the correct functioning of their algorithms and tools.

Ceballos Delgado et al. (2021) claim that realistic case studies are essential for successfully training digital forensic examiners but also stress that creating realistic data sets is both time- and resource-consuming. Hughes and Karabiyik (2020) state the importance of reference data representing the full range of conditions expected during the analysis. By carefully compiling reference data, the discipline enables peer review and reproducibility of testing and provides some traceability measures during validation testing. They further point out that conducting meaningful black box studies or proficiency testing is not feasible without curating a collection of test images.

Horsman and Lyle (2021) argue that there can never be too many data sets; provided they are structured effectively. They point out that anyone conducting research should consider creating a data set as a natural part of their research and development process, as creating and disseminating good data sets benefits everyone working in the field. The same work by Horsman and Lyle (2021) outlines and discusses a list of minimum requirements for data set creation for three specific data set types (cf. Sec. 2.2). Other good guidelines for creating data sets are provided on the NIST website along with templates to use (OSAC Digital Evidence Subcommittee Task Group on Dataset Development, 2022). Both works were considered by us and have influenced our work.

2.2. Classifications and taxonomies for data sets

Zheng et al. (2018) presented a taxonomy of cybersecurity research data sets. They divided data into four categories: attack-related data

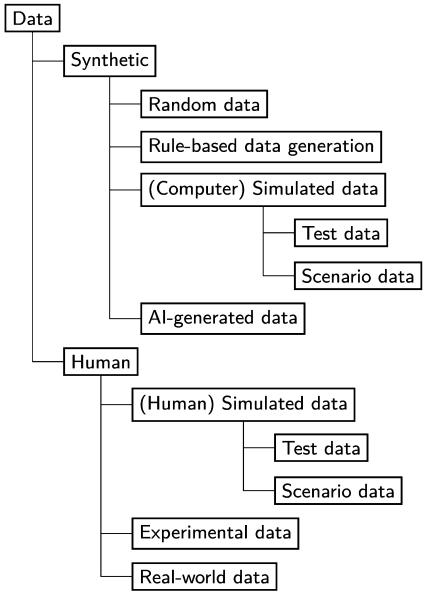

Fig. 1. Proposed taxonomy about the different types and origins of data sets (Breitinger and Jotterand, 2023).

set, defender artefacts, end user and organisation characteristics, and macro-level Internet characteristics. Each category contains several subcategories. Each of the data sets found was then manually assigned a category and a subcategory and it was indicated whether the data set already existed or was created by the authors and whether the data were made publicly available.

Using digital forensics as an example, the community has discussed various categories or data types, resulting in a taxonomy for data sets. For instance, Garfinkel et al. (2009) developed a taxonomy and differentiated between five categories of data: (1) Test data, (2) Sampled data, (3) Realistic data, (4) Real and restricted data, and (5) Real but unrestricted data. Five years later, Yannikos et al. (2014) condensed the number of categories and only distinguished on the top level between real-world data (like the *Enron email data set*) and synthetic data, which are separated between manually reproducing real-world actions and tool-supported synthetic data corpus generation. Grajeda et al. (2017) proposed three categories: (1) experiment-generated, (2) user-generated, and (3) computer-generated data sets. In contrast, Horsman and Lyle (2021) suggested (1) tool/process evaluation data sets, (2) actions data sets and (3) scenario-based data sets. Most recently, Breitinger and Jotterand (2023) developed the taxonomy depicted in Fig. 1 and defines the two main categories *synthetic* and *human generated data*, which we explain in what follows.

2.2.1. Synthetic generated data

Synthetic data is created by software with a certain degree of autonomy. Depending on the exact procedure, further categories may exist. For instance, *rule-based generation* describes software creating data deterministically way. *Simulated data generation*, on the other hand, uses system tools and thus, the respective outcomes vary, e.g., a file copied to a disc image will end up in different sectors/offsets.

An area receiving significant attention is *scenario data*. Various data generation frameworks (a.k.a. data synthesis frameworks) have been proposed in recent years to automate the generation of forensically relevant reference data sets based on an underlying scenario/story (Fragg, 2014; Visti, 2015; Park, 2018; Göbel et al., 2020; Du et al., 2021; Ceballos Delgado et al., 2021; Michel et al., 2022; Göbel et al., 2022; Demmel et al., 2024). Most of these frameworks are open source and aim to emulate user interactions within VMs, e.g., to generate disc, memory or mobile device images and network traffic captures, including synthetic traces. Scanlon et al. (2017) proposed a slightly different approach where differences (forensic challenges) to a previously distributed base image are distributed as evidence packages consisting of the modified artefacts and associated metadata. Further related work focuses on the application and evaluation of the existing data synthesis frameworks (Göbel et al., 2024) and on relevant extensions to them (e.g., for the synthesis of suspicious traces in the file system (Göbel et al., 2025) or for malware traces (Lukner et al., 2022)) as well as on the question of how the data quality of the synthetically generated images can be improved to be as realistic as possible (Schmidt et al., 2023; Wolf et al., 2024).

In addition, artificial intelligence (AI) offers promising opportunities to produce synthetic *AI-generated data*. For example, ChatGPT[1] can be used to create a meaningful storyboard for all kinds of cybercriminal scenarios and cases. AI can also be used to create synthetic evidence in the generated data sets, e.g., within disc images in the form of appropriate data in the file system, a chat conversation, notes, emails (Scanlon et al., 2023) as well as other relevant artefacts that correspond to the actual scenario, such as coherent background activity (Voigt et al., 2024).

2.2.2. Human generated data

The second part of the taxonomy in Fig. 1 is *human data generation*, which refers to data resulting from one or more humans interacting with a system in any way. This is further subdivided into *(human) simulated data*, *experimental data* and *real-world data* (Breitinger and Jotterand, 2023).

(Human) simulated data is the equivalent of synthetic simulated data and refers to data sets created by researchers, e.g., to test or validate the functionality of software. This category is further subdivided into *test data* and *scenario data*, with the latter having a higher complexity (e.g., disc images with full scenarios instead of a single file category). *Experimental data* is orchestrated by an individual/small group and requires a group of actors to produce the data. For *real-world data*, Breitinger and Jotterand (2023) use the description from Garfinkel et al. (2009) and define it as "data created by humans with no intention to create a forensic data set" (e.g., malware samples or data sold on the darknet).

2.3. Data set repositories in digital forensics

Researchers are well advised to publish the data on public repositories to obtain reproducible research results. A good source for data sets is the *Computer Forensic Reference Data Sets (CFReDS)*[2] platform maintained by NIST (2023a), which contains various data sets submitted by researchers and companies (Park et al., 2016). The repository enforces a minimum of documentation and thus supplies the community with documented samples as stated by Mombelli et al. (2024).

Further prominent repositories are *Digital Corpora*[3] (Garfinkel et al., 2009; Garfinkel, 2012) and *Data sets For Cyber Forensics*[4] (Grajeda et al., 2017). According to the CFReDS website, data sets from these two repositories have been incorporated into the CFReDS platform.

Moreover, it is still better to provide the forensic community with custom forensic artefacts instead of complete data sets than not sharing any data. Sharing custom artefacts can help other practitioners unfamiliar with these types of artefacts to solve similar forensics cases. Well-known platforms for sharing forensic artefacts include the *Artifact Genome Project (AGP)*[5] (Grajeda et al., 2018, 2023) and *MAGNET Artifact Exchange*.[6]

[1] https://chat.openai.com (last accessed 2024-11-30).
[2] https://cfreds.nist.gov (last accessed 2024-11-30).
[3] http://digitalcorpora.org (last accessed 2024-11-30).
[4] https://datasets.fbreitinger.de (last accessed 2024-11-30).
[5] https://agp.newhaven.edu/about/start (last accessed 2024-11-30).
[6] https://www.magnetforensics.com/artifact-exchange (last accessed 2024-11-30).

