Forensic Science International: Digital Investigation 48 (2024) 301681

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

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

DFRWS EU 2024 - Selected Papers from the 11th Annual Digital Forensics Research Conference Europe

FAIRness in digital forensics datasets' metadata – and how to improve it

Samuele Mombelli [a], James R. Lyle [b], Frank Breitinger [a,*]

[a] School of Criminal Justice, Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, 1015, Lausanne, Switzerland
[b] National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD, 20899, United States

ARTICLE INFO

Keywords:
Digital forensics datasets
FAIR principles
CFReDS
Sharing practices
Metadata
Repository
Survey

ABSTRACT

The availability of research data (datasets) and compliance with FAIR principles—Findability, Accessibility, Interoperability, and Reusability—is critical to progressing digital forensics. This study evaluates metadata completeness and assesses the alignment with the FAIR principles using all 212 datasets from NIST's Computer Forensic Reference DataSet Portal (CFReDS). The findings underscore deficiencies in metadata quality and FAIR compliance, emphasizing the need for improved data management standards. Based on our critical review, we then propose and discuss various approaches to improve the status quo.

1. Introduction

Data and its sharing are crucial for science. To help data creators and publishers improve the digital discovery and sharing of research data, Wilkinson et al. (2016) introduced FAIR, a quartet of guiding principles (Findability, Accessibility, Interoperability, and Reusability) designed to enhance the digital discovery and distribution of research data. These guidelines aim to promote the creation of datasets during the research process and their sharing with the scientific community, as well as to standardize their structure and accompanying documentation (through comprehensive metadata) to be transparent and produce reproducible and reusable results. This is a strong push towards an increasingly open and collaborative science.

A year later, a study by Grajeda et al. (2017) revealed that the digital forensics community often follows poor practices and that researchers do not share their datasets. Consequently, the authors proposed a "single, centralized, curated, well-organized repository" which shall encourage researchers to link their datasets. Following this study, the community has seen positive trends in the sharing of datasets which is mostly due to online repositories (Findability). For instance, the new Computer Forensic Reference DataSet Portal (CFReDS v2.0), a centralized repository that aims to enable the community to find and share datasets (Park et al., 2016), was released in 2021.

While there are now several well-known locations providing datasets that help to increase the Findability aspects of FAIR, the remaining aspects (Accessibility, Interoperability, and Reusability) have not received significant attention. Metadata standards for datasets, which serve as specific implementations of the FAIR Principles, exist for some disciplines, but digital forensics is currently not one of them (Garfinkel et al., 2009; Horsman and Lyle, 2021). One challenge is that the majority of data in this field is semi-structured or unstructured (Breitinger and Jotterand, 2023), a factor that complicates metadata standardization. Consequently, this study was conceived and developed to assess the situation and examine the following three questions:

RQ1: To what extent is the metadata available and comprehensive for digital forensics datasets?
RQ2: How well does digital forensics dataset metadata comply with the FAIR principles?
RQ3: What strategies and approaches can enhance the compliance of digital forensics dataset metadata with the FAIR principles?

To answer these questions, we examined all 212 datasets referenced in the CFReDS Portal as of January 2023. In our analysis, we looked for specific criteria that we believe each dataset should have. This empirical analysis allowed us to address the above questions and conclude that the community is currently using poor practices. In summary, our work provides the following contributions:

- Identification of Metadata Deficiencies: By systematically reviewing and assessing the metadata of published datasets, we provide valuable insights into the metadata quality. This contributes to the

* Corresponding author.
E-mail addresses: samuele.mombelli@unil.ch (S. Mombelli), james.lyle@nist.gov (J.R. Lyle), frank.breitinger@unil.ch (F. Breitinger).
URL: https://www.FBreitinger.de (F. Breitinger).

https://doi.org/10.1016/j.fsidi.2023.301681

Available online 15 March 2024

broader understanding of data quality and the challenges associated with metadata in the research community.
- Recommendations for Metadata Improvement: We present a set of recommendations for enhancing metadata quality. These recommendations can serve as a practical guide for dataset creators, curators, and data users to improve the quality and completeness of metadata associated with datasets.

2. Background and related work

This section first defines some key terms (Sec. 2.1), then highlights challenges for data sharing (Sec. 2.2), followed by an introduction to The FAIR Principles. Next, we briefly discuss several existing Data repositories and archives and go into more detail about The CFReDS Portal and The Computer Forensic Tool Testing (CFTT) Project.

2.1. Definitions

Data is a fundamental concept for any scientific field because it embodies the "representation of facts, concepts, or instructions in a manner suitable for communication, interpretation, or processing by humans or by automatic means" (National Institute of Standards and Technology, 2022b). It allows information to be collected, organized, structured, analyzed, and interpreted, thus enabling the constant development of scientific knowledge. Data elements can be grouped to form a *dataset*. In the context of digital forensics, a dataset is defined as "a collection of files or data (obtained from a digital device) created to have a desired set of attributes and known content" (OSAC Digital Evidence Subcommittee Task Group on Dataset Development, 2022, p. 1). As a result of its polymorphic nature, the data contained in a dataset cannot be understood without it being associated with a complete and comprehensive description, in the form of metadata.

According to Breitinger and Jotterand (2023), the term *metadata* has various meanings. In the context of digital forensics, metadata is additional data available for analysis, such as EXIF information. In the context of data and datasets, metadata describes the data object. Metadata may also be used as a synonym for descriptive information that helps locate a dataset within a larger repository. For this work, the term metadata is defined as any information describing the dataset and is further discussed in Sec. 6.1.

Metadata must be sufficiently comprehensive to allow a user to determine whether a particular dataset is suitable for use in their analysis (FORCE11, 2020). Furthermore, "a well-documented dataset facilitates more rigorous testing and reliable results. […] Thorough documentation of the dataset and dataset development process is critical to reliable and effective use of the data" (OSAC Digital Evidence Subcommittee Task Group on Dataset Development, 2022, pp. 1–2).

2.2. Data sharing challenges

Data generation and management in digital forensics is a domain that suffers from a variety of problems that have been discussed by several authors over the years. These difficulties can be summarized in three categories:

Creation of datasets: According to Horsman and Lyle (2021) and OSAC Digital Evidence Subcommittee Task Group on Dataset Development (2022), the community needs to engage in the creation of appropriate and quality datasets that can be used for multiple purposes, including testing tools and methods, training and education, research, or machine learning (Göbel et al., 2023). These datasets should be constructed in an accurate and structured manner and should be well described so that repeatable conclusions can be drawn once the data is used. Defining a 'complete' ground truth is difficult (impossible) for large datasets (Roussev, 2011).

Standards: According to Garfinkel et al., 2009 and Horsman and Lyle (2021), there is a shortage of standardized datasets. This is due to the lack of standardized metadata and schemas that govern the creation and description of datasets and the elements within them.

Release and sharing of datasets: According to Grajeda et al. (2017) and Horsman and Lyle (2021), many researchers do not share their datasets. Grajeda et al. (2017) identified four reasons: (1) they may not possess the capacity to do so (e.g., due to a lack of resources), (2) they may be concerned with privacy and proprietary rights, (3) they may not have considered how crucial data sharing is, and (4) they may choose not to share datasets due to concerns about protecting their intellectual property.

2.3. The FAIR principles

If data elements are the fundamental building blocks upon which science can evolve, the cyclicality with which this information is produced and used must be properly managed. This is especially true in modern science, where the sheer volume of data being generated poses a significant challenge (Starr et al., 2015). There is a growing need to develop a unified approach to improve the transparency of data, its openness, its reuse, and the reproducibility of the results it leads to (Starr et al., 2015; Wilkinson et al., 2016). To this end, Wilkinson et al. (2016) proposed four basic principles that, if carefully followed, can enable humans and machines to discover, evaluate, and reuse the information generated by one research effort for the next, thus contributing to the fundamental cyclicality of science.

