

# Data to Infinity and Beyond: Examining Data Sharing and Reuse Practices in the Computer Security Community

Anna Crowder  
University of Florida  
annacrowder@ufl.edu

Allison Lu  
University of Florida  
allison.lu@ufl.edu

Kevin Childs  
University of Florida  
k.childs@ufl.edu

Carson Stillman  
University of Florida  
carson.stillman@ufl.edu

Patrick Traynor  
University of Florida  
traynor@ufl.edu

Kevin R.B. Butler  
University of Florida  
butler@ufl.edu

**Abstract**—Sharing high-quality research data specifically for reuse in future work helps the scientific community progress by enabling researchers to build upon existing work and explore new research questions without duplicating data collection efforts. Because current discussions about research artifacts in Computer Security focus on reproducibility and availability of source code, the reusability of data is unclear. We examine data sharing practices in Computer Security and Measurement to provide resources and recommendations for sharing reusable data. Our study covers five years (2019-2023) and seven conferences in Computer Security and Measurement, identifying 948 papers that create a dataset as one of their contributions. We analyze the 265 accessible datasets, evaluating their understandability and level of reuse. Our findings reveal inconsistent practices in data sharing structure and documentation, causing some datasets to not be shared effectively. Additionally, reuse of datasets is low, especially in fields where the nature of the data does not lend itself to reuse. Based on our findings, we offer data-driven recommendations and resources for improving data sharing practices in our community. Furthermore, we encourage authors to be intentional about their data sharing goals and align their sharing strategies with those goals.

## 1. Introduction

The creation, use, and analysis of data is an integral component of experimental science. The White House Office of Science and Technology Policy has written that access to data provides societal benefits such as “accelerating discovery, fostering collaboration, advancing equity” and maximizing the return on investment for basic research funding [1]. Within the Computer Security community, data has been generated to support research efforts for diverse use cases from Internet data packets used for network measurement research to collections of binaries for assessing system security to collections of survey and interview data from user studies.

However, to date, there has been little investigation of how datasets within the Security community have been

created, how they have been shared, or how they have been used by others. While the community has begun to recognize the value of code artifacts to ensure reproducibility and has provided mechanisms such as evaluation badges for work that provide artifacts, datasets tend not to be evaluated and recognized in a similar way. This lack of recognition stands in contrast with the Measurement community, which has incentivized the creation of datasets through awards [2], or the Machine Learning community, which has demonstrated a critical need for data to train models and where recent work has provided guidance on how to assure the quality of datasets [3]. Given the considerable time and resources involved with generating, curating, and maintaining datasets, it is important to understand whether these investments are being leveraged by the Security community and if not, whether lessons from related fields can provide guidance on improving the state of Security datasets to facilitate research that advances our field.

In order to answer these challenging questions about the state of dataset generation and reuse in the Security community, we performed an extensive study of over 3,000 papers over the past five years, spanning quantitative and qualitative research, identifying datasets, and further analyzing critical properties of them to provide a first characterization. Because the Measurement community has developed incentives for dataset creation, we also look across publication venues in this community to compare the state of dataset sharing and reuse with the Security community. We make the following contributions:

- **Comprehensive Longitudinal Study:** We perform the largest known study of data sharing behavior across quantitative and qualitative research in the Computer Security and Measurement communities. This covers the past five years yielding observations on 948 papers that claim the creation of a dataset as a contribution of their work. We find that 60% of these papers do not provide any statement about the availability of their data.
- **Evaluate Data Sharing Practices:** We further as-

sess the 265 publicly accessible datasets across metrics of data sharing. We find that no standard practices exist for the documentation or format of publicly available Security or Measurement datasets. As a result, 16% of datasets are not understandable to independent researchers even with the original paper, and 44% are only understandable within the context of the original paper.

- **Investigate Data Reuse:** We evaluate dataset reuse based on dataset type, identifying which types see limited reuse and which are frequently reused. Overall, the reuse of Security and Measurement papers is low. 41% have never been reused and the median reuse of the remaining datasets is 2. This analysis aims to direct researchers' data sharing efforts toward intentional dataset creation for reuse based on the community's needs.
- **Data-driven Recommendations:** To establish a standard for data sharing documentation, we first assess the suitability of the questions raised by Gebru et al.'s widely-cited *Datasheets for Datasets* paper [3] used as a *de facto* standard in the Machine Learning community for dataset creation. Our analysis reveals that certain *Datasheets* questions are not widely applicable, as even the most reused datasets in our study—and widely adopted Machine Learning datasets—lack documentation for these areas. We refine our recommendations around the pertinent *Datasheets* questions, enhancing them with additional questions and guidelines to cover the issues we observe in existing datasets.

Our investigation, spanning over 7 TB of data, raises significant questions about data sharing and reuse in the Security community. While certain areas within the community are more amenable to using existing datasets than others, it is an open question as to why these differences exist. This work represents a significant first step in improving the state of dataset creation and reuse, but while such an effort may be arduous, the ability to advance the field by drawing on the work of others could have transformative benefits.

The rest of this paper is structured as follows: Section 2 provides background information on data sharing policies and research; Section 3 outlines our methodology; Section 4 discusses the results of our initial investigation of data sharing behavior; Section 5 presents the observations from our analysis of data reuse; Section 6 provides our assessment of *Datasheets for Datasets* as a guideline for documentation; Section 7 outlines our recommendations on sharing data for reuse; Section 8 presents limitations of our study; Section 9 highlights related work both from within Computer Security and the broader academic community; Section 10 provides concluding remarks.

## 2. Background

Publicly available datasets support reproducibility and benefit new work through dataset reuse. In this work, we

focus on data sharing practices specifically for reuse, which is defined as applying a dataset outside its original context [4]. When a dataset is shared for reuse, the required documentation level is extensive as it needs to comprehensively explain the dataset's content and context [4], [5].

Many publication venues outside of Computer Security have mandatory data sharing policies for publication. In 2011, top journals in evolution and ecology adopted the Joint Data Archiving Policy (JDAP), which states that a condition for publication is that the data necessary for reproducibility is available in a public archive [6]. The journal *Cognition* implemented a similar policy in 2015 [7] and work by Hardwick et al. [8] revealed that the introduction of the policy increased the presence of data availability statements by 53% and reusability of datasets by 40%. The policy for all *PLOS ONE* journals, a publisher for journals in engineering, biology, psychology, and more, also requires authors to make their data available for reproducibility and requires them to include a Data Availability Statement in the publication [9].

Computer Security conferences foster reproducibility and accessibility through Artifact Evaluation Committees (AECs) [10]. While AECs encourage data sharing because data is a necessary part of reproducibility, they do not specifically require data sharing as a requirement to receive an artifact badge, with artifact functionality and reproducibility instead being the focus. As a result, AEC members do not assess the quality of shared datasets at all let alone their suitability for reuse. Papers published at a venue with an AEC have the option to submit their artifacts for review, and then the committee awards badges based on the quality and extent of the artifacts. For example, the AEC at *USENIX Security* offers three badges: *Artifact Available*, *Artifact Functional*, and *Artifact Reproduced*.