2.4. Digital forensic tool testing

The *Computer Forensics Tool Testing (CFTT)* program, also from NIST (2023b), offers methodologies for testing forensic software by publishing general tool specifications, test procedures, test criteria, test sets, and test hardware. The published test results provide the information necessary for toolmakers to improve tools, for users to make informed choices about acquiring and using specific computer forensics tools, and for interested parties to understand the tool's capabilities and functionalities.

2.5. Standardised corpora for digital forensics

Efforts have been made to create standardised forensic corpora that provide accessible data sets tailored to testing specific data structures or application-oriented tools. In addition, concerted efforts have been made to disseminate these curated data sets within the forensic community. Examples of corpora are listed in the following paragraphs:

Garfinkel et al. (2009) started developing representative standardised corpora for research. In the meantime, the research corpus grew substantially and includes files and disc images, `Govdocs1`[7] (a corpus of 1 million documents that are freely redistributable), the `Real Data Corpus (RDC)`,[8] RAM dumps, network captures, the prominent `m57-patents` synthetic data set (consisting of multiple storage images, memory dumps and network packets acquired during the generation process spanning 17 days) (Woods et al., 2011), and more (Garfinkel, 2012).

Davies et al. (2021) emphasised that the `Govdoc1` corpus as well a prominent subset of it called `t5-corpus`,[9] which is often used for approximate matching, are outdated and miss or underrepresent files such as modern Office document types, archive files, encrypted files and others. To facilitate research and in hopes of developing a modern and freely available standard data set, they published the mixed file data set called `NapierOne`[10] that aims at ransomware detection and forensic analysis research (Davies et al., 2022).

Back in 2010 Carrier (2010) released file system and disc images for testing analysis and acquisition tools.[11] Although this platform was last updated in 2010 and does not yield images providing recent file systems or partitioning schemes like Ext4, Btrfs, GPT, it is still worth using it due to its edge cases. Furthermore, Vidas (2011) introduced the memory analysis corpus `MemCorp` that consists of memory dumps acquired from physical and virtual machines and may be used by educators in many academic settings.

Nemetz et al. (2018) introduced a standardised forensic corpus that provides SQLite database files to evaluate strengths, weaknesses, and different analysis methods and tools in this scope.[12] The corpus comprises 77 databases grouped into five categories according to their peculiarities. The various databases use particular features of the SQLite file format or contain potential pitfalls to detect errors in forensic tools. As an extension, Schmitt (2018) performed various manipulations introducing anti-forensic aspects tested against different SQLite analysis tools.

Park (2018) proposed a methodology on how to generate a reference Windows registry data set called `cfreds2017-winreg` which includes user-generated and system-generated reference data extracted from different versions of Windows from Vista to 10.

Another prominent and frequently discussed problem is the existence of deepfakes, as they threaten the trustworthiness of online information (Thies et al., 2016). In research, and especially to detect such deepfakes, it is essential to have valid and comprehensive data sets available, such as the `FaceForensics++` facial forgery data set (Rossler et al., 2019) or the `Celeb-DF` deepfake video data set for deepfake forensics (Li et al., 2020).

The same applies to forensic research when analysing image and video files. As portable devices (especially smartphones) have become the preferred means of taking photos and videos in recent years, they pose new challenges for digital forensics. Determining a picture's or video's origin becomes even more challenging since the content is often distributed via social media platforms (e.g., Facebook, instant messenger, YouTube, etc.). For example, high-quality and comprehensive data sets are required to improve research and thus the detection of CSAM materials, such as shown by Gloe and Böhme (2010); Shullani et al. (2017).

3. Methodology

We used the following methodology in this study:

Search Strategy: We began by conducting an extensive search of relevant literature in digital forensics. Our search was conducted across multiple platforms, including Elsevier's ScienceDirect, Springer Link, and IEEE Xplore, as well as using search engines such as Google Scholar and ResearchGate. To capture a wide range of relevant studies, we used search terms including digital forensics corpora, synthetic data, data set generation, data set repositories, data set properties, characteristics, and issues. The search terms were designed to reflect key topics associated with data set creation, use, and challenges in both fields.

Inclusion and Exclusion Criteria: We focused on publications from the last 10 years to ensure the inclusion of contemporary research, though we also considered foundational works, such as the seminal papers from 2007 that highlighted the importance of data sharing in the digital forensics community. Studies directly addressing the creation, evaluation, or use of data sets in digital forensics were prioritised, while papers with a focus on unrelated subfields or too general in scope were excluded.

Data Extraction and Synthesis: After selecting relevant articles, we reviewed each paper, focusing on two key aspects: (1) the specific properties that data sets are expected to fulfil (Sec. 5), and (2) the common challenges (issues) encountered in the creation and maintenance of these data sets (Sec. 5).

Principle Formulation: Building on the challenges and expectations identified, we formulated in Sec. 6 a set of principles aimed at guiding the creation, documentation, and maintenance of data sets in digital forensics. These principles were designed to be contemporary, addressing gaps in the existing literature, and generally applicable across a wide range of use cases, including tool testing and educational purposes. Our goal was to ensure that these principles could be applied during the data set creation process to enhance the quality and longevity of the data.

4. Eligible properties of data sets

Beyond the FAIR principles (Wilkinson et al., 2016), researchers and practitioners in digital forensics have specific expectations for data sets, which this section summarises from existing literature as gathered by our methodology from Sec. 3. Since Garfinkel (2007) was among the first who highlighted the limited availability of large corpora, many of the core requirements for data sets trace back to his work. In summary, we identified 15 properties (P1 to P15) as listed in Table 1 and explained below:

[7] https://digitalcorpora.org/corpora/file-corpora/files/ (last accessed 2024-11-30).

[8] https://digitalcorpora.org/corpora/disk-images/real-data-corpus/ (last accessed 2024-11-30).

[9] http://roussev.net/t5/t5.html (last accessed 2024-11-30).

[10] http://napierone.com/Website/index.html (last accessed 2024-11-30).

[11] https://dftt.sourceforge.net (last accessed 2024-11-30).

[12] https://faui1-files.cs.fau.de/public/sqlite-forensic-corpus/ (last accessed 2024-11-30).

Table 1
Community expectations of the requirements, properties and characteristics of digital corpora and their correlation with the properties we have identified and described in Sec. 4.

Author(s)	Stated data set requirements/properties/characteristics	Properties correlation
Garfinkel (2007)	Representative, Complex, Heterogeneous, Annotated, Available, Distributed in open file formats, Maintained	P1, P2, P3, P4, P5, P6
Garfinkel et al. (2009)	Standardised corpora, Differing modalities of corpora, Corpora sensitivity, Restrictions on corpora use, Describing corpora with metadata	P2, P3, P4, P12, P13, P14, P15
Woods et al. (2011)	Multi-modal, Answer-keys, Realistic wear and depth, Realistic background data, Sharing and redistribution, Instructional materials	P2, P3, P4, P5, P9, P10, P11
Grajeda et al. (2017)	Quality, Quantity, Availability	P1, P2, P3, P5, P7, P8, P10, P11, P13
Nemetz et al. (2018); Schmitt (2018)	Development and use of standardised corpora, Documented ground truth, Metadata for traceability and reproducibility, Simple distribution	P4, P5, P9, P12, P15
Horsman (2019)	Sufficiently comprehensive, Fully tested, Documented, Containing 'evidence', Maintained (due to rapid technological development)	P1, P2, P3, P4, P6, P7, P8, P10, P11, P12, P13
NIST (2023b)	Known, documented structure, Description of the methodology used to create the data set, Record of the expected output, Statement on the scope of test case	P4, P9, P12

P1 *Representativeness*: A corpus should contain data typically encountered during criminal investigations, civil litigation, and intelligence operations (Garfinkel, 2007).

P2 *Complexity*: Large-scale forensic corpora should be complex, incorporating interlinked information from various sources (Garfinkel, 2007; Garfinkel et al., 2009).

P3 *Heterogeneity*: Data sets should reflect a diverse range of IT systems and usage patterns to encompass the multitude of technological environments (Garfinkel, 2007; Grajeda et al., 2017).