Fig. 1 details these four related but independent principles—known as FAIR Principles—which allow data and metadata to be (GO FAIR, 2022):

Findable: Finding data is the first step in using and reusing it. For easy access by both humans and automated systems, metadata and data should be easily findable. Therefore, machine-readable metadata is essential for the automatic discovery of datasets and services, along with a unique identifier that ensures that data can be found globally and persistently.

Accessible: Upon locating the needed data, the user must be provided with information on how to access it, which may involve authentication and authorization procedures. This information must remain accessible even if the data is no longer available, along with the rest of the metadata.

Interoperable: All the information needed to integrate the dataset with other data elements and enable interoperability at multiple levels (analysis, storage, and processing) must be present and universally comprehensible.

Reusable: The fundamental aim of the FAIR principles is to maximize the reusability of data. This requires a comprehensive description of both data elements and metadata to enable their replication and/or integration in multiple contexts. The presence of a license governing the use of the data is a key feature to allow data reuse, and thus data FAIRness.

2.3.1. Unique identification and licensing

There are two elements without which the FAIR principles cannot be implemented: a unique, global, and persistent identifier for the information (a prerequisite for Findability) and the presence of an associated license (a prerequisite for Reusability), which are briefly described here.

First, access to data and metadata must be ensured by an identifier, i.e., a "unique name given to an object, property, set or class" (Juty et al., 2020, p. 31), in a way that is resolvable by machines and humans on the Web using a free and open protocol (e.g., HTTPS) (Starr et al., 2015). This identifier must be accompanied by a minimal amount of descriptive metadata to assist the user in discovering and verifying the information (Juty et al., 2020). There are many types of identifiers and structures to

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Fig. 1. Overview of the FAIR principles from Wilkinson et al. (2016). The term '(meta)data' refers to both data and metadata.

follow, but their basic purpose always remains to enable the discoverability and accessibility of a specific resource. A more detailed approach and implementation of these identifiers is discussed by Starr et al. (2015) and Juty et al. (2020). One type of identifier that needs to be briefly discussed is the Digital Object Identifier (DOI), the most common type used by the scientific community. It is a stable resource identification mechanism that can be freely accessed via the `prefix` https://doi.org/, followed by the identifier itself. The registration of a DOI and its association with an entity (e.g., article, dataset) is done by Registration Agencies (RAs) at the request of a publisher or repository. From the researcher's perspective, obtaining a DOI is primarily a matter of publishing one's work through an authority that provides this option.

Once the information is found, accessed, and the content acquired, the reusability of the information is the key principle for closing the research loop: in this respect, the presence of a license that explicitly states how the data it protects can be used is crucial (Labastida and Margoni, 2020). There are many options to choose from, and a custom waiver is also an alternative, but the important point is that at least one of them is explicitly associated with the data and metadata. A more detailed approach by Labastida and Margoni (2020) develops the legal context and suggests specific licensing options that can be implemented. If no license is provided, data may be protected by other laws and thus is not useable (Breitinger and Jotterand, 2023).

2.3.2. *Landing pages and FAIR repositories*

Good data and metadata management also includes the use of repositories (or landing pages) that support and enable all the functionalities necessary to host FAIR data and metadata.

According to Starr et al. (2015), the identifier associated with the data should point to a landing page rather than to the data itself, because: (1) metadata could potentially be accessed for longer than data, and it is necessary to have a landing page that provides continuous access to metadata; (2) a landing page enables access to metadata even for users who do not have permission to access the actual data (if access to the data is restricted); and (3) a landing page provides a point of access to the data that is independent of its encoding, making it easier to understand and interact with. The authors then suggest a list of elements that must be present on these landing pages, how they must be encoded, what protocols must be implemented to access them, and how their persistence must be ensured. Similarly, the Swiss National Science Foundation (2017) defines a set of minimum criteria that repositories must meet to be considered FAIR-compliant (one requirement to obtain funding).

2.4. *Data repositories and archives*

Although some authors distinguish between repositories and archives (repositories are considered more temporary, while archives promise long-term availability), in this article we use the terms as synonyms.

As a result of the increasing amount of data and the importance of sharing it, many repositories (often domain-specific) have been created over time. To archive data and guarantee availability to a specific community, one can upload data to highly specialized repositories such as GenBank or the Worldwide Protein Data Bank (wwPDB). Because these repositories are established in their field, findability is often ensured. If a specialized repository is not available, general-purpose data repositories such as `dataverse.org`, `FigShare.com`, `Zenodo.org`, `DataHub.io`, `SWISSUbase.ch`, or `EUDat.eu` can be used. Due to the generic nature of these repositories, the possibilities to describe the data using the provided submission forms are less specific and domain-standard.

As an illustrative example, we refer to SWISSUbase,[1] which is presented as "a comprehensive platform for the curation, preservation, and dissemination of scientific research data and metadata" (SWISSUbase, 2023). It is a FAIR-compliant portal, recognized by the Swiss National Science Foundation, dedicated to the collection and organization of research data and associated metadata. The publication of new research datasets is gradually supported by the system, which requires complete metadata structured according to the FAIR principles: it asks for a title, a language, and a license, as well as some optional metadata (e.g., dataset description, documentation notes, etc.). It also allows for the reservation of a unique identifier (DOI) that will be associated with the uploaded data.

Depending on the repository, sensitive data may be uploaded and access restrictions may be applied. For example, SWISSUbase requires the user to attest, during the upload process, that the anonymization of the data has been carried out as completely as possible.

From a digital forensics perspective, a downside of these repositories is that they often limit the size of the dataset. For instance, Zenodo states that it accepts up to 50 GB per dataset (one can upload multiple datasets) and FigShare accepts files up to 20 GB. SWISSUbase does not list a

[1] https://www.swissubase.ch/.

maximum file size but informs the users that they aim to also support large datasets.[2] While these quotas are reasonable for most domains, digital forensics research produces very large datasets.

As many of these repositories exist, `re3data.org`, a registry of research data repositories, can be used to help identify appropriate repositories.

2.5. The CFReDS portal

With respect to digital forensics, the most comprehensive repository is the Computer Forensic Reference DataSet Portal (CFReDS) which is maintained by the National Institute of Standards and Technology (2022a) (NIST) and provides a large collection of datasets (Park et al., 2016). Here, dataset refers to digital information that has been generated for a specific purpose, either randomly or collected from digital devices in the physical world. The purposes that these datasets can serve include testing tools, studying tool behavior, training general practitioners, educating, and conducting proficiency tests (Park et al., 2016; Horsman and Lyle, 2021; National Institute of Standards and Technology, 2022a).

CFReDS includes data provided by NIST but also allows users to freely upload datasets or link datasets that are already hosted in other (external) repositories. In either case, adding a dataset requires filling out several descriptive fields that serve as metadata for the dataset: year of creation, title (10–100 chars), short description (1–200), and long description (0–1000)[3] as well as details about the uploader (name, institution, and email address). In a subsequent step, tags can be added to the dataset to ease searches. Tags are based on a hierarchical taxonomy that allows a detailed classification and characterization. As an example, the first hierarchical sub-level for each of the three root tags is as follows:

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; etc. (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. Some examples of useful tags for digital forensic classification are: forensic functions (e.g., Acquisition, File Carving, String Search), Operating Systems, File Systems, and application data (list of application artifact sets present, e.g., Facebook, Google Maps).