In 2017, the *ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)* and the *Annual Computer Security Applications Conference (ACSAC)* were the first Computer Security conferences to implement Artifact Evaluation Committees, and *USENIX Security* was the first Tier 1 conference in 2020 followed by the *ACM Conference on Computing and Communications Security (CCS)* in 2023 and the *Network and Distributed System Security Symposium (NDSS)* in 2024. Within the Measurement community, the *ACM Internet Measurement Conference (IMC)* does not have an AEC, but starting in 2023, artifact availability was a requirement for paper submissions. If authors did not plan to make artifacts publicly available, they were required to provide a "legitimate reason." Upon acceptance, the shepherding process ensured that promised artifacts were available. In 2025, *USENIX Security* introduced a similar policy [11] which requires all accepted papers to make their artifacts publicly available or provide a detailed justification for not sharing. In 2024, *CCS* introduced a requirement that submitted papers make their artifacts accessible to reviewers but submission to the AEC remains optional [12].

Guidelines for documenting machine learning datasets have been proposed in the form of *Datasheets* [3] and *Data Cards* [13]. The goal of these guidelines is to assist dataset

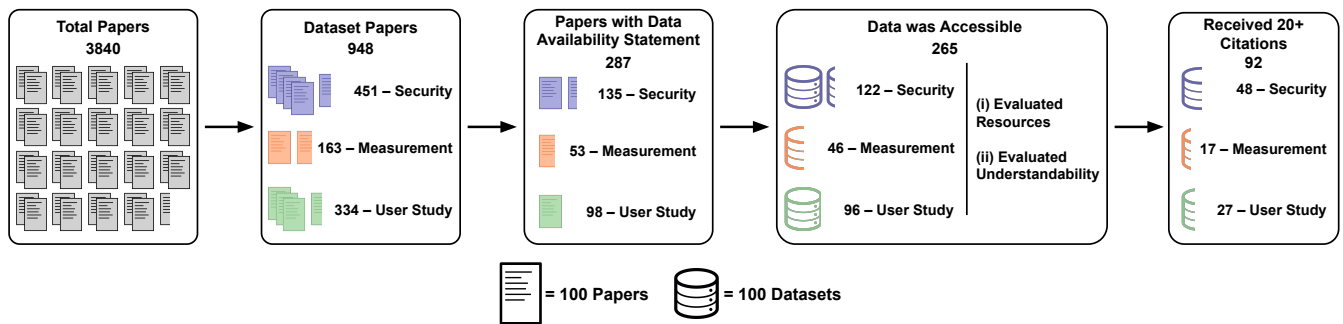


Figure 1: **The pipeline for our study methodology.** First, we survey quantitative Security papers, qualitative Security papers (User Study), and Measurement papers from seven conferences, and select the papers that create a dataset as one of their primary contributions. Then, we identify which papers state that their data is publicly available. For the papers that provide a statement, we check the dataset’s accessibility and evaluate the accessible datasets on their provided resources (i.e., documentation and analysis source code) and understandability by an independent researcher. The final stage of our study investigates the reuse of the accessible datasets with 20+ citations.

creators in sufficiently writing or structuring documentation for their dataset so that future users have the necessary information to use the datasets correctly. *Datasheets* [3] provides a set of 57 questions related to the motivation, composition, collection process, cleaning or labeling, use, distribution, and maintenance of the dataset. The *Data Card* [13] is a physical template for documenting information about the stages of a dataset’s lifecycle: origins, factuals, transformations, experience, and examples. Because *Datasheets* focuses specifically on what questions should be answered within a dataset’s documentation and less on the structure of the documentation, we use it as a guide for comprehensively examining the level of documentation provided by the most reused datasets in our work.

### 3. Methodology

We seek to characterize the current state of publicly available datasets created by the Computer Security community and their suitability for reuse in future work. To do so, we examine the Security community’s datasets from the last 5 years including quantitative work and qualitative User Study papers, and compare them with datasets created by the Computer Measurement community. In the past, the Measurement community has incentivized data sharing by offering an award recognizing the creation of a dataset. To be eligible, papers were required to share their data publicly [2]. Based on their existing attitude toward data sharing, we hypothesize that the data sharing practices for structure and documentation within the Measurement community will benefit the Security community.

Based on data sharing research in other fields [8], [14], we organize our analysis of the datasets into four stages: (1) Stating that the data is available; (2) Making the data accessible; (3) Providing resources to support data reuse; (4) Ensuring that the data is understandable to an independent researcher. We investigate how well the datasets from the

papers we survey perform in each of these stages to gather a baseline understanding of data sharing behavior in the Computer Security and Measurement communities. We then measure the reuse of well-cited datasets and perform a deeper analysis of the quality of documentation for the most reused datasets using *Datasheets* as a guide.

In this section we outline our paper selection process, analysis methodology for the four stages, and how we quantify the reuse of well-cited datasets. The steps of our methodology are outlined in Figure 1. We formulate the following research questions to direct our study:

- RQ1** What is the current state of data availability, accessibility, and understandability in the Computer Security community?
- RQ2** What is the current level of reuse among publicly available datasets?
- RQ3** What data sharing qualities and gaps exist in the data sharing practices of the most reused publicly available datasets?

#### 3.1. Paper Selection

Our analysis includes papers where the creation of a dataset is one of the primary contributions of the paper. We further clarify that the data must exist beyond the body of the paper. Data external to the body of the paper is not required for publication and therefore requires motivation to share beyond publication incentives. Examples of datasets include raw data such as network traffic, survey responses, and audio files, as well as processed data such as annotated website links and qualitative codebooks from User Study papers. We exclude papers that create a dataset solely for training, testing, or evaluating an artifact developed in the paper unless the authors emphasize the dataset’s creation as a contribution.

We collect papers from the four Tier 1 Security conferences (*USENIX Security*, *CCS*, *NDSS*, and the *IEEE*

<i>Factor</i>	<i>Variable</i>	<i>Question</i>
<b>Data Accessibility</b>	Access	Could data be accessed/opened?
	Location	Where is the data made available?
	Format	What file format is the data in?
<b>Resources</b>	Documentation	Is documentation about the dataset provided and to what extent?
	Source Code	Is source code for the data provided with or without instructions?
<b>Understandability</b>	Level	Is the dataset understandable either with or without the context of the paper?
	Reason	Why is the dataset not understandable?

TABLE 1: The three factors we use to analyze each dataset from a paper with a Data Publicly Available statement. Data Accessibility, Resources, and Understandability are all vital data sharing elements.

*Symposium on Security & Privacy (IEEE S&P)*) published between 2019 and 2023 (5 years). We also collect papers from two measurement conferences (*IMC* and *ACM SIG-METRICS*) and a usable security conference (the *Symposium on Usable Privacy and Security (SOUPS)*) from the same years, creating a corpus of 3,840 papers.