P4 *Annotation, ground truth*: In addition to the data and descriptive information (metadata), ground truth data are needed (Garfinkel, 2007; Garfinkel et al., 2009; Breitinger and Jotterand, 2023; Horsman, 2024). That is data documenting the relevant content of a data set, e.g., what artefacts are found in a disc image. This type of labelling may exist in the form of log files or specific data specification languages and helps others to reuse the corpus and validate research results.

P5 *Distribution*: Digital corpora should be distributed in open file formats and provided with tools to allow easy manipulation (Garfinkel, 2007). In the best case, the distribution is not restricted in any way (Nemetz et al., 2018).

P6 *Maintenance*: Maintaining a corpus may be essential to prevent obsolescence (Garfinkel, 2007; Horsman, 2019).

P7 *Quantity*: Data sets must contain a reasonable amount of data to train and validate approaches/tools (Grajeda et al., 2017). We point out that this is closely related to P1.

P8 *Quality*: Data sets must guarantee accurate and generalisable results through correct labels and similarity to real-world data (Grajeda et al., 2017). This is closely related to P1, P2, P3, P10, and P11.

P9 *Answer Keys*: Each digital artefact should include an answer key to explain what information can be found, where that information is located, and how the problems should be solved (Woods et al., 2011). We consider answer keys a special case of annotation/ground truth (P4).

P10 *Realistic Wear and Depth*: Digital artefacts should simulate realistic wear patterns and depth, resembling typical computer usage (Woods et al., 2011).

P11 *Realistic Background Data*: Incorporate a reasonable amount of non-case relevant background data to avoid scenario-based data dominance (Woods et al., 2011).

P12 *Metadata*: (Standardised) metadata or schema to describe corpora or elements within corpora should be provided, such as The Simple Dublin Core Metadata Element Set (DCMES)) (Garfinkel et al., 2009), the template for ground truth data by Horsman (2024), or at least specific descriptions, tags and persistent identifiers that help users locate data sets (Mombelli et al., 2024) (e.g., as applied by *CFReDS* (NIST, 2023a)).

P13 *Diversity*: Data sets should be diverse and comprehensive in terms of the amount of data and the traces they contain (Horsman and Lyle, 2021). For example, data sets must be available for both legacy systems and modern devices. While diversity includes the temporal component, heterogeneity (P3) primarily addresses different contemporary IT systems.

P14 *Sensitivity*: Corpora need to contain both sensitive and non-sensitive information. Access to real and restricted data sets should be controlled (Garfinkel et al., 2009).

P15 *Standardisation*: The development of representative standardised corpora is needed to further research and is essential for the long-term scientific health and legal reputation of the field (Garfinkel et al., 2009; Nemetz et al., 2018), describe as scenario data but less for a set of pictures.

Naturally, these properties may overlap, vary in granularity, and reflect the specific types of data sets that their respective authors had in mind when defining them. For instance, maintained (P6) or realistic background data (P11) make more sense for what Breitinger and Jotterand (2023)

5. Typical issues with data sets

In this section, we present ten common challenges and issues (I1 to I10) that have been discussed by several authors over the years and complement the expectations for data sets outlined in the previous section.

I1 Time costs

Preparing disc images, smartphone images, network captures, and other relevant corpora is time-consuming (Grajeda et al., 2017). The process involves creating a list of actions that simulate a 'security-relevant' scenario (often referred to as a *story*) executed manually in a virtual machine or sandbox environment. During the creation process, relevant artefacts are recorded and then made available. However, this workflow of (human) simulated or experimental data is a tedious, time-consuming and error-prone process (Woods et al., 2011). The time required to create a sufficient data set that can be used for tool testing should not be underestimated (Garfinkel, 2012). One of the biggest challenges in testing tools is generating and maintaining comprehensive and documented data to exhaustively test the functionality of tools. Therefore, maintenance is a key issue of manually created corpora (Horsman, 2019). Woods et al. (2011), for example, state that creating realistic corpora that are plausible, consistent and useful is a complex task requiring extensive planning.

I2 Missing wear and tear (a.k.a. background noise)

Data sets are often scenario-focused, resulting in limited scope. For example, a scenario-based image only covers a short time frame or misses a sufficient volume of non-pertinent wear and tear or realistic

background noise on the suspect device, which means that the resulting images are inherently limited and often appear too simplistic (Du et al., 2021) (note, in some cases, this may be desired, e.g., beginner-friendly). Data sets may not include sophisticated user actions (e.g., involving the GUI), which would lead to other artefacts (Park, 2018). Manual image generation is often limited to scenario-specific artefacts, as mimicking numerous actions to fully represent complex scenarios is labour-intensive. This is not in line with a typical incident, which often resembles the search for a needle in a haystack (Göbel et al., 2020).

I3 Shortage of standardised data sets

According to Garfinkel et al. (2009) and Horsman and Lyle (2021), there is a lack of standardised data sets. This is due to the fact that there are no standardised metadata and schemas for the creation and description of data sets (Mombelli et al., 2024). Without standardised corpora, "researchers at different organisations must waste time and money amassing their low-quality data" (Garfinkel, 2007), which are also often not released (Grajeda et al., 2017). Consequently, research performed on such data sets is not reproducible, has limited impact and is lost for future research (Schmitt, 2018). Even if data are published, comparing techniques, research results and findings is often difficult when the data are self-selected and inadequately documented.

I4 Outdated/bad timeliness

Once an image is created, it is static, i.e., it cannot be adjusted without recreating the entire image (Göbel et al., 2022). Consequently, released data sets may not reflect the state of the art (Garfinkel, 2007). While this is essential for testing purposes, it poses other challenges. For instance, reusing the same data sets in a (graded) educational setting may not be possible as corresponding write-ups or walkthroughs become available. In addition, they become outdated, i.e., they no longer contain the latest versions and can therefore lose their relevance (Grajeda et al., 2017).

I5 Legal barriers

Legal barriers may restrict the use or sharing of data sets, especially in the case of real-world data (where all types of sensitive data would first need to be anonymised or redacted). For example, forensic investigators with access to real-world data are subject to legal and practical restraints that prevent the data from being used in research (Garfinkel, 2007). In addition, Breitinger and Jotterand (2023) state that data may be protected by copyright, special law, contractual provisions, or privacy or data protection laws that may impose rules and restrictions on sharing personal data.

I6 Limited completeness

Horsman and Lyle (2021) note as part of the results of their practitioner survey that while the data sets available in the repositories provide a good basis for forensic tool testing, they are non-exhaustive and do not provide the depth necessary to test the complete functionality of all kind of tools effectively. Garfinkel (2010) also drew attention to the fact that there is a lack of complex, realistic training data, which means that most classes are taught using simplistic manufactured data. Also, poor transferability may occur, as corpora are tied to a specific local environment, operating system, apps, or even a specific country or language (Yannikos et al., 2014). For instance, the *Enron email corpus* may be valuable to linguists in English-speaking countries but less for researchers focusing on other languages. The same applies to a corpus developed to test specific tools or applications that may not be available or widely used in different regions.

I7 Bad adaptability

Adapting a data set to different settings or an updated software, application, or operating system version can be cumbersome or even impossible. For instance, changing the language from EN to ES of a disc image. Consequently, the data set often has to be recreated, which is costly (see I1) (Göbel et al., 2020, 2023).

I8 Missing/insufficient ground truth (data)

Knowing the ground truth is essential to use the data to evaluate (new) tools and methods using objective metrics (Garfinkel et al., 2009), i.e., without ground truth data, no confidence can be established in an existing data set. However, defining it is difficult and time-consuming. It may even be infeasible to establish the ground truth on any set of non-trivial sizes. This is why some research undertakes a controlled study rather than a real-world data study (Roussev, 2011). Even if some ground truth data are provided, it may not be sufficient (depending on the use case), as it would be necessary to organise a corpus with accurate logs and timelines of performed actions with as much detail as possible (Park, 2018). Furthermore, not all traces in public data sets are as intended, and inconsistencies can be discovered because, in many cases, the exact creation process of the data set is not known or documented (Woods et al., 2011).