Fig. 2 provides an absolute number of datasets released each year.

2.6. The Computer Forensic Tool Testing (CFTT) project

The objective of NIST's CFTT is to create a standardized framework for assessing computer forensic software tools. This framework encompasses the development of comprehensive tool specifications, testing procedures, evaluation criteria, test datasets, and requisite testing equipment. CFTT datasets are then used internally by NIST researchers, as well as forensic tool vendors, academia, etc. The datasets provided with CFTT focus on measuring three things: tool behavior, practitioner ability to perform a forensic task, or the behavior of applications in a software environment. These datasets can be either static or created on-the-fly. Some examples of both types of datasets:

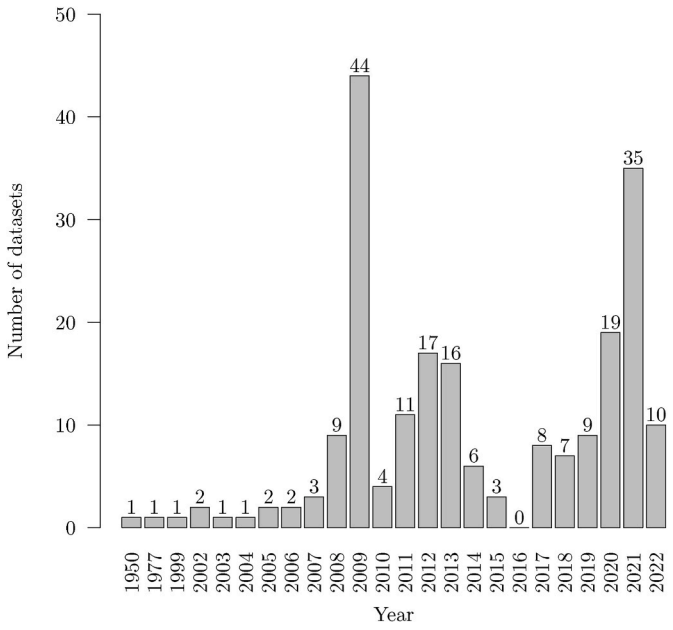

Fig. 2. Evolution of the number of datasets published on the CFReDS according to the year of creation (National Institute of Standards and Technology, 2022a).

Disk imaging: uses an on-the-fly dataset[4] created by running a program to write unique content to each sector of a test disk. A tool is provided to perform a byte-by-byte comparison of the original source to the captured data, characterizing the differences between the two.
Mobile Device and Cloud: uses an on-the-fly dataset created by tracking a detailed sequence of user actions to populate a mobile device and cloud services.
String Search: uses a static image created on a removable drive. The removable drive is first erased (`0x00`), then partitions are created for file systems of interest to the test, files with search targets are copied to each partition, and other actions are taken to prepare the required state of the dataset.

To create a test report for a tool or technique, other elements—such as requirements that the tool must meet, a set of test cases to evaluate the requirements, and a dataset for each requirement—may be needed beforehand.

3. Methodology

It is important to note, as stated by Wilkinson et al. (2016), that the FAIR principles consist of a series of advanced ideas that are not fully prepared for practical application. They can be seen as a guide to assessing the findability, accessibility, interoperability, and reusability of a single data object or collection. Thus, to answer our research questions and to capture the metadata associated with the datasets, it is necessary to develop a customized implementation of these principles that is both comprehensive and easily operationalizable. The methodology consisted of three steps:

Developing a checklist

The first step involved developing a list of criteria in the form of a checklist that would be as comprehensive as possible. This was done by

[2] We asked if they would be willing to accommodate a 5 TB dataset and received a positive response.

[3] Numbers in parentheses define the minimum and maximum number of characters per field.

[4] Dataset generated by following a detailed sequence of user actions or by running a program to establish a state of interest on a device (e.g., hard drive, mobile device, or remote storage such as a cloud service).

reviewing various references/resources on FAIR principles and their implementation, mainly of three types: (1) high-level definitions and conceptualizations, (2) general or domain-specific implementations, and (3) data management reports and guidelines. The resulting list of criteria is presented in Appendix A. As the outcome was too comprehensive, various elements were combined to reduce the checklist to a reasonable number of criteria which are presented in Sec. 4.

Collecting data

As an operational implementation of these criteria, a form was developed to allow for quick and standardized collection of the various metadata.[5] Along with this data, a personal comment, if any, and the timestamp of the dataset analysis were recorded. The data collection was done manually and findings were stored in a dedicated MySQL database. The resulting dataset is publicly accessible (Mombelli et al., 2023).

In several cases, it was necessary to perform an additional analysis of the dataset structure, its context (i.e., the specifics of the repository in which it was hosted), its accessibility, and its content to complete the contextual information describing the data.

Analyzing results

Having all the details in a database allowed us to query it and retrieve multiple statistics, which are summarized in Sec. 5.

4. Checklist development

The aim was to find the best possible compromise between time spent on collection and the quality of data collected. With this goal in mind, we reduced the list and collected the information for each dataset (some criteria were recorded twice for data found in various locations). As the CFReDS Portal was used for this article, some criteria directly relate to it.

In general, we made the following two distinctions: (1) data and metadata directly accessible through the CFReDs Portal vs. data and metadata located on external resources (e.g., another website), and (2) information related to the dataset (general criteria) vs. information related to the metadata (metadata-related criteria).

The first five criteria relate to the *general information* found directly within CFReDS. This information was obtained by examining the short/long description:

Dataset title: Datasets have a unique title, which often is a summary of the content or description used by the community as an identifier (e.g., the M57 Patents Scenario[6]). A title is a mandatory field within the CFReDS Portal.

Presence of a persistent identifier: The presence of a globally unique identifier associated with the dataset was recorded, as was its type, based on a non-exhaustive list inspired by Starr et al. (2015): Archival Resource Key (ARK), Digital Object Identifier (DOI), Handle System (HDL), Persistent Uniform Resource Locator (PURL), and Other.

Presence of a license: The presence of a license associated with the dataset was recorded, as was its type, based on a non-exhaustive list inspired by Labastida and Margoni (2020) and The FAIRsharing team (2023): Apache License, BSD License, Creative Commons (CC), ISO Privacy and Copyright, GNU General Public License (GPL), Open Data Commons License, W3C Document License, Other, and Unspecified type.

Dataset repository type: The repository in which the dataset was hosted was identified and recorded. We differentiated between the following four types: Cloud Storage Provider (CSP) (e.g., Google Drive, Dropbox), institutional repository (e.g., university, company, association), personal environment (e.g., personal website), and public repository (e.g., GitHub, SWISSUbase, NIST, Digital Corpora). This list has been developed from personal knowledge in an attempt to be comprehensive. However, the categories presented are not necessarily mutually exclusive, so an evaluation specific to each case was necessary.

Presence of an integrity value (hash): The presence of an integrity value (one or more hash values) was recorded.

The next two criteria focus on the *metadata-related information* found within CFReDS:

Metadata scope: This criterion is recorded as an integer value that represents the total length in characters (excluding spaces) of the dataset name, the dataset description (long + short), and any other descriptive metadata contained in the dataset that is directly accessible via CFReDS. The motivation of this criteria is that longer descriptions are generally more exhaustive and, thus, FAIR-compliant. Of course, this does not say anything about the quality, which is discussed later.

Associated tags (based on CFReDS taxonomy): Tags associated with the dataset were recorded to document the findability of the data and metadata, and the indexing by search keywords.

This next section of criteria focuses on external data that relates to a dataset. For instance, in case the description within CFReDS is kept minimal but there is comprehensive documentation on a project website or external repository.