To identify papers that focus on dataset creation, two raters independently review the set of papers from each year/conference pair. They first examine the *Abstract* for dataset-related keywords (e.g., dataset, measurement, collection, longitudinal, etc.) along with other nuanced indicators of dataset creation that require contextual understanding. If needed for clarification, the raters also read the *Introduction*. Agreement between raters is achieved through in-person discussion until full consensus is reached [15].

### 3.2. Data Collection

For each of the resulting 948 dataset papers, we collect general information on the paper and its dataset. This information includes the paper’s citation count, the origin of the dataset (real-world, synthetic, or User Study), and data collection status (ongoing or one-time collection). The data collection process is conducted by a team of raters consisting of graduate-level Computer Science students with a primary rater overseeing the process and clarifying scope.

**(1) Data Availability Statement.** Publicly available data should be directly linked to the paper that introduces it to ensure accessible long-term access [14]. Therefore, our analysis focuses on papers that include a statement indicating that their data is available and provide the location of the data. For each paper, we determine whether an availability statement is present and categorize it as follows:

- Data publicly available
- Data not publicly available
- Data will be publicly available
- Data publicly available upon request
- Non-data artifacts available<sup>1</sup>
- No statement

For the papers that have a “Data publicly available statement,” we evaluate the paper’s dataset on the remaining

three data sharing stages: Data Accessibility, Resources, and Understandability. We outline in detail each category and its factors in Table 1 and discuss them in the following sections. Rater 1 collects the general information on each paper, determines the presence or lack of an availability statement, and codes each available dataset on the three stages. Rater 2 only performs the coding for the datasets with a data availability statement. Similar to paper selection, any discrepancies in coding are discussed at length between the raters until a full consensus is reached. The data collection process is completed using separate Google forms for Rater 1 and 2. Our full dataset is available through our website.<sup>2</sup>

**(2) Data Accessibility.** We attempt to access each dataset stated as available, noting its location, the file format(s), and any issues encountered during the process. We also document the type of data available: raw, processed, qualitative codebook, or any combination of these. *Raw* data refers to data that has not been modified beyond the original measurement process. In contrast, *processed* data is raw data that has been transformed, analyzed, or annotated. A *qualitative codebook* consists of themes that emerge during the thematic analysis of raw qualitative data and are subsequently used to analyze that data [16]. Although qualitative codebooks can act as a non-data artifact, they also serve as processed data when accompanied by participant counts or raw participant responses. We include all qualitative codebooks in our analysis, as the data sharing practices for codebooks remain the same regardless of whether they contain data.

Finally, to determine whether a dataset is actively maintained or updated, we document the most recent modification date for the dataset or a related artifact. This metric is not limited to the last dataset update, as some datasets originate from one-time data collections and do not require regular updates. However, updates to related artifacts indicate ongoing maintenance efforts by the dataset authors. To identify the most recent modification date, we examine the latest commit or version for artifacts hosted on repositories. For artifacts hosted on websites, we check the `Last-Modified` field in the HTTP header. In cases where this field is absent, we are unable to determine the last modification date for those websites.

1. Non-data artifacts include source code for a tool or model or interview or survey questions for a user study.

2. <https://sites.google.com/view/data-to-infinity-beyond/home>

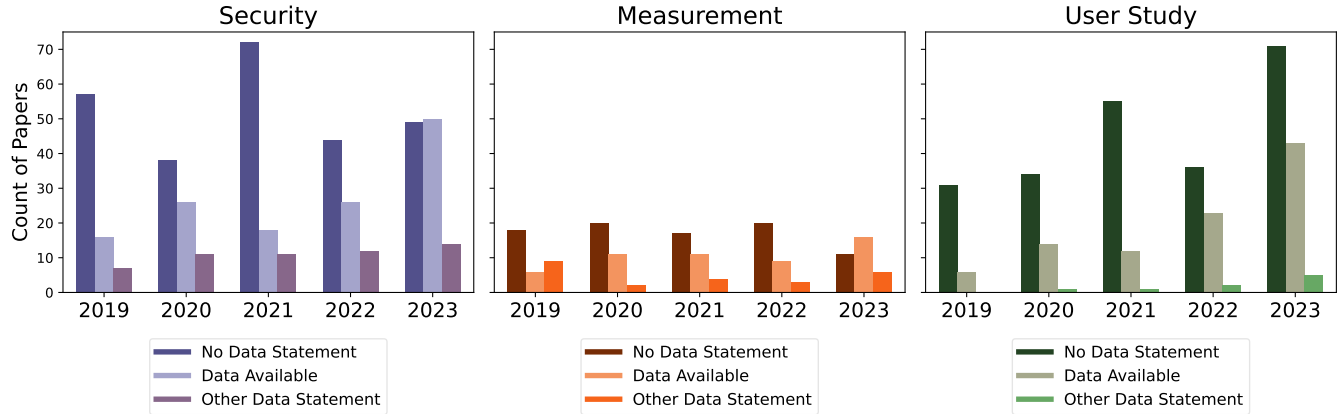


Figure 2: **Count of the types of availability statements in Security, Measurement, and User Study papers from 2019 to 2023. Security and Measurements have a slight downward trend in the number of papers with no statement, and an upward trend for all other statement types indicating generally more papers making an explicit statement about the availability of their data. Usability papers show an upward trend in all statements indicating more papers.**

**(3) Resources.** Access to a dataset is often not enough for a dataset to be reusable. Documentation describing the dataset provides valuable context for researchers trying to reuse the dataset. We investigate the existence and level of documentation provided for publicly accessible datasets. We code the documentation as Thorough, Minimal, Documentation for a different artifact, or None. If the documentation provides a general description of the dataset, it is labeled as *Minimal* whereas if the documentation provides specific details about the dataset and its contents, it is labeled as *Thorough*.

In some cases, the source code used to process the dataset also acts as a form of documentation to help the user better understand the best way to handle the data [14]. We examine source code for the datasets either with or without instructions or comments to help run the code.

**(4) Understandability.** Finally, we assess whether a dataset is comprehensible to an independent researcher, with or without the original paper’s context. A dataset is labeled as *not understandable* if any component—such as the file naming scheme, column headers, abbreviations, units, labels, or variables—is left unexplained. If an unclear component can be interpreted by referring to the paper, the dataset is labeled as *understandable only with the paper*. If all elements are clearly defined, the dataset is labeled as *understandable without the paper*. For datasets labeled as not understandable, raters provide a rationale for their assessment.

### 3.3. Dataset Reuse Analysis

Through our data collection process, we establish a baseline understanding of the availability of Security, Measurement, and User Study datasets. Next, we seek to measure at what level the accessible datasets are used for new research or influence new research in a substantial way, as well as who is using these datasets.

Our analysis includes each accessible dataset whose original paper has 20 or more citations, as this number allows us to maintain a manageable sample size while prioritizing papers that are growing in influence. For a given dataset paper (original paper) we first collect all the papers that cite the original paper (citing papers) by scraping the original paper’s Google Scholar “Cited by” link and downloading the citing papers as PDFs. Next, we remove non-English publications and anything that is not a full research article such as school projects, poster abstracts, or opinion articles.