I9 Lack of variety

Grajeda et al. (2017) points out that the available data sets have poor variety, e.g., many images originate from IoT or smartphone images that only come from the same widespread vendor. Data sets from devices considered less likely to be used in cyber incidents (even if there is evidence that they have been used in the past) are rare. These include, for example, data sets from game consoles, Smart TVs, IoT devices, drones and voice assistant devices. Therefore, in a world with a rapidly growing amount of interconnected devices and the era of big data, there is a need for a large number and variety of available data sets to provide comprehensive insights into the different devices under investigation, including various types of digital evidence.

I10 Lack of metadata

Many of the existing data sets lack comprehensive metadata, especially if they do not have external, supplemental data to complement the information found in data repositories such as *CFReDS* (Mombelli et al., 2024), and there is still no standardised metadata or schema to describe forensic corpora or elements within a corpora (Garfinkel et al., 2009; Horsman and Lyle, 2021). Mombelli et al. (2024) assessed the completeness of the metadata and compliance with the FAIR principles using 212 data sets from NIST's *CFReDS* and highlighted deficiencies in metadata quality and FAIR compliance, emphasising the need for improved data management standards.

6. Definition of principles and related properties and characteristics for data set creation

Some of the previously highlighted aspects of data sets are contradictory, e.g., a standardised data set cannot be updated frequently, or a representative image for a particular case does not necessarily require a large number of wear patterns. Consequently, aspects depend on the use case for the data set (i.e., its context). This section first highlights four common use cases of data sets and then describes in detail five principles and their associated data set properties and characteristics that should be considered when creating a data set to be of value to the community. We then discuss the implications of each data set principle and its properties and characteristics for the respective data set use case and the community's expectations of valuable data sets in Sec. 7.

Table 2
Common use cases that require data sets (adapted; originally published by Göbel et al. (2023)).

Item	Use Case	Key aspects
1	Method/Tool Testing and Validation	Adaption to recent software, hardware, concepts Evaluation of error rates and limitations Tool's functionality and ability in handling both modern/legacy systems Assessment with respect to cyber incidents/anti-forensics
2	Practitioner Training/ Education	Incident training in a contemporary environment For educational use, knowledge competitions, competency/proficiency tests One-time tasks for exercises and exams in university education
3	Research & Reproducibility	Scientific and forensically sound proof of a hypothesis Verification/peer-review of artefacts/traces and their interpretation Re-usage by the community Enhancing trust in research results
4	Machine Learning	Large-scale training data to build machine learning models Unbiased training data to get models close to reality

6.1. Data set use cases

Horsman and Lyle (2021) state that data in the context of digital forensics are utilised for the following three purposes: (1) Training, (2) Tool/process evaluation, and (3) Data exploration and reverse engineering (research & development). Two years later, Göbel et al. (2023) expanded this work and described the four common use cases (1) Method/Tool Testing and Validation, (2) Practitioner Training/Education, (3) Research & Reproducibility, and (4) Machine Learning, which we summarised and adapted for our needs in Table 2. In their article, Göbel et al. (2023) also illustrate why the particular use case is relevant and, more importantly, explain why and what kind of data is required. The authors provide information on which characteristics the data sets should fulfil to be most useful for the respective use case.

The use cases presented all rely on data and data sets, but as can be seen from the key aspects in Table 2, the requirements for a specific data set (and thus the creation and development of the data set) vary depending on the use case for which the data is employed. In some cases, knowledge of the exact ground truth data is essential (e.g., in forensic practitioner training and education, i.e., in an academic setting, or competency, certification or proficiency tests), in other cases, it is less important or the subject of interest (e.g., in research). On the other hand, sometimes large amounts of data are needed (e.g., for machine learning), while there are cases where one sample may be sufficient (e.g., to test a specific functionality of a tool to detect a certain artefact).

6.2. Principles for data sets and its creation process

This section introduces a set of principles designed to address key challenges in data set creation, complementing existing guidelines like the FAIR principles (Wilkinson et al., 2016). These principles highlight crucial steps to ensure data sets are high-quality, reproducible, and valuable for the digital forensics community. Each principle is explained with a focus on addressing common issues and improving the overall data creation process.

Principle 1: Decide on a use case and a goal/objective/purpose for the data set

Before creating the data set, the creator has to decide on the objective or purpose of the data set, which generally is related to the use case. This must be done as a first step as this principle impacts subsequent principles. For instance, one may want to create a data set to test a new parser for JPG images or to validate if a tool can handle non-ASCII characters.

Principle 2: Given the taxonomy of data set types, one settles for the most appropriate data origin

The taxonomy in Fig. 1 defines how the data set is created, i.e., the origin of the data set (e.g., a statement about whether the data were created manually by a human or synthetically or whether it is real-world data). The exact method (e.g., synthetic scenario data, human experiment, or rule-based data generation) should be carefully considered as it affects further characteristics of the resulting data set, such as:

Determinism/Repeatability: Does the procedure allow recreating the data set? This is difficult to define and depends on the use case and the objective. For instance, a disc image containing ten deleted photos can be reproduced. However, it is unlikely that the offsets of all fragments are identical. A rule-based data generation process should ensure this. Besides its basic scientific importance, repeatability is also important if (periodic) updates are desired. In this case, only synthetic data approaches are feasible.

Scalability/Adaptability: Is it easy or difficult to scale a data set in size or adapt its content? This depends on the origin of the data set. For example, if there is a possibility to generate (computer) simulated data (e.g., using a data synthesis framework), then multiple data sets with slight changes (e.g., evidence hidden at different offsets) can be generated, which would require significantly more time if generated manually. Another example would be the ability to create a snapshot of a state in a VM, which serves as the basis for different scenarios. Existing data sets are also easier to update or extend using automated tools. On the other hand, initially implementing such a generation framework can also be complex and may only make sense if scalable data sets are needed, e.g., in forensic tool testing.

Data source: Is the exact data source specified, i.e., is it indicated from which device, hardware, software, etc., the data were collected so that one has access or means to reproduce the data acquisition process?

Principle 3: The data set should be of sufficient quality and representative

We consider a broad meaning of the term *quality* and subsume apparent aspects like complexity, heterogeneity, and proximity to real-world traces to represent actual threats and cyber incidents, but also volume-related aspects like quantity under this term. Depending on the objective, it may need other artefacts. Characteristics that fall under this principle are:

Data Volume: Define the data set volume by including relevant digital objects such as appropriate image types (e.g., disc images, network traffic captures, mobile phone device data, IoT device, drone dumps, etc.) and concrete data set content specification (e.g., type of files, file names, number of files, data set size, etc.), to ensure completeness.

Scope: The scope of a data set should be described. For instance, an SQLite database may have the purpose of testing a tool's detection mechanisms for deleted records. However, the scope of the data set may be limited to a particular version of SQLite.

Post-processing: This step may involve actions like filtering elements or data reduction to refine the data set's content and improve its relevance for specific purposes or research goals.

Validation: The validation of the data generation process and the resulting data aim to assess and ensure its accuracy, completeness, and adherence to predefined standards, enhancing the data set's quality, reliability and comparability.

Edge cases: When creating a data set, it may be essential to consider and include edge cases, which are instances or scenarios that are atypical or less common but may be crucial for comprehensive testing and research in digital forensics.

Seed/Randomness: Depending on the objective or use case, some seed or randomness is required to create slightly or completely different samples. For instance, in forensic education, one may need multiple images in an exam that are somewhat different, triggered by a seed value.