Dataset and metadata availability: We recorded whether the dataset and metadata were still available. This criterion allowed us to understand how many of the datasets were still available, but more importantly, if the metadata was still present after a dataset disappeared. For example, the metadata on CFReDS may still be accessible while the GitHub repository (dataset and/or metadata) has disappeared.

External metadata: When we found a reference to external metadata, we analyzed that information and recorded the five criteria listed under 'general information' (explained above: dataset title, presence of a persistent identifier, presence of a license, dataset repository type, and presence of an integrity mechanism) as well as the metadata scope referring to external metadata.

Presence of machine-readable metadata: The presence of machine-readable and open-format metadata associated with the dataset was recorded, as was its format based on a non-exhaustive list inspired by Starr et al. (2015): Comma Separated Values (CSV), JavaScript Object Notation (JSON), Resource Description Framework (RDF), Extensible Markup Language (XML), and Other. To find this information, we analyzed both the landing page and the dataset itself (e.g., a dataset may be zipped and contain a description inside the zip file). Note: The CFReDS Portal is not considered machine-readable, but if we found machine-readable metadata inside the dataset, we still considered it accessible from the CFReDS Portal.

Regardless of where and how the dataset was hosted, we also captured if the dataset was directly accessible or required additional steps:

Accessibility: If an additional step was required to access the dataset, we captured what had to be done. An example would be to register or request credentials.

5. Data collection results

We conducted the survey between December 21, 2022, and January 28, 2023. A total of 212 datasets were listed in the CFReDS Portal out of

[5] A PHP-based webpage was developed allowing to capture the data and store it in a database.

[6] https://digitalcorpora.org/corpora/scenarios/m57-patents-scenario/.

which 26 (12.3 %) were no longer available. The following two subsections summarize our findings.

5.1. Information within CFReDS

As required by CFReDS, every dataset came with a title allowing us to identify it and also weakly reference it in work. In contrast, none of the datasets had a persistent identifier in their metadata. Two datasets included a license. The first mentioned several licenses of multiple types (Apache License, Artistic License (Perl), CC, GNU AGPL, GNU GPL, GNU LGPL (+ unRAR restriction), and several proprietary licenses), while the second mentioned a license of unspecified type. With respect to the repository type, Table 1 shows the distribution between the four categories. Of the 126 datasets hosted by a public repository, 16 were hosted directly by NIST, 105 were listed in Digital Corpora, and 5 were listed in other public repositories. Of all the datasets, 19 included an integrity value.

Metadata contained in CFReDS

As it is difficult to classify the quality of metadata, we decided to use an objective value, i.e., a numerical threshold of 600 characters. The choice of this value was based on the total length possible, which is limited to 200 (short description) plus 1000 (long description) within CFReDS. Consequently, metadata was considered exhaustive if the total number of characters (without spaces) was larger than 600, and non-exhaustive otherwise. However, as this is still somewhat arbitrary, further research is needed to determine a good-metadata-length threshold (if one exists). Only 15.6 % (33/212) of datasets included descriptive metadata longer than 600 characters (i.e., classified as exhaustive); the calculated median for this variable is 195 characters. For the 26 datasets that were no longer available, all metadata was still accessible via the portal.

Two datasets listed on the portal were associated with metadata available in an open machine-readable format (included in the datasets themselves), namely XML and Microsoft Excel Open XML Spreadsheet (XLSX).

Concerning the associated tags, the distribution of datasets under each root entry (see Sec. 2.5) is very imbalanced (note that multiple tags can be given to one dataset):

- 159 under Data/Forensic Related,
- 90 under IT System Type, and
- 20 under Simulated Cases/Scenarios.

For a comprehensive view of the taxonomy and the number of datasets in each category, see Appendix B.

5.2. Information externally available

Out of the 212 datasets, 107 (50.5 %) included an explicit reference to external metadata. Commonly, the title was similar/identical to the CFReDS reference. One dataset came with a persistent identifier of the type Digital Object Identifier (DOI). Of the 107 datasets, 43 included license information in the external metadata. A summary of the different license types is shown in Table 2. The category "unspecified types" summarizes custom descriptions of licenses, e.g., the dataset 'M57 Patents Scenario' contains a license file stating: "Contains information derived from copyrighted materials. For use only for research, education, training only, and the production of educational materials. All other uses require the permission of the copyright holders".

In contrast to the metadata scope within CFReDS, the external metadata was significantly more exhaustive. A total of 67 datasets out of the 107 considered (62.6 %) had descriptive metadata longer than 600 characters (i.e., classified as comprehensive), and the calculated median for this variable was 916 characters. For the 26 unavailable datasets, we were only able to access the metadata of two (7.7 %) in their respective repositories.

A total of eight datasets included metadata in an open machine-readable format:

- 4 datasets had metadata in XLSX format;
- 2 datasets had metadata in JSON format;
- 1 dataset had metadata in XML format;
- 1 dataset had metadata in Extensible Hypertext Markup Language + Resource Description Framework in attributes (XHTML + RDFa 1.0) and JSON formats.

Table 1
Repository types for the datasets listed on the CFReDS Portal.

	Public	CSP	Institutional	Personal
#	126	52	30	4

Table 2
License types found externally. License types not mentioned in the checklist (compare Sec. 4) are included in the "Other" category.

	#
Creative Commons License (CC)	24
Apache License (Apache)	1
General Public License (GPL)	1
Other	8
Unspecified types	9

5.3. Accessibility

A total of 11 datasets (11/212) had access restrictions such as the need to accept a license agreement, request credentials via email, password protection, need to register for a service or dataset set to private.

6. Discussion

In the Introduction we raised three research questions which are discussed in the upcoming paragraphs:

6.1. To what extent is the metadata available and comprehensive for digital forensics datasets?

As mentioned in Sec. 2.1, the term metadata has various meanings. First, *metadata* can refer to a set of information that describes the artifacts present in the dataset in such a way that the results can be verified once the dataset is utilized (e.g., in the case of a string search tool, a list of strings to be used as a test and the specific expected result). This type of metadata needs to be distinguished from information that more generally describes what a particular dataset is about to help a user find a relevant dataset within a collection/repository (e.g., the keywords selected by the CFReDS taxonomy to categorize and describe a dataset). We propose to use the term *category*. During our analysis, these two types were considered as a single set, but distinguishing between them may allow for a more detailed analysis of the availability and comprehensiveness of different types of information. *Availability* refers to the simultaneous assessment of the presence/absence of the various metadata elements and their accessibility, while *comprehensiveness* involves evaluating the degree to which these elements are complete, with some requiring only presence (e.g., a persistent and globally unique identifier) for validation.

A summary of the findings is provided in Table 3. It is obvious that many datasets lack exhaustive metadata especially if they have no external (supplemental) data complementing the information found within CFReDS; if there is an external reference, it is on average more complete and includes additional details, such as a license. We also did

Table 3
Main survey results summarizing the items found within CFReDS (columns 2 and 3) and the metadata found externally (columns 4 and 5).

	CFReDS ($n = 212$)		External ($n = 107$)	
	Total	Percentage	Total	Percentage
Exhaustive metadata	33	15.6 %	67	62.6 %
Persistent identifier	0	0.0 %	1	0.9 %
License	2	0.9 %	43	40.2 %
Machine-readable metadata	2	0.9 %	8	7.5 %
Integrity value (hash)	19	9.0 %	60	56.1 %

not see a positive trend of metadata over time, i.e., regardless of when the dataset was submitted (see Fig. 2), the metadata details—such as length or presence of a license—are similar.