Next, we remove citing papers that only cite the original paper in the *Introduction*, *Background*, *Related Work*, *Discussion*, *Limitations*, *Future Work* or *Conclusion* as these sections of a publication do not explain the details of a publication’s methodology or results. For the majority of the papers, we implement a Python script to convert the PDF to text and locate the relevant citation within the text. However, for PDFs with uncommon formatting, we perform the filtering manually. Finally, we manually read the remaining citing papers and document any that use the original paper’s dataset or create a similar dataset.

We examine whether the authors of the citing papers are also the original authors or collaborators with those authors. We hypothesize that direct connections could allow citing authors to access the dataset directly from its creators. If all citing authors had such access, the dataset may not be truly publicly available to those without those connections. To identify collaborators of the original authors, we query the dblp computer science bibliography [17] using the `dataprep` python library [18]. We then match the original authors’ collaborators to any of the authors of the citing papers. We manually confirm any matches due to the possible existence of duplicate names in dblp.

## 4. Dataset Analysis

We find that sharing practices can significantly differ between real-world or machine-generated datasets and human-sourced datasets (e.g., user studies). We present our analysis of the collected datasets and discuss them separately. We refer to **Security papers** ( $n = 451$ ) as quantitative papers from the Tier 1 security conferences (*USENIX Security*, *CCS*, *NDSS*, *IEEE S&P*) and *SOUPS*, **Measurement papers** ( $n = 163$ ) as papers from *IMC* and *SIGMETRICS* (which both only include quantitative work), and **User Study papers** ( $n = 334$ ) as qualitative papers from the Tier 1 security conferences and *SOUPS*. We analyze these datasets with regard to data availability, data accessibility, available resources, and understandability.

### 4.1. Security and Measurement Datasets

**(1) Data Availability Statement.** Of the 451 Security dataset papers and 163 Measurement papers we survey, 136 (30%) Security and 53 (33%) Measurement provide a statement that data is publicly available, while 260 (58%) Security and 86 (53%) Measurement papers do not discuss the availability of their dataset. The remaining papers provide other statements about data availability. Examples include claims that the data will be available in the future, the data is available upon request, or the data is not available for a specified reason.

Within the papers that provide other statements about data availability, many reasons exist for not making datasets publicly available such as privacy concerns or limited resources. Authors may choose to make datasets available upon request to ensure greater control over their data. For some datasets we survey, access is contingent on users completing a data usage agreement stating that the user will not use the data for commercial purposes. Other authors require users to complete an agreement committing to properly care for sensitive information within the dataset. Still, not all authors with datasets that require granting of access provide a reason for limiting access. We focus the additional stages of our analysis on papers that state their dataset is publicly available as we should be able to access such data without the assistance of the original dataset creators.

Figure 2 illustrates the yearly count of Security, Measurement, and User Study papers that either make no statement about their data, specify that their data is available, or provide another type of statement about the data. We address the results for User Study papers in Section 4.2. Both Security and Measurement papers show a minor downward trend in the amount of papers with no statement about data over the five years. Additionally, the number of papers with a data availability statement produces a minimal upward trend. In contrast to our assumption made in Section 3, Security and Measurement papers have similar trends in the number of papers with or without statements about the availability of their data. Due to limited data points and year-to-year fluctuations in our analysis, further research is needed to verify if these trends persist over time. With the increased

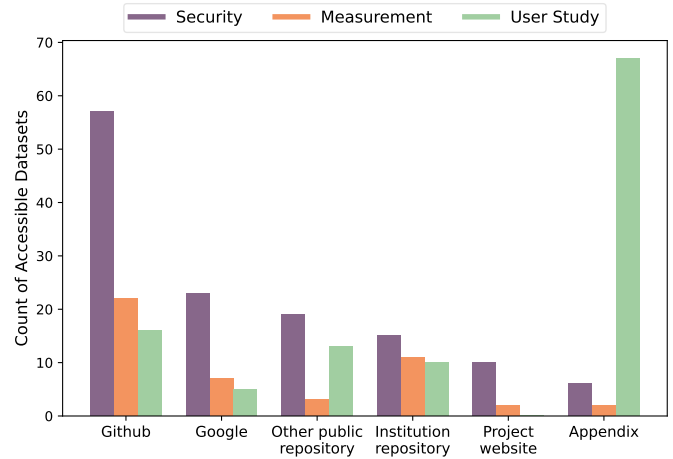


Figure 3: **The number of accessible datasets hosted on each type of platform. The Security and Measurement datasets primarily use Github, while the User Study papers often provide their data in the Appendix of the original paper.**

presence of AECs and mandatory availability statements at key publication venues which prompt authors to think about data sharing at least for reproducibility, we anticipate and encourage continued progress toward consistent data sharing statements across these research communities.

**Finding 1.** 58% of Security papers and 53% of Measurement papers from the last five years do not provide a statement about the availability of their data.

**(2) Data Accessibility.** Of the 136 Security and 53 Measurement papers that state their data is publicly available, 20 (13 Security and 7 Measurement) do not have accessible datasets. The datasets are not accessible because the linked repository or website does not have the data that the paper claims it does (11), the link is broken or has restricted access (6), no link is provided (2), or the compressed data files cannot be unzipped (1). We input the 11 repositories with missing data and the 6 broken or restricted links into the Internet Archive Wayback Machine [19], a tool that regularly collects snapshots of publicly available web pages, to see if the datasets were ever accessible. One of the broken links was archived in March of 2023 but was still broken at the time. Another link was archived repeatedly between October 26, 2021, and January 13, 2024, but the snapshots show a broken link starting February 6, 2022. Finally, one of the links to an empty repository had source code in August 2021 but still did not have data. The remainder of the links are not archived by the Wayback Machine.

Three of the datasets that were accessible at the start of our study became inaccessible due to the link no longer being active. Two of the links were connected to university repositories, and it is unclear why they were removed.

Additionally, 15 of the Security datasets and 6 of the Measurement datasets are only partially available. These



datasets are partially available because the full dataset was available upon request (7), the authors did not share all of the datasets described in the paper (6), the full dataset could not be shared due to privacy or proprietary concerns (3), the authors provide a measurement framework that needed to be run to collect the data (2), the authors state that the full dataset will be available in the future (1), a link to some of the data files is broken (1), or a portion of the files cannot be unzipped (1). We include the partially available datasets in the remainder of our analysis.

For papers with accessible datasets, the dataset's location often differs from that of other associated artifacts, and some datasets are distributed across multiple platforms. Figure 3 illustrates the frequency with which each platform is used. GitHub is the primary hosting platform for Security and Measurement datasets, with 57 (47%) Security and 22 (48%) Measurement datasets accessible there. The category "Other public repositories" includes platforms that are used infrequently, such as Zenodo, OSF, Bitbucket, Dropbox, and Hugging Face. Table 5 in Appendix A shows the full distribution of hosting platform use.