Principle 4: Ensure adequate documentation and transparency

Careful documentation of the data set, its creation process and constraints is crucial for reliable, versatile and effective use of the data. Sufficient ground truth (data) is essential for most use cases that should be considered during the data set creation process. Sample questions to consider are: Is it possible to automatically log all activities (which is often the case when using data synthesis frameworks such as `TraceGen` (Du et al., 2021) or `ForTrace` (Göbel et al., 2022; Wolf et al., 2024))? Can the data set be created with a script that allows reproducibility (rule-based data generation according to Breitinger and Jotterand (2023)) and perfect ground truth? Determinism and repeatability therefore play an important role here. Characteristics in the context of documentation and transparency are:

Metadata: Descriptive information should be provided to describe the data set or elements within a corpus that allows reusing it (e.g., used system specifications, technical environment, tools, processes, software versions, the meaning of variables, etc.), its creators, intended use, constraints/limitations, whether/what kind of ground truth data exist (Horsman and Lyle, 2021). A standardised metadata description language or schema is used in the best case. For example, when uploading a new data set to *CFReDS*, several descriptive fields must be filled. These serve as metadata for the data set, e.g., year of creation, title, short/long description, details about the uploader, other tags for detailed classification and characterisation that facilitate the search (Mombelli et al., 2024). Horsman (2024) proposed 20 metadata fields that should accompany a data set. These include information about the creators, format/name/hash value of the data set, hardware details, date and time of creation/acquisition, and acquisition methods/tools, among others.

Ground truth (data): In addition to the metadata that generally describes a data set and its content, it is preferred and often beneficial to have labelled data (or ground truth data) available. This involves, in particular, contemporaneous records of all user actions during the data set creation, including timestamps. For instance, it includes logon/logoff processes, sent/received data and messages, start/end of applications, file system events, changes to settings, etc. Labelled data are required for tool validation. Otherwise, it cannot be proven that the software works as expected. In an educational setting, one also needs to know what kind of malware, traces, artefacts, evidence, etc. can be found in the training data sets and where to find it. In machine learning, it depends on whether it is a supervised algorithm (where training data are required) or an unsupervised algorithm. Horsman (2024) proposed a template with detailed records of the schedule of activities conducted during the data set creation to support the creation of ground truth data. Besides general information about the service or software being used, he suggests documenting the interaction type (e.g., visiting a website), the input data (website address), the time of the activity and the primary trace considered (Internet history record).

Reproducibility: To be most valuable to the forensic field, research results performed on data sets need to be traceable and reproducible (note that this property correlates to *Determinism/Repeatability* of Principle 2). This is best done when the data set itself is reproducible, i.e., its structure, the data it contains, and how the data set was created/collected/acquired must be documented in detail (e.g., Nemetz et al. (2018) provide all SQL statements necessary to reproduce their SQLite corpus). If customised or automated methods, processes, tools, technical environment, frameworks, etc., have been used during the creation process, access to these means should be provided so that the method itself is understood and can be reused. Otherwise, researchers may be unable to reproduce, verify, or build upon results obtained based on an unknown data set.

Principle 5: Ensure compliance with ethical and legal regulations

Adhere to relevant legislation and ethical restrictions when creating and publishing the data set, particularly when handling personal or sensitive data. Characteristics of this principle are:

Legislation: Legislation of the created data set must be clearly defined. The creator of a data set should provide a statement as to whether and to what extent the data set contains PII or copyright-restricted data (e.g., licence keys) and whether the data set can be used worldwide without restrictions or has geographic limitations.

Shareability: The shareability of a data set (which ensures reproducibility and a thorough (peer-)review) is directly influenced by ethical and legal compliance. Prior information indicates whether or not a data set can be shared and, if so, to what extent.

Unbiasedness: The data set creation process should reflect if ethical aspects are relevant and must be respected to be unbiased. For instance, if personal attributes like sex or skin colour are relevant, this characteristic shall avoid restricted representativeness.

7. Discussion and limitations

The previously defined principles show that consideration of the actual purpose or goal of a data set plays an important role before/when creating the data set, as this influences whether and to what extent other principles and characteristics are met. In this section, we discuss the proposed principles, properties, and characteristics and their limitations by describing their impact on metadata (Sec. 7.1) and by correlating them with the community's expectations and requirements for valuable data sets. Next, we summarize common issues encountered with data sets (Sec. 7.2). Lastly, we provide an integration of the defined data set properties and characteristics with the relevant use cases (Sec. 7.3).

7.1. Principles impact on metadata

Since many existing data sets lack comprehensive metadata (Mombelli et al., 2024), each data set should be accompanied by a description that outlines its content (see Principle 4). These metadata may also be needed to ensure findability (cf. FAIR principles (Wilkinson et al., 2016)), e.g., it may be indexed and thus is searchable.

The principles for creating records significantly impact the creation of good metadata. Following these principles, metadata creation becomes a structured and purpose-driven process. For example, the *scope* and *data volume* properties of Principle 3 require metadata to include information about file/image types, the number of files, and the size of the data set and guide metadata creators to provide essential details about the content of the data set. Principle 4 (*adequate documentation and transparency*) encourages appropriate documentation of metadata creation methods and decisions, ensuring transparency and traceability in metadata creation. Furthermore, it provides information on how a data set should ideally be described so that it can be reproduced. In addition, the *validation* property of Principle 3 emphasises the need to validate metadata alongside the data set to ensure its accuracy and completeness.

Table 3
Mapping of principles to eligible data set properties (from Sec. 4) and typical data set issues (from Sec. 5) (Brackets indicate whether the treatment of properties/issues depends on the respective data origin, i.e., human vs. synthetic data origin).

Proposed Principle	Addressed Data Set Properties	Addressed Data Set Issues
Principle 1	–	–
Principle 2	P1, P2, P3, P6, P7, P13	I1, (I2), I4, I5, (I6) I7, I8, I10
Principle 3	P1, P2, P3, P7, P8, P10, P11, P13, P15	I2, I3, I4, I6, I7, I9
Principle 4	P4, P9, P12, P15	I3, I7, I8, I10
Principle 5	P3, P5, P12, P13, P14, P15	I3, I4, I6, I9

In summary, the aforementioned principles guide metadata creators in creating accurate, comprehensive and well-documented metadata consistent with the data set's intended use, thereby improving the overall quality, utility, comparability, reproducibility and usability of the data set for its intended use case.

7.2. Principles in relation to the desired properties and common issues of data sets

In Sec. 6.2, adequate principles and properties for data set creation were presented. This section highlights our suggestions on how these principles can be used as a guide when creating data sets to address the general data set requirements and existing issues, as described in Sec. 4 and Sec. 5, respectively. Table 3 illustrates which of the expected data set properties (previously annotated with P1 to P15 in Sec. 4) and described issues (previously annotated with I1 to I10 in Sec. 5) are fulfilled if the data set creators comply with the individual principles.

Principle 1: Decide on a use case and a goal/objective/purpose for the data set. As mentioned, compliance with the subsequent eligible properties is also ensured to some extent by defining the actual use case and goal of the data set before it is created. An objective could be, for example, to forensically analyse and compare different IoT devices. This also allows the appropriate data origin to be specified more precisely (Principle 2). Interestingly, from Table 3, we deduce that Principle 1 seems to address a new aspect of data set generation as we do not find any mapping from Principle 1 to previously raised properties or issues.

Principle 2: Given the taxonomy of data set types, one settles for the most appropriate data origin. To stay in the scope of IoT devices, an appropriate data origin is next determined. If the IoT device can be emulated, synthetic data would be preferred. If not, only real-world data (collected from existing devices) or experimental data (created with real devices in a dedicated testbed) can be used. The given taxonomy and thus the origin of the data set in turn determines whether a data set is adaptable and scalable and thus if the generation process is easily repeatable. Therefore, depending on the correct choice of the most appropriate data origin, one may address several of the typical data set issues and desired properties.

Whenever it is feasible to automate the data set creation process, this should be considered, thereby addressing in particular the data set issues I1 (time costs), I4 (outdated/bad timeliness), I5 (legal barriers), I7 (bad adaptability), I8 (missing/insufficient ground truth), and I10 (lack of metadata). On the other hand, it must be considered that I2 (missing wear and tear) and I6 (limited completeness) are general issues of synthetically generated data sets and are typically better fulfilled by, e.g., real-world data (which is why they are in brackets).

Following Principle 2 may help to produce a data set addressing P1 (representativeness), P2 (complexity), P3 (heterogeneity), P6 (maintenance), P7 (quantity), and P13 (diversity).