6.1.1. Repositories and usability

The usability of repositories varies widely and is mainly influenced by the internal structure of each repository. Some repositories require users to register and fill out extensive forms that ask for a lot of information, which slows down the process. On the other hand, others are simpler and likely considered to be more user-friendly. This "user-friendliness" of a repository may conflict with its ability to capture detailed metadata and thus indirectly with its FAIRness.

For instance, the CFReDS Portal displays records and associated metadata on a single page. This is a positive feature because it makes the connection between these two elements clear. The metadata is then organized into two main sections, namely *short description* and *long description*, which represent all the information about the dataset. This approach is visually clear, concise, and quick for a user to complete. However, this loose format has its drawbacks: because there is no distinction between metadata elements, searching and filtering are more complicated. In other words, metadata is made understandable to humans, but not to automated systems, which require well-structured information designed in standard formats. In addition, metadata may be incomplete since it is difficult to verify whether free text includes all necessary information, such as a license or a persistent identifier. This is in contrast to repositories that require a lot of detailed information, making it more tedious to submit a dataset.

Note: We are not saying that one is better, but that both have a right to exist. It is important to be aware of these differences and consider them when sharing a dataset. Dataset creators may include additional information within the dataset itself or, more often, link to external resources.

6.1.2. Creator practices

Data shows that researchers may create and share rich metadata if the platform supports or requires it. According to Table 3, researchers are likely to explicitly associate rich descriptive metadata and a license with their dataset when publishing to external sources. A closer look at the metadata, sorted by repository type in Table 4, reveals a mixed situation. It can be concluded that when data is shared only through a Cloud Storage Provider (CSP), it is less likely to contain rich descriptive metadata. This can be explained by the fact that it is easier to upload a dataset without any documentation being created or associated with it. On the other hand, if the data is shared via an online repository or resource, it is likely to be accompanied by some additional information. The presence of machine-readable metadata seems to be mostly independent of the repository type and is generally very limited. The presence of hash values is highly variable, but more favored in public repositories, slightly less in CSPs, and even less in institutional repositories. The presence of a persistent and globally unique (meta)data identifier and the influence of personal environments cannot be commented on due to the insufficient amount of data.

6.1.3. Continuous availability

In terms of dataset availability, based on the data collected, the most persistent repository type is the institutional repository with 6.7 % (2/30) of datasets no longer available, followed by Cloud Storage Provider with 11.5 % (6/52), public repository with 13.5 % (17/126), and personal environment with 25.0 % (1/4). In addition, there were only two cases where metadata was still available after the dataset was rendered inaccessible, and both were found in an institutional repository. These results suggest that the most persistent and reliable repository type is indeed the institutional repository. However, to ensure long-term availability, dataset archives (see Sec. 2.4) should be favored.

6.2. How well does digital forensics dataset metadata comply with the FAIR principles?

This section combines the survey findings with the FAIR principles to identify the strengths and weaknesses of the current community practices.

The survey revealed the following weaknesses:

- non-exhaustive metadata, i.e., less than 600 characters (see Sec. 5) (*FAIR Principles F2 and R1*),
- lack of licenses (*FAIR Principle R1.1*),
- lack of persistent and globally unique identifiers (*FAIR Principles F1, F3, and A1*),
- lack of integrity values and mechanisms (*part of FAIR Principle R1*),
- lack of metadata in machine-readable formats (*FAIR Principle I1*), and
- presence of unqualified references to other metadata (*FAIR Principle I3*).

In terms of strengths, the datasets are registered and indexed in a searchable resource (*FAIR Principle F4*), and the portal demonstrates data persistence and metadata independence from the repository, which makes metadata accessible even if the data is no longer available (*FAIR Principle A2*).

Regarding the use of standards for data and metadata (*FAIR Principles I2 and R1.3*), the specific context of digital forensics currently complicates their proper implementation. Therefore, this criterion is not covered by this consideration.

In some cases, there was a problem with the accessibility of the datasets due to access restrictions. Although this conflicts with the principle of Accessibility (*FAIR Principle A1.2*), a positive aspect is that even in these cases the metadata was accessible and searchable via CFReDS, although the data was not. These specific cases should be evaluated to find a solution that allows, if possible, full access to the data through the portal, or at least to the entirety of the associated metadata.

Table 4
Influence of the repository type on metadata.

Repository type	Exhaustive metadata	Identifier	License	Machine-readable metadata	Hash
CSP ($n = 46$)	28.3 %	0.0 %	10.9 %	6.5 %	52.2 %
Institutional ($n = 28$)	92.9 %	0.0 %	57.1 %	7.1 %	35.7 %
Public ($n = 32$)	81.2 %	3.1 %	9.4 %	9.4 %	78.1 %
Personal ($n = 1$)	100.0 %	0.0 %	0.0 %	0.0 %	100.0 %

6.3. What strategies and approaches can enhance the compliance of digital forensics dataset metadata with the FAIR principles?

We believe that several factors will help with compliance:

6.3.1. Creation of datasets

As outlined in Sec. 2.2, it starts with the creation of datasets, where two references are particularly relevant: (1) OSAC Digital Evidence Subcommittee Task Group on Dataset Development (2022)'s Guidelines for Dataset Development and (2) Horsman and Lyle (2021)'s list of minimum requirements to be followed when creating datasets in digital forensics.

As outlined in Sec. 2.6, it may not always be necessary to release a dataset, but tools that can generate on-the-fly datasets and help evaluate the test results are available. These, when used in combination with appropriate documentation, can be an alternative to sharing datasets.

6.3.2. Creation of metadata

Raising awareness among dataset creators about the importance of metadata is crucial in any context that deals with research and data production. Efforts must be made by all members of the community to overcome major difficulties before the ever-increasing volume of data produced becomes unmanageable.

More specifically, the following two recommendations should be considered: (1) Starr et al. (2015) propose a landing page for a dataset, including guidelines on structure, the information contained, page coding, page links and references, and guarantees of persistence; and (2) the guiding questions formulated by the Swiss National Science Foundation (2017) may be used to assess whether a repository conforms to the FAIR Principles.

There are also several tools—often in the form of checklists—that allow researchers to assess the FAIRness of their research data and metadata. Two examples applicable to digital forensics research (as they are not domain-specific) are the ones developed by Jones and Grootveld (2017) and Australian Research Data Commons (2022).

With respect to the information that should be included, we argue that a released digital forensics dataset should at least have (a):

Dataset description: While it is difficult to describe what makes a good description, a detailed description of the dataset should always be associated with it. The level of detail may vary depending on the type of dataset and use cases (Göbel et al., 2023), but some elements should always be present in a dataset, e.g., title, author, internal structure, scope and evaluation target, possible related publications, and other needed documentation such as tool requirements and test cases. Ideally, new related publications can be added.

Persistent identifier: An identifier that uniquely and globally identifies the dataset should be explicitly associated with the dataset and its metadata. While any widely used and recognized type is valid, we believe that DOI is currently best recognized within the community. This criterion represents Findability.

Publishing the dataset to a public and persistent repository: The dataset should be released and published in a portal or repository that is openly committed to maintaining the persistence of its data and metadata. This should be public, recognized by the community, and provide access to the data and metadata through free, open, and universal protocols. The Swiss Federal Institute of Technology in Lausanne (EPFL) provides a comprehensive list of various repositories.[7] This criterion represents Accessibility.

Explicit and accessible license: The dataset should be accompanied by a user license that specifies how the user may reuse the data and associated metadata. It is preferable to use common, community-approved types of licenses (e.g., CC, GPL), but a custom waiver may suffice. This criterion represents Reusability.