For accessible datasets, we check whether the year of the most recent modification is later than the publication year of the original paper. While we cannot determine with certainty how frequently a dataset is maintained by its authors, any updates made after publication indicate at least some level of maintenance. This analysis excludes datasets shared solely within the paper itself, as published papers cannot be updated. Among the accessible external datasets, 41 out of 113 (36%) Security datasets and 20 out of 42 (48%) Measurement datasets have been updated in the years following the original paper's publication. In some cases, ongoing dataset updates and maintenance are evident, particularly when new data is regularly collected and published, as seen with *Tranco* [20] and *Censored Planet* [21]. Conversely, other datasets show signs of a lack of maintenance, such as unresolved open issues on repository platforms that have remained open for years.

**Finding 2.** 90% of Security papers and 87% of Measurement papers that provide a data availability statement within their paper have partially or fully accessible data.

**(3) Resources.** Documentation for datasets is critical as it fosters reproducibility, increases transparency, and ensures that dataset users have the necessary information to use a dataset properly [3]. Improper use of a dataset can lead to biased machine learning models or misinterpreted results. Figure 4 shows the percentage of papers with accessible datasets from both the Security and Measurement communities that provide each level of documentation. Of these papers, 25% of the Security and 26% of the Measurement provide a Thorough level of documentation for their dataset. For example, Nair et al. [22] created a *Datasheet* [3] for their dataset, a Nutrition Card [23] filled with quick facts about the composition of the dataset, and additional information throughout their website. A larger amount of these papers, 47% of Security and 46% of Measurement, have

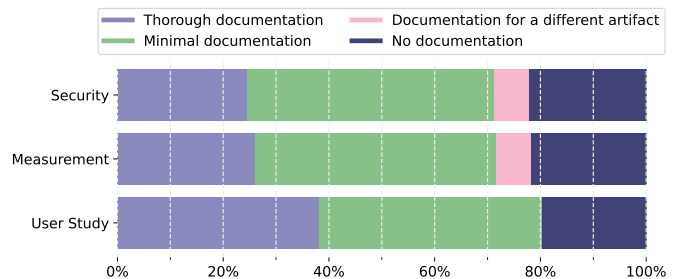


Figure 4: The percentage of accessible datasets with each level of documentation. "Thorough documentation:" all portions of the dataset had corresponding documentation. "Minimal documentation:" while some documentation is provided some portions are missing and/or insufficient. "Documentation for a different artifact:" documentation was provided for some other artifact (such as scripts for analysis) "No documentation:" documentation was not provided for any artifact.

only a Minimal level of documentation, often consisting of a single paragraph in a README file which is not enough to adequately describe the contents, file structure, or proper use of a dataset. 7% of the Security papers and 7% of the Measurement papers supply documentation for a different artifact from their paper, but include no information on the dataset. 22% of the Security papers and 22% of the Measurement papers provide no documentation with some repositories containing a README that still says "TODO" after several years. Again we see very little split between the Security and Measurement papers in terms of the level of documentation they provide.

Sharing source code specifically for datasets is not mandatory as some datasets do not require source code for processing or analysis. However, source code can be beneficial for future dataset users as it provides insight into best practices for handling the dataset. When source code is shared with a dataset, providing instructions and comments explaining how to run the code is always beneficial. We find that 48% of Security papers and 39% of Measurement papers with accessible datasets provide source code with instructions. A smaller portion, 10% of Security and 17% of Measurement papers, provide source code without any instructions or comments. The remaining papers either provide source code for a different artifact or tool or do not provide any source code.

**Finding 3.** Consistent practices do not exist for the structure or documentation of publicly available datasets.

**(4) Understandability.** We assess the accessible datasets for their understandability, categorizing them as understandable without the paper, understandable only with the paper, or not understandable at all. This analysis offers an additional metric to gauge how many datasets are practically reusable for independent researchers. As shown in Figure 5, 18% of the Security datasets and 33% of the Measurement datasets

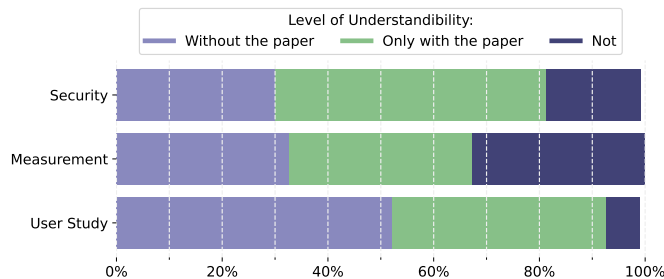


Figure 5: **The percentage of accessible datasets with each level of understandability.** “Without the paper:” the raters could understand the dataset with only the dataset and dataset-specific documentation. “Only with the paper:” the raters had to reference the accompanying paper to understand the dataset. “Not”: the raters did not understand the construction of the dataset even with the accompanying paper as a guide.

are found to be indecipherable by both of our independent raters. When raters deem a dataset not understandable, they provide detailed explanations, which help inform our recommendations for improving data sharing practices. The main barrier to understanding is the absence or inadequacy of documentation. In 28 papers, documentation is too limited for an independent researcher to grasp the dataset’s structure or content. Out of the 28 datasets, 9 include some documentation, yet fail to clarify key variables or fields—for example, using abbreviated column headers or numerical codes without a key. Three have understandable content, but their organization is confusing, with thousands of files and no clear file structure or naming convention. Sixteen lack both file structure and data content explanations.

Another frequent issue, found in eight datasets, was the use of non-human-readable formats without accompanying documentation. Non-human-readable formats are sometimes necessary, and our analysis shows that they are highly used. Of the accessible datasets, 48 (39%) Security and 20 (43%) Measurement are shared fully or partially as compressed files (e.g., zip, tar.gz, etc.)<sup>3</sup>. Because compressed files are not human-readable in their current form, users must first download and process the file before previewing the data. When files are too large to download on an average user device, the data becomes impossible to understand without major effort. In Section 7, we provide recommendations for sharing non-human readable data files in an accessible way.

The existence of documentation for datasets is high for the accessible datasets. However, because both the Security and Measurement communities have no standards or guidelines for data sharing, the quality of the documentation is widely variable. Furthermore, the presence of documentation alone is not enough to guarantee an understandable dataset. To help fill this gap, we provide data sharing recommendations in Section 7.

3. The full list of data file types is shown in Table 4 in Appendix A.

**Finding 4.** *Lack of understandability for a portion of the accessible datasets is due to insufficient documentation of file content and structure as well as poor dataset organization.*

## 4.2. User Study Datasets

User Study papers primarily involve conducting surveys and interviews but can include additional types of human-centered studies such as ethnographies. Raw data from these research methods include survey responses, interview transcripts, or field notes. Processed data from these studies include demographics or qualitative codebooks that capture the essence of the raw data without extraneous or sensitive details. Raw data from User Study papers is challenging to share because it often contains personal information. A codebook provides a simple artifact that represents the data for reproducibility and future work.

**(1) Data Availability Statement.** Within our User Study papers, we find that 98 of 334 (29%) have a data availability statement and 227 (68%) make no statement about the availability of their data. A majority of the papers that do not provide a data statement (156/227), instead provide non-data artifacts in the form of protocol documents such as interview or survey questions.