Principle 3: The data set should be of sufficient quality and representative. Principle 3 and its related characteristics (i.e., the data set's volume and scope, information on post-processing, validation steps, edge cases, provision of a seed, randomness or other minor changes to the data set) help to create synthetic data sets reasonably so that they address, in particular, P1 (representativeness), P2 (complexity), P3 (heterogeneity), P7 (quantity), P8 (quality), P10 (realistic wear and depth), P11 (realistic background data), and P13 (diversity). Following Principle 3 may help to solve issues like I2 (missing wear and tear), I3 (lack of standard data sets), I4 (outdated data sets), I6 (limited completeness), I7 (bad adaptability), and I9 (lack of variety).

Principle 4: Ensure adequate documentation and transparency. Based on adequate documentation and thus respecting the properties of Principle 4, in particular, P4 (annotation, ground truth), P9 (answer keys) and P12 (metadata) are addressed, thereby solving the issues I8 (missing/insufficient ground truth (data)) and I10 (lack of metadata). Typically, the provision of adequate documentation and the achievement of transparency are easier to achieve if the data are generated synthetically.

However, there are data set objectives for which human-generated data are best suited (cf. Sec. 7.3) as synthetic data generation is not always feasible or practical. For example, a typical scenario in forensic research would be that current software, hardware, or any emerging technologies are of interest to provide the community with new insights into the forensic processing of this modern technology. No data generation tool is likely available at the beginning of such a new research project, so people have to create the data manually, or an appropriate data synthesis tool must be developed or adapted first. Especially for data sets with human origin (as it is typically more difficult to achieve here), compliance with Principle 4 is fundamental, as this helps to address the requirements P4 (annotation, ground truth), P9 (answer keys), and P12 (metadata). Because this kind of documenting and describing information is typically missing or lacking (mostly in human-generated data sets and especially in real-world data), issues such as I3 (shortage of standardised data sets), I7 (bad adaptability), I8 (missing/insufficient ground truth (data)), and I10 (lack of metadata) appear as to why data sets and research results obtained with them are typically not reproducible.

Principle 5: Ensure compliance with ethical and legal regulations. As long as there are no legal issues, in particular, human-origin data sets should always be considered for sharing with the community so that the field does not continually suffer from outdated (I4), incomplete (I6), not varied enough (I9), or even completely missing (standard) data sets (I3). Synthetically generated data generally cause fewer problems when sharing, i.e., their publication should always be considered so that research results can be reproduced more efficiently. In general, Principle 5 and its related characteristics address the eligible properties P3 (heterogeneity), P5 (distribution), P12 (metadata), P13 (diversity), and P14 (sensitivity).

7.3. Principles in relation to the use cases

In what follows, we specify a relation of our data set principles and properties from Sec. 6.2 to the four common data set use cases from Göbel et al. (2023) as presented in Sec. 6.1. Table 4 provides a complete

Table 4
Assignment of specific data set properties to the presented data set use cases (Evaluation matrix: ++ = absolutely necessary; + = necessary; 0 = borderline; - = not necessary; - - = not necessary at all).

Properties/Characteristics	UC1	UC2	UC3	UC4
Determinism/Repeatability	++	+	++	+
Scalability/Adaptability	++	+	+	++
Scope	++	0	++	+
Post-processing	0	0	++	++
Validation	++	+	++	+
Edge cases	++	0	++	+
Seed/Randomness	0	++	-	-
Metadata	+	++	++	+
Ground truth	++	++	++	++
Reproducibility	+	++	++	+
Legislation	+	++	++	++
Shareability	0	+	++	0
Unbiasedness	-	-	+	+

overview of our mapping. Our assessment is based on an established five-level rating scheme, starting with ++ for a high level of compliance and ending with - - for no compliance. Overall, we conclude that many properties are generally required, regardless of the use case, which is a good result, as it shows that many properties are relevant overall. We remark that there is no double minus score (i.e., it is impossible to neglect a property completely) and only a few negative scores. In what follows, we explain sample assessments of our mapping.

UC1: Method/Tool Testing and Validation

When testing new and existing tools and methods, sufficient test data are needed to exhaustively test countless functions built in many different software suites on the market. Therefore, the test data sets need to be comprehensive in terms of size and variety, including data for modern and legacy systems. Only based on suitable test data can we ensure that forensic software consistently delivers accurate results in a reasonable time.

However, it is challenging to provide appropriate and enough test data sets and keep them up-to-date to test the ever-increasing capabilities of modern forensic software. Therefore, the preferred and most appropriate data origin for this use case is synthetic data (e.g., rule-based data generation or (computer) simulated test or scenario data) to generate test data sets faster without the great manual effort of creating a multitude of sometimes only slightly differing data sets. However, due to the ever-changing trends, it can be difficult to generate a synthetic data set that immediately keeps up with new findings. Therefore, the use of real-world data for testing software can be considered if the exact content of the image is known. Concerning the defined properties, a well-documented data set (in the form of existing *metadata* and *ground truth data*) is, therefore, an essential part of comparing the tool's result against the expected result. In addition, the two properties, *repeatability* and *scalability*, play an important role. Moreover, more *edge cases* and thus somehow inconsistent data sets may be needed to prove that a tool's detection mechanisms are also valid for certain unexpected inputs.

UC2: Practitioner training/education

To keep up with the continuously changing challenges, forensic practitioners depend on training data covering a broad spectrum of criminal activities to train their skills. Depending on the training, the data may be either specific and narrow (e.g., a single PCAP file of a data breach to be analysed with Wireshark) or comprise a complex scenario (e.g., the M57-patents case[13] including multiple disc, network, and RAM dumps with a complex and realistic storyline (Woods et al., 2011)). Typically, training and education focus on the latest trends and developments, so training data must be up-to-date (e.g., a workshop on the acquisition and analysis of modern devices, such as IoT devices or drones, is currently more in demand than a training session on file carving).

While in commercial training centres, it is preferred that everybody works on the same case (for reasons of comparability), forensic training in educational institutions may require flexible and easily adaptable challenges, as solutions to exams and practical exercises can be leaked and quickly spread over the Internet which would limit the learning effect (Göbel et al., 2023). Hence again, the data origin plays a role here. As for super complex scenarios, only (human) simulated scenario data, (human) experimental data, or even real-world data (in most cases, however, not possible due to legal concerns) come into question. The latter academic setting benefits from synthetic data generation and its typical characteristics of better *scalability*, *adaptability* and *repeatability*. In addition, using some *seed*, *randomness*, or different order of instructions in an automatic data generation process can help produce the desired slightly different data sets. In addition, at least the instructor in an educational setting must know the exact *ground truth* and thus the expected findings of a data set for grading to be possible.

UC3: Research & reproducibility

For verification and replicability of research results, especially the property *repeatability* for data set creation is essential, as it is often relevant to reuse data sets to compare one's findings with the research efforts and results of others. Sometimes adjustments need to be made to existing data sets, so the ability to repeat the data set creation is helpful (cf. with the properties *scalability* and *adaptability*). This requires thorough *documentation*, i.e., providing metadata for the data set. *Validation* of the content of the data set is critical, e.g., to verify that the creation of the data set was done correctly and that relevant artefacts and required data are present in the data set. Previous post-processing of a data set should be stated and later adaptations should be possible, as this may be relevant depending on the research goals. *Shareability* is essential to ensure reproducibility and a thorough (peer-)review. The *specificity* of the data set also plays a role here, so it is typically better for initial research to use a series of smaller samples than complex scenarios. The data origin in research depends (on human vs. synthetic), as it is likely that no data synthesis tool exists at the beginning of a new research project. Consequently, one mainly encounters real-world data or experiment-generated data in this use case.

UC4: Machine learning and deep learning

Appropriate labelling and thus high-quality ground truth data are required for data sets used as training data in machine learning and deep learning. Furthermore, the need for a large amount of data is noteworthy. The need for novel data sets (including current trends, use and impact) also plays a major role in developing machine learning models. A common class of machine learning is supervised algorithms, where the algorithm creates its model based on a labelled (training) data set and where feature selection is relatively easy. In unsupervised learning, the data are not labelled, so the algorithm must automatically select features and find patterns and relationships in the data on its own. There are also semi-supervised algorithms, which are a mixture of supervised and unsupervised methods.