Machine-readable metadata independent of the dataset: The dataset should be accompanied by metadata in machine-readable form (in a free and open format), and the structure of this metadata should follow standards specific to the community. Unfortunately, this structure has not yet been developed for digital forensics.

6.3.3. Repositories

As pointed out in Sec. 2.4, there are various kinds of repositories, where some are highly specialized and others are more generic; some are kept minimalist, requiring only a little mandatory information (we define them as *Link-Repositories*), while others are more complex. The key is to find a good balance between mandatory and optional attributes. Ideally, repositories are flexible, allowing users to add fields (e.g., from a predefined set), and do not limit the length of input fields. In fact, extensive documentation should be encouraged.

Researchers need to consider the repository type (link-repository vs. archive) when making their datasets available. If the chosen repository is only a link-repository, comprehensive metadata should be provided either on a landing page dedicated to the dataset, within the dataset itself (e.g., if uploaded directly to NIST), or by first uploading the dataset to a more complete repository (archive). Nevertheless, digital forensics datasets should be included in the CFReDS Portal to have a central repository which greatly improves findability.

It would certainly be helpful to support the researchers with tools that make the task of releasing and sharing data easier. These could be features built into the portal that allow for metadata standardization, e.g., automatic conversion of metadata into machine-readable formats, help for the researchers to complete missing metadata, automatic provision of a persistent identifier if one is not already associated with the dataset. This would reduce the researcher's workload and thus potentially increase release and sharing rates, in addition to FAIRness. However, research to assess the correlation between these elements should be conducted once the tools are implemented.

Lastly, we should consider the possibility of updating datasets and adding references. For instance, if a dataset has been used in a study that was then published as an article or within a blog entry, knowing about this work can be beneficial for researchers as the new references may include additional findings.

6.3.4. Standards and compliance

As discussed, heterogeneity exists in the nature of datasets and data (e.g., formats, creation context, purpose, use) and in the metadata describing them. This richness is also evident in the distribution of the analyzed datasets in the CFReDS taxonomy (see Appendix B).

Attempting to devise a one-size-fits-all metadata standard would risk oversimplifying this intricate ecosystem, potentially undermining the accurate representation of datasets. Moreover, the multidisciplinary nature of digital forensics, which draws from computer science, law enforcement, and legal domains, among others, further compounds the challenge. A unified metadata standard demands a collective endeavor wherein domain experts, data scientists, legal professionals, and technologists collaborate to develop a framework that accommodates the varying needs and nuances of different subfields. This collaborative approach is indispensable to ensure that the resulting metadata standard captures the breadth and depth of digital forensics datasets, fostering effective data sharing, interoperability, and the overall advancement of the field.

Ensuring that the dataset is shared and that it contains acceptable metadata (e.g., license, description) could be a requirement for publication and thus be validated (enforced) by the reviewers and publication venues.

[7] https://www.epfl.ch/campus/library/services-researchers/data-publication/data-repositories-and-related-platforms/.

7. Limitations

The entire data collection process was performed by one person to reduce inter-operator variability. Some decisions were based on personal interpretations and choices, e.g., the selection of information collected in the metadata scope point. It is important to note that there was considerable diversity in the structure and content of the datasets, the information presented, and the external repositories. This led to some difficulties in applying the survey as it was developed and presented in Sec. 4. Minor adjustments to the specific cases, combined with reasoned choices, were made to reduce the impact on the validity and quality of the data, which remains the product of a human-led survey and is therefore inherently imperfect.

The nature of the research conducted did not allow for an evaluation of the internal structure of the datasets, the accuracy with which they were constructed, or their overall quality, as this was beyond the scope of the study. Similarly, it is beyond the scope to assess the accuracy and quality of the descriptions associated with dataset creation, as illustrated by Horsman and Lyle (2021) and OSAC Digital Evidence Subcommittee Task Group on Dataset Development (2022).

Overall, the considerations we discussed regarding metadata FAIRness in digital forensics (see Sec. 6.2) are based on generalizations from empirical observations. Although the statistical significance is limited due to the small number of datasets analyzed, the study considers CFReDS as a representative sample.

The generalizability of our results is also potentially affected by the fact that some of the criteria in our checklist are directly related to the CFReDS Portal and its structure (e.g., the threshold for evaluating the metadata scope criterion). However, the list in Appendix A, from which this checklist is derived, remains valid in a more general scope.

8. Conclusion

This work has shed light on the current practices for sharing datasets in digital forensics. While there are well-established repositories providing datasets that contribute to the findability of data, we have identified a significant deficiency in addressing the holistic FAIR principles in this field.

Our study began by addressing three fundamental research questions, namely, the availability and comprehensiveness of metadata for digital forensics datasets, the compliance of this metadata with the FAIR principles, and strategies to enhance such compliance. Through a meticulous examination of all 212 datasets referenced in the CFReDS Portal, we have uncovered a sobering reality: current practices in this domain are far from ideal.

Our contributions to the field are twofold: Firstly, we have identified the existing deficiencies in metadata quality, thereby expanding the understanding of data quality challenges within the research community. This insight serves as a valuable starting point for addressing these issues comprehensively. Secondly, we offer a practical set of recommendations aimed at improving metadata quality. These recommendations can be instrumental for dataset creators, curators, and data users, enabling them to enhance the completeness and quality of metadata associated with datasets.

Our contributions serve as a foundation for fostering better data practices, ultimately advancing the FAIR principles within the digital forensics community. We hope that this research will stimulate further discussion, innovation, and collaboration.

Author contribution statement

Samuele Mombelli: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation, Writing - Original Draft. **James R. Lyle:** Writing - Review & Editing. **Frank Breitinger:** Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision.

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.

Appendix A. Comprehensive checklist

Table A.5 through Table A.8 present the complete list of criteria developed to conduct the survey on the datasets included in the CFReDS Portal. This is a custom implementation of the FAIR principles, adapted to the research objective.

The criteria in Table A.5 are directly related to Principles F1. through F4. described in Fig. 1. The same applies to Table A.6 (Principles and sub-principles A1. to A2.), Table A.7 (Principles I1. to I3.) and Table A.8 (Principle R1. and sub-principles).

Remark: The tables cite Breitinger and Jotterand (2023). While the findings in the tables are based on a first draft of this paper, the original paper was never published. Findings might be slightly different when reviewing/using the published reference.

Table A.5
Checklist - Findability

Criterion	Reference(s)
General metadata:	–
└── Dataset title	Breitinger and Jotterand (2023), Starr et al. (2015)
└── Dataset author/creator	Breitinger and Jotterand (2023), Starr et al. (2015)
└── Creator identifier	Starr et al. (2015)
└── Data collection method	Breitinger and Jotterand (2023), Horsman and Lyle (2021)
└── Dataset publisher/contract	Starr et al. (2015)
└── Dataset release date	Starr et al. (2015)
└── Data quality	Breitinger and Jotterand (2023)
└── Dataset version	European Commission Directorate-General for Research & Innovation (2016), Starr et al. (2015)
Dataset content description and data definition:	Breitinger and Jotterand (2023), Wilkinson et al. (2016)
└── Dataset description	Starr et al. (2015); Breitinger and Jotterand (2023)
└── Data origin	European Commission, Directorate-General for Research and Innovation (2016)
└── Data formats	European Commission Directorate-General for Research & Innovation (2016)
└── Naming conventions	European Commission Directorate-General for Research & Innovation (2016)
└── Dataset structure	Horsman and Lyle (2021); Breitinger and Jotterand (2023)
(Meta)data identifier(s):	–

(continued on next page)

Table A.5 (*continued*)