Figure 2 shows an increase across the five years for the number of papers that provide each type of statement. These trends are likely the result of the number of user study projects within Security growing overall. However, each year, the majority of User Study papers do not provide any statement about the availability of data.

**Finding 5.** *68% of User Study papers do not provide a statement about the availability of their data, but the majority of those papers make non-data artifacts available.*

**(2) Data Accessibility.** The datasets from User Study papers with data availability statements are highly accessible. Of the 98 papers with data availability statements, only two are fully inaccessible due to broken links, and five are partially accessible due to broken links, proprietary data, or partial request access data.

As shown in Figure 3, most User Study datasets are shared in the Appendix of the original paper. Of the 96 accessible datasets, 53 share their data solely in the Appendix. This sharing method serves as a guaranteed way to provide additional materials related to a paper without the risk of links being deprecated, mis-referenced, or left empty. Qualitative research lends itself to sharing information in the Appendix through formatted codebooks and graphics or tables that represent survey responses. The papers that share only non-data artifacts often also only share the material in the Appendix. It is unclear why so many User Study papers over the last five years share only non-data artifacts when the method for sharing data in the form a codebook or graphic



is the same and does not require a substantial increase in effort.

Other User Study papers share partial data, such as codebooks or graphics, in the Appendix while providing external links for larger datasets. For instance, Perry et al. [24] investigated whether AI-assisted coding leads to more insecure code and included statistics and demographic details in the Appendix. Additionally, they shared anonymized interaction logs via a third-party link. This approach—sharing processed data in the Appendix and raw data through external sources—was observed in 14 of the accessible User Study papers we analyzed.

The remaining 29 User Study papers with accessible data are shared only through external links. They share information like interaction logs, survey results, or extensive details relating to their codebook that are too large to share in an Appendix. Five out of seven CCS User Study papers chose to share their codebook through an external link. This choice may be due to the 15-page limit at CCS that includes citations and appendices. Table 5 in Appendix A illustrates the large portion of datasets that share their data in the Appendix as well as the distribution of the rest of the hosting platforms used by User Study datasets.

In contrast to the Security and Measurement datasets, the User Study datasets that are shared externally are mostly shared in an easily accessible, human-readable format. 21/43 share their external data as a pdf. Others (12) share the data as csv files, and only a few (5) share the data as compressed files. The remainder of the external User Study datasets use a variety of formats such as txt, json, or xlsx as shown in Table 4 in Appendix A.

**Finding 6.** *97% of User Study papers that provided a data availability statement within their paper have partially or fully accessible data.*

**(3) Resources** Most User Study papers share survey and interview questions, but additional shared resources are often needed to understand the context in which qualitative research was conducted. This includes documentation for the data such as descriptions of codes within a codebook or details about how an interaction log was collected. We see that 37 of our 96 papers (39%) with accessible datasets share Thorough documentation while 19 (20%) have no documentation and 40 (42%) have Minimal documentation. While the practices are comparably better (see Figure 4) than other disciplines, there is still a need for improvement of documentation. Only 10% of the User Study papers with accessible datasets provide source code, with or without instructions, likely because data from User Study papers does not typically require computationally intensive processing.

For the User Study datasets that consist of just a codebook, we label documentation as Minimal or None if the descriptions of the codes are not provided. An exemplar of Thorough documentation is work by Bernd et al. [25] who explored power dynamics with smart homes and nannies. The authors externally provide over 100 pages of additional documentation sharing extensive codebooks with descrip-

tions, and research protocols including recruitment tools, interview protocols, and interview questions. Additionally, the artifacts are well documented with a table of contents allowing specific resources to be easily accessible.

**Finding 7.** *While the only documentation required for qualitative codebooks is code descriptions, consistent practices do not exist for creating and sharing the documentation.*

**(4) Understandability.** Data from User Study papers often has a higher level of understandability compared to Security and Measurement papers. Figure 5 highlights that over half of the shared datasets from User Study papers can be understood without the context of the original paper and less than 10% suffer from being incomprehensible. This is driven by the format of codebooks and other qualitative sharing methods. Because codebooks are often small and straightforward, the only factor influencing their understandability is the existence of quality descriptions and example quotes.

Work by Stephenson et al. [26] that investigated IoT-enabled intimate partner violence contributes an example of an easily understandable codebook. They provide a breakdown of codes using strong code descriptions and counts of participants per code. We did identify a portion of the codebooks that are not understandable primarily due to non-existent code descriptions or not easily readable structures.

Across User Study papers, we observe accessible, well-documented, and understandable datasets when compared to Security and Measurement papers. Still, there exist areas for improvement related to the frequency at which codebooks are made available as well as thoughtful descriptions for documentation. As a community, further effort toward improving codebooks from User Study papers can help to provide additional value and context to shared work. To assist in this effort we provide recommendations and templates for sharing qualitative codebooks in Section 7.2.

**Finding 8.** *User Study data, primarily in the form of codebooks, are more accessible and understandable compared to Security and Measurement datasets, but they still lack standard sharing methods.*

## 5. Dataset Reuse Analysis

To quantify the extent of data reuse in Security, Measurement, and User Study papers, we analyze the citations received by 92 datasets (48 Security, 17 Measurement, and 27 User Study). The full results of our dataset reuse analysis are provided in Figure 6. We manually categorize the Security, Measurement, and User Study datasets based on the type of data they consist of to guide our investigation. Dataset reuse within these communities is overall very low. 41% of the datasets have not been reused at all, and the median reuse of the remaining datasets is 2.

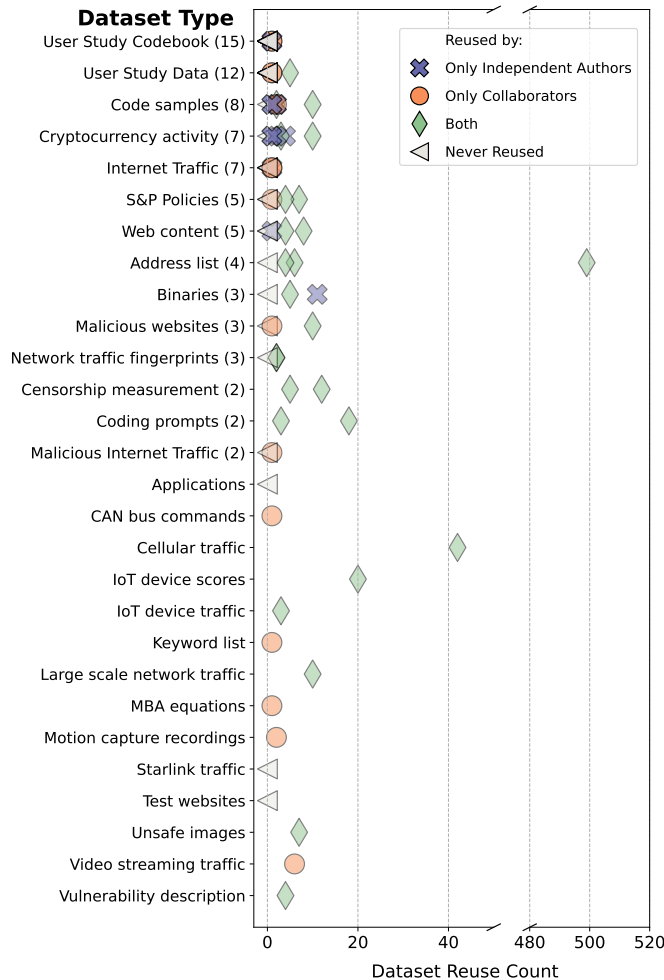


Figure 6: Results of the dataset reuse analysis. Each y-axis label represents a dataset type, with the number of datasets of that type indicated by (*n*). If no (*n*) is shown, one dataset of that type was present. The data points indicate how often the dataset was reused as well as if the reuse was conducted by only independent authors, collaborators, or both. The most reused dataset is an address list with over 10x the amount of reuse compared to the second most reused.