While practitioners prefer to use real-world data or at least experimental data (according to the data type taxonomy) for training to obtain realistic machine learning models, law enforcement typically lacks

[13] https://digitalcorpora.org/corpora/scenarios/m57-patents-scenario/ (last accessed 2024-11-30).

publicly available and labelled training data, making the creation of machine learning models a difficult task. Therefore, mainly the defined properties *ground truth*, *scalability*, *adaptability* and *data volume* suffer in machine learning if it only relies on real-world data with a limited amount and perhaps even unknown content.

So, the actual decision for the ideal data origin in machine learning depends on various factors. While real-world data can be noisy, inconsistent, or biased, synthetic data can be generated with specific properties, ensuring the quality and consistency of the data (e.g., the data can be labelled appropriately directly when the data set is created). On the other hand, the synthetic generation of training data (which generally has to be up-to-date due to increasing technical progress) can be more cost-efficient than the collection and annotation of real-world data. In addition, synthetic data can usually be used without privacy concerns. However, depending on the actual classification task and the required threshold of the machine learning algorithm, synthetic data may not be as close to the real-world data as necessary. However, synthetic data are the only option anyway when real-world data are not available at all or not available in sufficient quantity.

7.4. Enforcement of the principles

We do not believe that the principles we propose can be actively enforced. However, it is important to raise awareness within the community of the implications of the data sets created and how the creation process and thus the usability and value of the data sets can be improved taking into account the proposed principles and the respective properties. Considerable progress has already been made with the introduction of DOIs for data sets so that they can be properly cited and researchers who contribute valuable data are recognized.

To promote adherence to these principles, the peer review process for research articles could include an assessment of data sets, including their metadata if the work is data-dependent, to ensure that the data meet established quality and documentation standards. Repositories that publish data sets could support these efforts by publishing guidelines and requirements for publishing data sets, possibly incorporating community-defined documentation standards, such as those discussed by Horsman (2024). In addition, conferences and organizations could incentivise the creation of high-quality data by introducing awards or quality seals for exemplary data sets to promote a culture of excellence and encourage the widespread adoption of best practices.

8. Conclusion and future work

Although there is a consensus in the community about the importance of making high-quality data sets publicly available within the field, projects such as NIST's *CFReDS* platform have recently been updated, and an increasing number of data synthesis frameworks have been published, there is still a distinct lack of standards and best practices for data set creation and description. At the same time, cyber incidents and threats have evolved significantly in recent years, changing the expectations of the community and other parties (e.g., funding agencies) for the availability and reliability of data sets as they are used in research, machine learning, training and education, tool testing and development.

This work identified major characteristics and properties of data sets used in digital forensics. First, we analysed the literature for data set (generation) characteristics, desired properties, and typical issues with data sets. We then complemented existing work by defining principles and associated properties that we believe data sets must satisfy to be valuable, useful, and applicable to critical tasks such as the validation of security-critical systems and software, or the education and training of the cybersecurity workforce. A peculiarity is that we have addressed the question of what properties a data set must have to be useful, depending on the specific use case for which the data is needed. We also considered how the data were created, i.e., its origin. The respective principles and properties were outlined and discussed to assist those who create them and to ensure that the data sets provide maximum value to the cybersecurity domain.

In the long term, we hope that the community will take our suggestions into account when creating data sets and that our discussions will lead to more and better-quality data sets being created, appropriately documented and published in the future. As a direct next step, we plan to evaluate published data synthesis frameworks in the field. Based on the proposed data set properties and characteristics, it is now possible to fairly evaluate these frameworks and thoroughly assess the data sets they produce.

CRediT authorship contribution statement

Thomas Göbel: Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Frank Breitinger:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Harald Baier:** Writing – review & editing, Supervision, Conceptualization.

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.

Data availability

No data was used for the research described in the article.

References

Abt, S., Baier, H., 2014. Are we missing labels? A study of the availability of ground-truth in network security research. In: 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), pp. 40–55.

Baggili, I., Breitinger, F., 2015. Data sources for advancing cyber forensics: what the social world has to offer. In: 2015 AAAI Spring Symposium Series.

Breitinger, F., Jotterand, A., 2023. Sharing datasets for digital forensic: a novel taxonomy and legal concerns. Forensic Sci. Int. Digit. Invest. 45, 301562. https://doi.org/10.1016/j.fsidi.2023.301562.

Carrier, B., 2010. Digital forensics tool testing images. http://dftt.sourceforge.net. (Accessed 3 November 2024). Online.

Ceballos Delgado, A.A., Glisson, W.B., Grispos, G., Choo, K.K.R., 2021. Fade: a forensic image generator for Android device education. WIREs Forensic Sci. 4, e1432. https://doi.org/10.1002/wfs2.1432. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wfs2.1432.

Davies, S.R., Macfarlane, R., Buchanan, W.J., 2021. Exploring the need for an updated mixed file research data set. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET), pp. 1–5.

Davies, S.R., Macfarlane, R., Buchanan, W.J., 2022. Napierone: a modern mixed file data set alternative to govdocs1. Forensic Sci. Int. Digit. Invest. 40, 301330. https://doi.org/10.1016/j.fsidi.2021.301330.

Demmel, M., Göbel, T., Gonçalves, P., Baier, H., 2024. Data synthesis is going mobile—on community-driven dataset generation for Android devices. Digit. Threats, 5. https://doi.org/10.1145/3688807.

Du, X., Hargreaves, C., Sheppard, J., Scanlon, M., 2021. Tracegen: user activity emulation for digital forensic test image generation. Forensic Sci. Int. Digit. Invest. 38, 301133. https://doi.org/10.1016/j.fsidi.2021.301133.

Fragg, M., 2014. Forgeosi. https://github.com/maxfragg/ForGeOSI. (Accessed 3 November 2024). Online.

Garfinkel, S., 2007. Forensic Corpora: a Challenge for Forensic Research. Electronic Evidence Information Center, pp. 1–10.

Garfinkel, S., 2012. Lessons learned writing digital forensics tools and managing a 30 TB digital evidence corpus. In: The Proceedings of the Twelfth Annual DFRWS Conference. Digit. Investig. 9, S80–S89. https://doi.org/10.1016/j.diin.2012.05.002.

Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. In: The Proceedings of the Ninth Annual DFRWS Conference. Digit. Investig. 6, 2–11. https://doi.org/10.1016/j.diin.2009.06.016.

Garfinkel, S.L., 2010. Digital forensics research: the next 10 years. In: The Proceedings of the Tenth Annual DFRWS Conference. Digit. Investig. 7, S64–S73. https://doi.org/10.1016/j.diin.2010.05.009.

Gloe, T., Böhme, R., 2010. The 'Dresden Image Database' for benchmarking digital image forensics. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1584–1590.

Göbel, T., Baier, H., Breitinger, F., 2023. Data for digital forensics: why a discussion on "how realistic is synthetic data" is dispensable. Digit. Threats, 4. https://doi.org/10.1145/3609863.

Göbel, T., Baier, H., Türr, J., 2025. Usable and assessable generation of forensic data sets containing anti-forensic traces at the filesystem level. In: Kurkowski, E., Shenoi, S. (Eds.), Advances in Digital Forensics XX. Springer International Publishing, Cham.

Göbel, T., Baier, H., Wolf, D., 2024. Scenario-based data set generation for use in digital forensics: a case study. In: INFORMATIK 2024. Gesellschaft für Informatik e.V, Bonn, pp. 355–370.

Göbel, T., Maltan, S., Türr, J., Baier, H., Mann, F., 2022. Fortrace - a holistic forensic data set synthesis framework. Forensic Sci. Int. Digit. Invest. 40, 301344. https://doi.org/10.1016/j.fsidi.2022.301344. Selected Papers of the Ninth Annual DFRWS Europe Conference.