Criterion	Reference(s)
└— Must be machine-actionable	Starr et al. (2015)
└— Must be persistent	Breitinger and Jotterand (2023), European Commission Directorate-General for Research & Innovation (2016), FORCE11 (2020), Horsman and Lyle (2021), Starr et al. (2015), Wilkinson et al. (2016)
└— Must be globally unique	Starr et al. (2015), Wilkinson et al. (2016)
└— Must be widely used (by a community)	Starr et al. (2015)
└— Must resolve on the web to the dataset	Starr et al. (2015)
Metadata must clearly include the (meta)data identifier(s)	Wilkinson et al. (2016)
(Meta)data findability:	–
└— Must allow search by keywords	European Commission Directorate-General for Research & Innovation (2016)
└— Must be registered or indexed in a searchable resource	Wilkinson et al. (2016)
Metadata must follow discipline-specific standards	Breitinger and Jotterand (2023), European Commission Directorate-General for Research & Innovation (2016)

Table A.6
Checklist - Accessibility

Criterion	References
Statement on open availability of (meta)data	European Commission Directorate-General for Research & Innovation (2016)
Description of how data are made available	European Commission Directorate-General for Research & Innovation (2016)
Standardized access and download of (meta)data:	FORCE11 (2020)
└— Open, free, universally implementable protocols	Breitinger and Jotterand (2023), European Commission Directorate-General for Research & Innovation (2016), Wilkinson et al. (2016)
└— Software/tools specification and documentations	European Commission Directorate-General for Research & Innovation (2016), Starr et al. (2015)
└— Details on access and download methods	FORCE11 (2020)
(Meta)data specifications, documentation and repository	European Commission Directorate-General for Research & Innovation (2016)
Access-control:	European Commission Directorate-General for Research & Innovation (2016), FORCE11 (2020), Starr et al. (2015)
└— Authentication and authorisation protocols	FORCE11 (2020), Wilkinson et al. (2016)
└— Specification on how to retrieve data	Starr et al. (2015)
(Meta)data persistence statement and policies:	Starr et al. (2015), Wilkinson et al. (2016)
└— Data availability and disposition (if restricted or de-accessioned, metadata should persist)	–

Table A.7
Checklist - Interoperability

Criterion	References
Machine-readable metadata formats to:	FORCE11 (2020)
└— Describe datasets	Breitinger and Jotterand (2023)
└— Structure datasets	Breitinger and Jotterand (2023)
└— Combine datasets	Breitinger and Jotterand (2023)
Standards (data models and formats):	Wilkinson et al. (2016)
└— Vocabularies	European Commission Directorate-General for Research & Innovation (2016)
└— Methodologies	European Commission Directorate-General for Research & Innovation (2016)
└— Semantics	FORCE11 (2020)
└— Syntax	FORCE11 (2020)
Qualified references to other (meta)data	Wilkinson et al. (2016)

Table A.8
Checklist - Reusability

Criterion	References
Rich dataset description:	Wilkinson et al. (2016); Breitinger and Jotterand (2023)
└— Dataset creation	Breitinger and Jotterand (2023), Horsman and Lyle (2021)
└— Data formats	European Commission Directorate-General for Research & Innovation (2016)
└— Data collection	Breitinger and Jotterand (2023), Horsman and Lyle (2021)
└— How to reuse data and how to interpret results	Breitinger and Jotterand (2023), Horsman and Lyle (2021)
└— Author/Creator	Breitinger and Jotterand (2023), Starr et al. (2015)
└— Creator identifier	Starr et al. (2015)
└— Publisher/Contract	Starr et al. (2015)
└— Release date	Starr et al. (2015)
└— Related publication(s)	Breitinger and Jotterand (2023)
└— Integrity mechanism (hash)	Horsman and Lyle (2021)

(*continued on next page*)

Table A.8 (*continued*)

Criterion	References
└── Contextual information (explanations, guidance, caveats, documentation for data use)	Starr et al. (2015)
└── Standardised (discipline-specific)	Wilkinson et al. (2016)
About metadata:	–
└── Common open machine-readable format	Breitinger and Jotterand (2023)
└── Standardised (discipline-specific)	Breitinger and Jotterand (2023), European Commission Directorate-General for Research & Innovation (2016), Wilkinson et al. (2016)
Unambiguous and accessible license, and link to relevant license	Breitinger and Jotterand (2023), FORCE11 (2020), Starr et al. (2015), Wilkinson et al. (2016)
Details:	–
└── On data availability for re-use	European Commission Directorate-General for Research & Innovation (2016)
└── On data re-usability/restrictions	European Commission Directorate-General for Research & Innovation (2016), FORCE11 (2020)
└── On data re-usability time limits	European Commission Directorate-General for Research & Innovation (2016)
(Meta)data persistence statement and policies	Starr et al. (2015), Wilkinson et al. (2016)

Appendix B. CFReDS portal taxonomy with datasets count

Table B.9 provides a comprehensive overview of the hierarchical taxonomy that organizes CFReDS datasets. The 'Count' column represents the cumulative number of records found in a given category. For example, the count associated with the 'Databases' category includes all datasets explicitly associated with the 'Databases' category, as well as all datasets associated with its child categories (in this case, 'SQL'), without counting duplicates if the dataset is explicitly associated with both categories.

Table B.9
CFReDS taxonomy with datasets count (continues on next column).

Category	Count
Data/Forensic Related	159
└── Databases	8
└── SQL	7
└── Date, Time & Place Analysis	3
└── Place	3
└── Cell Site	2
└──GPS	3
└── Timeline	2
└── Email Searching	7
└── Evidence Collection & Integrity Management	49
└── Hashing	13
└── Imaging	38
└── Disk Images	29
└── RAM Images	10
└── Media Preparation	2
└── Write Blocking	2
└── File Recovery	10
└── DFR	4
└── File Carving	10
└── Database Carving	9
└── Image Carving	9
└── Other Carving	9
└── Video Carving	9
└── Internet	14
└── Browser	4
└── Normal Browsers	4
└── Chrome	4
└── Internet Explorer	3
└── Private Browsers	3
└── TOR	3
└── Cryptocurrency	2
└── Bitcoin	2
└── Peer To Peer File Sharing	2
└── Search History	2
└── Telecommunications	12
└── Network Packets	12
└── Multimedia	55
└── Audio	15
└── Deep Fakes	3
└── Images & Photographs	25
└── Steganography	11
└── Video	21
└── Text	17
└── Social Media & Messaging	5
└── Messaging	5
└── Facebook Messenger	3
└── In Game Messaging	3

(*continued on next page*)

Table B.9 (*continued*)

Category	Count
└── WhatsApp	3
└── Instant Messaging	4
└── Social Media	2
└── Facebook	2
└── Instagram	2
└── LinkedIn	2
└── Twitter	2
└── SnapChat	2
└── Pinterest	2
└── TicToc	2
└── String Searching	3
└── Character Sets	3
└── Container Types	3
└── File Type	94
└── 3G2	4
└── 3 GP	3
└── ASF	3
└── AVI	4
└── FLV	3
└── MOV	5
└── MP4	4
└── MPG	5
└── WMV	4
└── MP3	12
└── WAV	3
└── PNG	2
└── JPEG	5
└── ZIP	34
└── XML	17
└── TXT	6
└── DMP	3
└── PDF	7
└── SQLITE	6
└── GEN	3
└── EO1	17
└── AFF	7
└── DMG	3
└── RAW	3
└── TAR	7
└── UFD	3
└── XLSX	3
IT System Type	90
└── File System	7
└── APFS	1
└── Ext2/3/4	1
└── FAT	4
└── FAT32	4
└── exFAT	2
└── HFS	2
└── NTFS	2
└── Alternate Data Streams	2
└── Other Devices & Systems	27
└── Cameras	3
└── Canon	2
└── Cars & Infotainment	0
└── Car Maker	0
└── Car Model	0
└── Cloud & Remote Systems	0
└── General Purposes Cloud System	0
└── AWS	0
└── Adobe Creative Cloud	0
└── Google Drive	0
└── iCloud	0
└── Other Cloud System	0
└── Drones	1
└── Gaming Systems	2
└── Xbox	1
└── Xbox 360	1
└── Xbox One	0
└── IOT	15
└── Memory	6
└── PC & Operating Systems	27
└── Linux/UNIX	8
└── Ubuntu	6
└── Mac	5
└── Mac Artifacts	4