## 5.1. Security and Measurement Datasets

We first examine what variables may influence dataset reuse, analyzing the effects of publication year, conference, dataset file count, and dataset size. For dataset size, we use Spearman’s rank correlation coefficient [27] to test for relationships with dataset reuse. Dataset size is highly variable across datasets ranging from 5KB to over 4TB.

We consider file count a categorical variable and categorize it as either having more than five files or five or fewer. Calculating the exact file count in large datasets is challenging and our focus is on whether having a smaller, quickly readable file count (i.e., fewer than five files) af-

Categorical Variable	Fisher’s Exact <i>p</i> -value	Mood’s Median <i>p</i> -value
Publication Year	0.65	0.83
Publication Venue	0.58	0.5
Number of Files	0.57	0.25

Continuous Variable	Spearman $\rho$	<i>p</i> -value
Dataset Size	-0.05	0.73

TABLE 2: Results of the statistical analysis on what variables affect dataset reuse. We do not find evidence that any of the variables have an effect as all of the *p*-values are above 0.005.

fects reuse rates. For the categorical variables – file count, publication year, and conference – we apply Fisher’s Exact Test [28], [29] and Mood’s Median Test [30]. Fisher’s Exact Test assesses whether the proportion of reused versus never reused datasets differs across categorical values, while Mood’s Median Test evaluates if the median dataset reuse count is consistent across these values.

Based on our statistical tests (shown in Table 2), we do not find evidence that any of the variables have an effect on the reuse level of a dataset as all the *p*-values are well above 0.005.<sup>4</sup> This indicates that size and complexity might not hinder dataset reuse.

Next, we consider how dataset type may influence reuse. Internet traffic is one of the most common dataset types in our analysis, but they have one of the lowest reuse rates. Each of the seven benign and two malicious Internet traffic datasets has been used at most once. In contrast, we observe several follow-on studies that cite these original works but still craft their own datasets. The lack of reuse of Internet traffic datasets may be due to the dynamic nature of the Internet. Researchers may require up-to-date information when studying Internet behavior and, as a result, collect their own data. Additionally, collecting Internet measurement data was simplified in 2013 with the release of the customizable Internet scanner, ZMap [32], [33].

The most reused dataset is *Tranco*, a ranked list of the one million most popular domains [20]. The dataset which was reused 499 times, is simple, and regularly updated. The next most reused dataset is the *Lumos5G* dataset created by Narayanan et al. [34] which was reused 42 times. It is the only cellular traffic dataset in our analysis, and none of the citing papers create a similar cellular traffic dataset. Because a scanner tool like ZMap does not exist for cellular traffic, it is challenging to collect which encourages researchers to reuse the existing dataset. Reuse patterns vary for other dataset types. For example, our analysis includes three datasets composed of binaries; one was reused 11 times; another was not reused at all and the most frequently reused binary dataset was published more recently.

Creating and sharing datasets specifically for reuse is a challenging effort that currently offers few incentives and,

4. Modern experts encourage using  $p < 0.005$  to determine statistical significance as opposed to the less restrictive  $p < 0.05$  [31]

as our results show, often yields limited success. While the research community provides some motivation for data sharing through AECs, this only applies to sharing data for reproducibility purposes. Although all research data should be shared for reproducibility, data sharing for reuse should be done selectively and intentionally.

**Finding 9.** *Dataset size, structure and publication details do not have a significant effect on overall dataset reuse. However, the nature of specific dataset types influences the level of reuse (e.g., Internet traffic datasets experience almost no reuse).*

## 5.2. User Study Datasets

As shown in Figure 6, reuse patterns for data from User Study papers are particularly low whether the studies share a codebook or additional data. The only User Study dataset reused more than once is from a paper by Redmiles et al. [35], who share a database of security advice and a codebook generated from their interviews with security experts. Although this dataset was initially available at the start of our work, the link is now inactive, but the dataset was reused five times before the link became inactive. The most reused User Study artifact is a non-data tool developed by Faklaris et al. [36] consisting of a six-question metric to evaluate the security attitudes of user study participants. The metric has been reused by 26 citing papers.

For popular User Study topics, datasets often remain unused, despite high research interest. For example, the work by Utz et al. [37], which examines how users interact with GDPR consent notices, had over 400 citations at the time of our analysis with no reuse of their codebook.

Some artifacts from User Study papers have broader reuse, as demonstrated by the six-question metric developed by Faklaris et al. [36]. However, the reuse of User Study codebooks for future research appears to be uncommon, even when related research is conducted. Despite this, codebooks offer valuable context for User Study papers. To promote their utility, we present specific recommendations for improving codebook documentation in Section 7.2.

**Finding 10.** *Reuse of qualitative codebooks is not common practice across similar studies.*

## 6. Datasheets Analysis

Documentation for the accessible datasets in our analysis is highly variable because the community does not have commonly used best practices for data sharing. We assess the suitability of the *Datasheets for Datasets* questions [3] to act as a community standard for dataset documentation by attempting to answer the questions for the 10 most reused datasets in our analysis. We also attempt to answer the questions for two well-known Machine Learning (ML) datasets, MNIST [38] and CIFAR-10 [39], to investigate whether

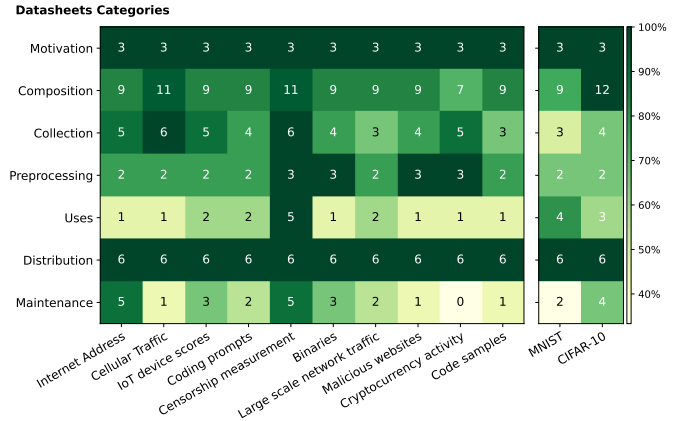


Figure 7: The proportion of answerable questions within each category of the *Datasheets for datasets* questions for the top ten most reused datasets from our analysis along with two highly used Machine Learning datasets. The amount of answerable questions is similar across all of the datasets regardless of reuse level.

widely reused datasets provide enough documentation to answer the *Datasheets* questions.