Göbel, T., Schäfer, T., Hachenberger, J., Türr, J., Baier, H., 2020. A novel approach for generating synthetic datasets for digital forensics. In: Peterson, G., Shenoi, S. (Eds.), Advances in Digital Forensics XVI. Springer International Publishing, Cham, pp. 73–93.

Gonçalves, P., Dološ, K., Stebner, M., Attenberger, A., Baier, H., 2022. Revisiting the dataset gap problem – on availability, assessment and perspective of mobile forensic corpora. Forensic Sci. Int. Digit. Invest. 43, 301439. https://doi.org/10.1016/j.fsidi.2022.301439.

Grajeda, C., Berrios, J., Benzo, S., Ogunwobi, E., Baggili, I., 2023. Expanding digital forensics education with artifact curation and scalable, accessible exercises via the Artifact Genome Project. Forensic Sci. Int. Digit. Invest. 45, 301566. https://doi.org/10.1016/j.fsidi.2023.301566.

Grajeda, C., Breitinger, F., Baggili, I., 2017. Availability of datasets for digital forensics — and what is missing. Digit. Investig. 22, S94–S105. https://doi.org/10.1016/j.diin.2017.06.004.

Grajeda, C., Sanchez, L., Baggili, I., Clark, D., Breitinger, F., 2018. Experience constructing the artifact genome project (agp): managing the domain's knowledge one artifact at a time. Digit. Investig. 26, S47–S58. https://doi.org/10.1016/j.diin.2018.04.021.

Horsman, G., 2019. Tool testing and reliability issues in the field of digital forensics. Digit. Investig. 28, 163–175. https://doi.org/10.1016/j.diin.2019.01.009.

Horsman, G., 2024. A template for creating and sharing ground truth data in digital forensics. J. Forensic Sci. https://doi.org/10.1111/1556-4029.15524. n/a, https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.15524.

Horsman, G., Lyle, J.R., 2021. Dataset construction challenges for digital forensics. Forensic Sci. Int. Digit. Invest. 38, 301264. https://doi.org/10.1016/j.fsidi.2021.301264.

Hughes, N., Karabiyik, U., 2020. Towards reliable digital forensics investigations through measurement science. Wiley Interdiscip. Rev. Forensic Sci., e1367. https://doi.org/10.1002/wfs2.1367.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020. Celeb-df: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Luciano, L., Baggili, I., Topor, M., Casey, P., Breitinger, F., 2018. Digital forensics in the next five years. In: Proceedings of the 13th International Conference on Availability, Reliability and Security. Association for Computing Machinery, New York, NY, USA.

Lukner, M., Göbel, T., Baier, H., 2022. Realistic and configurable synthesis of malware traces in windows systems. In: Peterson, G., Shenoi, S. (Eds.), Advances in Digital Forensics XVIII. Springer International Publishing, Cham, pp. 21–44.

Michel, M., Pawlaszczyk, D., Zimmermann, R., 2022. Autopod-mobile—semi-automated data population using case-like scenarios for training and validation in mobile forensics. Forensic Sci. 2, 302–320. https://doi.org/10.3390/forensicsci2020023. https://www.mdpi.com/2673-6756/2/2/23.

Mombelli, S., Lyle, J.R., Breitinger, F., 2024. Fairness in digital forensics datasets' metadata – and how to improve it. In: dFRWS EU 2024 - Selected Papers from the 11th Annual Digital Forensics Research Conference Europe. Forensic Sci. Int. Digit. Invest. 48, 301681. https://doi.org/10.1016/j.fsidi.2023.301681.

Nemetz, S., Schmitt, S., Freiling, F., 2018. A standardized corpus for sqlite database forensics. Digit. Investig. 24, S121–S130. https://doi.org/10.1016/j.diin.2018.01.015.

NIST, 2023a. The cfreds project. https://www.cfreds.nist.gov/. (Accessed 3 November 2024). Online.

NIST, 2023b. Computer forensics tool testing program (cftt). https://www.nist.gov/itl/ssd/software-quality-group/computer-forensics-tool-testing-program-cftt. (Accessed 3 November 2024). Online.

OSAC Digital Evidence Subcommittee Task Group on Dataset Development, 2022. The organization of scientific area committees for forensic science (osac) - guidelines for dataset development v2. https://www.nist.gov/system/files/documents/2022/12/15/OSAC-DE-Guidelines%20for%20Dataset%20Development.pdf. (Accessed 3 November 2024). Online.

Park, J., 2018. Trede and vmpop: cultivating multi-purpose datasets for digital forensics – a windows registry corpus as an example. Digit. Investig. 26, 3–18. https://doi.org/10.1016/j.diin.2018.04.025.

Park, J., Lyle, J.R., Guttman, B., et al., 2016. Introduction to cftt and cfreds projects at nist. J. Korea Inst. Inf. Security Cryptogr. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=921807.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M., 2019. Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Roussev, V., 2011. An evaluation of forensic similarity hashes. In: The Proceedings of the Eleventh Annual DFRWS Conference. Digit. Investig. 8, S34–S41. https://doi.org/10.1016/j.diin.2011.05.005.

Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J.N., Sheppard, J., 2023. Chatgpt for digital forensic investigation: the good, the bad, and the unknown. Forensic Sci. Int. Digit. Invest. 46, 301609. https://doi.org/10.1016/j.fsidi.2023.301609.

Scanlon, M., Du, X., Lillis, D., 2017. EviPlant: an efficient digital forensic challenge creation, manipulation and distribution solution. In: dFRWS 2017 Europe. Digit. Investig. 20, S29–S36. https://doi.org/10.1016/j.diin.2017.01.010.

Schmidt, L., Kortmann, S., Hupperich, T., 2023. Improving trace synthesis by utilizing computer vision for user action emulation. Forensic Sci. Int. Digit. Invest. 45, 301557. https://doi.org/10.1016/j.fsidi.2023.301557.

Schmitt, S., 2018. Introducing anti-forensics to sqlite corpora and tool testing. In: 2018 11th International Conference on IT Security Incident Management & IT Forensics (IMF), pp. 89–106.

Shullani, D., Fontani, M., Iuliani, M., Shaya, O.A., Piva, A., 2017. Vision: a video and image dataset for source identification. EURASIP J. Inf. Secur. 2017, 1–16.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Niessner, M., 2016. Face2face: real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Vidas, T., 2011. Memcorp: an open data corpus for memory analysis. In: 2011 44th Hawaii International Conference on System Sciences, pp. 1–6.

Visti, H., 2015. Forge - forensic test image generator v2.1. https://github.com/hannuvisti/forge. (Accessed 3 November 2024). Online.

Voigt, L.L., Freiling, F., Hargreaves, C.J., 2024. Re-imagen: generating coherent background activity in synthetic scenario-based forensic datasets using large language models. In: dFRWS APAC 2024 - Selected Papers from the 4th Annual Digital Forensics Research Conference APAC. Forensic Sci. Int. Digit. Invest. 50, 301805. https://doi.org/10.1016/j.fsidi.2024.301805.

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.

Wolf, D., Göbel, T., Baier, H., 2024. Hypervisor-based data synthesis: on its potential to tackle the curse of client-side agent remnants in forensic image generation. In: dFRWS EU 2024 - Selected Papers from the 11th Annual Digital Forensics Research Conference Europe. Forensic Sci. Int. Digit. Invest. 48, 301690. https://doi.org/10.1016/j.fsidi.2023.301690.

Woods, K., Lee, C.A., Garfinkel, S., Dittrich, D., Russell, A., Kearton, K., 2011. Creating realistic corpora for security and forensic education. In: Proceedings of ADFSL Conference on Digital Forensics, Security and Law, pp. 123–134.

Yannikos, Y., Steinebach, M., Graner, L., Winter, C., 2014. Data corpora for digital forensics education and research. In: Peterson, G., Shenoi, S. (Eds.), Advances in Digital Forensics X. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 309–325.

Zheng, M., Robbins, H., Chai, Z., Thapa, P., Moore, T., 2018. Cybersecurity research datasets: taxonomy and empirical analysis. In: 11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18). USENIX Association, Baltimore, MD. https://www.usenix.org/conference/cset18/presentation/zheng.