(*continued on next page*)

Table B.9 (*continued*)

Category	Count
└── Mac Plists	4
└── Mac OS Version	4
└── Snow Leopard	4
└── Windows	15
└── Windows Artifacts	10
└── Windows Registry	10
└── Windows OS Versions	14
└── Windows 10	10
└── Windows 3.1	10
└── Windows 7	12
└── Windows 95	10
└── Windows XP	13
└── Phone, Mobile & Tablet	40
└── Android	26
└── Android OS	20
└── Android 10	14
└── Android 9	14
└── Android 8	14
└── Android 7	14
└── Android 1	13
└── Android 2	13
└── Android 4	13
└── Android 5	13
└── Android 6	13
└── Android Vendor	17
└── LG	12
└── Samsung	13
└── Google Pixel	12
└── Google Nexus One	12
└── Google Nexus S1	12
└── Google Nexus 5	12
└── HTC	12
└── Motorola	12
└── ZTE	11
└── Acquire Type	12
└── JTAG	12
└── Chip-off	12
└── Logical	11
└── Physical	11
└── Non Android or IOS Phones	5
└── Blackberry	0
└── Feature Phones	0
└── Other Mobile	4
└── Sony Ericsson	3
└── SE P800	1
└── SE T630	1
└── SE T68i	1
└── Nokia	1
└── Nokia 6230	1
└── Windows Mobile	1
└── iOS	11
└── iPad	5
└── iPhone	7
└── iPhone Hardware Version	7
└── iPhone 11	5
└── iPhone SE	7
└── iPhone OS Version	7
└── iOS 13.3.1	6
└── iOS 13.4.1	6
└── iPod	8
Simulated Cases/Scenarios	20
└── Data Leakage	0
└── Hacker Case	0
└── M57	0
└── M57 Patents	6
└── Nitroba U Harassment	2
└── Rhino Hunt	1
└── M57 Jean	2
└── NPS	4
└── Weapon Deletion	1
└── Weapons 2	1
└── Drug Traffic	1
└── Control	1
└── National Gallery DC	1
└── Lone Wolf	1
└── Narcos	1

(*continued on next page*)

Table B.9 (*continued*)

Category	Count
└── Owl	1
└── Tuck	1

References

Australian Research Data Commons, 2022. FAIR data self assessment tool | ARDC. URL: https://ardc.edu.au/resource/fair-data-self-assessment-tool/. .

Breitinger, F., Jotterand, A., 2023. Sharing Datasets for Digital Forensic: A Novel Taxonomy and Legal Concerns, vol. 45, 301562. URL: https://www.sciencedirect.com/science/article/pii/S2666281723000719, 10.1016/j.fsidi.2023.301562.

European Commission Directorate-General for Research & Innovation, 2016. 'H2020 programme: Guidelines on FAIR data Management in horizon 2020 version 3.0'. Guidelines. URL: https://repository.oceanbestpractices.org/handle/11329/1259. , 10.25607/OBP-774.

FORCE11, 2020. Guiding Principles for Findable, Accessible, Interoperable and Re-useable Data. Publishing version b1.0. URL: https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/.

Göbel, T., Baier, H., Breitinger, F., 2023. Data for digital forensics: why a discussion on 'how realistic is synthetic data' is dispensable. URL: 10.1145/3609863.doi:10.1145/3609863 Digital Threats. Presented at the 12th International Conference on IT Security Incident Management IT Forensics (IMF).

Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. Digit. Invest. 6, S2–S11. URL: https://www.sciencedirect.com/science/article/pii/S1742287609000346. doi: 10.1016/j.diin.2009.06.016.

GO, F.A.I.R., 2022. FAIR Principles - GO FAIR. URL: https://www.go-fair.org.

Grajeda, C., Breitinger, F., Baggili, I., 2017. Availability of datasets for digital forensics - and what is missing. Digit. Invest. 22, S94–S105. URL: https://www.sciencedirect.com/science/article/pii/S1742287617301913, 10.1016/j.diin.2017.06.004.

Horsman, G., Lyle, J.R., 2021. Dataset construction challenges for digital forensics. Forensic Sci. Int.: Digit. Invest. 38, 301264. https://www.sciencedirect.com/science/article/pii/S2666281721001815, 10.1016/j.fsidi.2021.301264.

Jones, S., Grootveld, M., 2017. How FAIR are your data? URL: https://zenodo.org/records/1065991, 10.5281/zenodo.1065991.

Juty, N., Wimalaratne, S.M., Soiland-Reyes, S., Kunze, J., Goble, C.A., Clark, T., 2020. Unique, persistent, resolvable: identifiers as the foundation of FAIR. Data Intelligence 2, 30–39, 10.1162/dint_a_00025.doi:10.1162/dint_a_00025.

Labastida, I., Margoni, T., 2020. Licensing FAIR Data for Reuse. *Data Intelligence, 2*, 199–207, 10.1162/dint_a_00042.doi:10.1162/dint_a_00042.

Mombelli, S., Breitinger, F., Lyle, J.R., 2023. Survey on FAIRness of CFReDS Portal Datasets' Metadata - 2022. https://doi.org/10.48657/k4vr-7z49.

National Institute of Standards and Technology, 2022a. The CFReDS project. URL: https://cfreds.nist.gov/.

National Institute of Standards and Technology, 2022b. Data - glossary | CSRC. URL: https://csrc.nist.gov/glossary/term/data.

OSAC Digital Evidence Subcommittee Task Group on Dataset Development, 2022. *'Guidelines for dataset development (V2)'*. Guidelines the organization of scientific area committees for forensic science (OSAC). URL: https://www.nist.gov/system/files/documents/2022/12/15/OSAC-DE-Guidelines%20for%20Dataset%20Development.pdf. .

Park, J., Lyle, J.R., Guttman, B., 2016. Introduction to CFTT and CFReDS projects at NIST. Journal of the Korea Institute of Information Security & Cryptography 26, 54–61. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=921807.Publisher:JungheumPark,JamesR.Lyle. Barbara Guttman.

Roussev, V., 2011. An evaluation of forensic similarity hashes. 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.

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R.R., Duerr, R., Haak, L.L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J.E., Lin, J., Nielsen, L.H., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., Taylor, M., Clark, T., 2015. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1, e1. https://peerj.com/articles/cs-1, 10.7717/peerj-cs.1.

Swiss National Science Foundation, 2017. Data Management Plan (DMP) - Guidelines for Researchers. Guidelines Swiss National Science Foundation (SNSF). URL: https://www.snf.ch/SiteCollectionDocuments/ORD_Research_Council_3May2017_E.pdf.

SWISSUbase, 2023. About Us. URL: https://resources.swissubase.ch/about-us/.

The FAIRsharing team, 2023. FAIRsharing | standards. URL: https://fairsharing.org/standards.

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., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. URL: https://www.nature.com/articles/sdata201618, 10. 1038/sdata.2016.18.