The *Datasheets* questions are organized into seven categories: Motivation (Q=3), Composition (Q=12), Collection Process (Q=6), Preprocessing/cleaning/labeling (which we refer to as Preprocessing) (Q=3), Uses (Q=5), Distribution (Q=6), and Maintenance (Q=6). To ensure consistency across the datasets we analyze, we exclude questions that are relevant only to datasets that relate to people. Figure 7 shows the number of answerable questions from each category for the most reused datasets and the two widely reused ML datasets. We are unable to answer all of the questions for any of the datasets because the information for the question was not provided, the information is unclear, or the question is not applicable to the given dataset. The full results for the *Datasheets* analysis are provided in Table 6 in Appendix C.

Across all datasets, we can answer all Motivation and Distribution questions. The Motivation questions, which cover the dataset’s purpose, authors, and funding sources, are consistently addressed in the paper introducing each dataset. Similarly, Distribution questions, which cover aspects like dataset location, ownership, and accessibility, are straightforward to answer because these datasets have already been publicly distributed.

Other categories present greater challenges. Some Composition questions apply only to ML, while others inquire about known issues or missing information in the dataset. When authors provide all known information about the dataset but are unaware of any issues or missing elements, these questions remain unaddressed. Among the Collection Process questions, most authors do not discuss ethical concerns, likely because most of the datasets we analyze do not involve human subjects and thus do not require Institutional Review Board oversight. The ML questions may not be relevant to all datasets, but ethical concerns should be addressed

whether a dataset involves human subjects or not.

Geburu et al. state that the information from the Uses questions is meant to provide guidance for future users to limit potential harm that may come from using the dataset incorrectly. Prior work by Layton et al. [40] has highlighted this issue further by showing how the imbalance in deepfake datasets has real-world impacts on the detectors that use the data. We find that dataset creators often center their documentation around the dataset’s content and current applications and lack information about the potential misuse of their dataset. While the inability to answer the Uses questions does not impede the reuse of the dataset, we encourage future dataset creators to address the Uses questions if they intend for the dataset to be reused in order to mitigate harm.

For datasets generated through ongoing measurement platforms, authors generally provide sufficient documentation to answer the Maintenance questions. Conversely, datasets resulting from a single data collection process are less likely to include maintenance plans, as future updates are typically not anticipated. As a result, the Maintenance questions that are relevant for measurement platforms differ from ones relevant to a one-time data collection dataset.

**Finding 11.** *Even highly reused datasets do not provide enough documentation to answer all of the Datasheets for Datasets questions.*

The questions we are able to answer for the 10 most reused datasets are comparable to the questions we are able to answer for MNIST and CIFAR-10. This suggests that the inability to answer some of the *Datasheets* questions for a dataset does not impede the reusability of that dataset.

**Finding 12.** *No strong correlation exists between providing enough information to answer all of the Datasheets for Datasets questions and the amount of dataset reuse.*

## 7. Recommendations

Our recommendations center on best practices for data sharing aimed at reuse. However, authors sharing data for reproducibility who wish to provide comprehensive information about their datasets will also find these guidelines valuable. We begin with general recommendations applicable to all types of datasets, followed by specific guidance on creating detailed qualitative codebooks.

### 7.1. General

Our general recommendations are a modified version of *Datasheets for Datasets*, but we also provide guidelines beyond documentation based on our experience evaluating 265 accessible datasets. A list of the questions we add to *Datasheets* is shown in Table 3. A full version of our modified *Datasheets* is provided through our website<sup>5</sup>, along

5. <https://sites.google.com/view/data-to-infinity-beyond/home>

<i>Datasheets Category</i>	<i>New Questions</i>	
<b>Motivation</b>	Do datasets of the same type exist? Describe how this dataset differs from related work.	
	Was the dataset shared for the purpose of reproducibility, reuse, or both?	
<b>Composition</b>	What is the file structure and file naming scheme of the dataset?	
	How do the instances within the dataset relate to the file structure?	
	For each file type, what are the fields/columns in each? Provide a brief description of each file/column.	
<b>Collection Process</b>	No changes	
<b>Preprocessing</b>	No changes	
<b>Uses</b>	Does source code exist for processing or analyzing the data? Please provide the location of any relevant source code.	
<b>Distribution</b>	Provide the location of all resources relevant to the dataset including links to project websites, documentation, and source code.	
<b>Maintenance</b>	Is the dataset the result of an ongoing measurement platform or one-time data collection?	
	<b>Measurement platform:</b>	<b>One time collection:</b>
	How often does data collection occur?	Will the dataset be maintained beyond the initial share?
	At what point after data collection is the data published?	Who will be hosting the dataset and how can they be contacted?
	Will older versions of the dataset be supported?	
	Who supports the operation of the measurement platform and hosts the data? How can they be contacted?	

TABLE 3: The new questions provided in our modified *Datasheet* developed based on our analysis of 265 datasets.

with an example *Datasheet* created for the dataset used in this work. During our analysis, we found that the lack of information to answer some of the *Datasheets* questions does not necessarily impede the reuse of a dataset. However, we do not remove any of the *Datasheets* questions from our example, as answering the questions would provide valuable information about a dataset. We encourage authors to answer as many questions as possible or acknowledge which questions are not relevant in order to best equip future users of their dataset.

Since data sharing for reproducibility and reuse requires different levels of detail, we add a question in the Motivation section that allows authors to clarify their data sharing goals. We also add a question asking authors to discuss similar existing datasets, as our analysis revealed some instances of new datasets being created without clear reason. We urge authors to differentiate from existing datasets.

For datasets we deem not understandable, the primary



obstacles are complex, unexplained file structure and unclear codes or variables within the data. To address this, we add two questions in the Composition category: one for detailing the file structure and naming conventions, and another for defining the variables used in the files.

In the Uses and Distribution sections, we add one new question each. For the Uses category, we recommend authors indicate whether analysis scripts are available. As mentioned, analysis scripts provide an important depth of information for datasets, and dataset users should have easy access to these resources. For the Distribution section, we include a question asking for a comprehensive list of all relevant links or web pages associated with the dataset. During our analysis of all accessible datasets, we found that resources for a single dataset are sometimes scattered across multiple web pages, with some not linked to the original paper in any way. Centralizing all links in one clearly labeled place helps ensure access for future users.

Finally, we restructure and adapt the original Maintenance questions to accommodate the different maintenance needs of measurement platforms versus one-time data collection efforts. Our revised structure lets authors specify whether data collection is complete or ongoing, with tailored follow-up questions based on the response. This ensures that maintenance expectations are clear regardless of the dataset collection status.

**Recommendation 1.** *Because quality documentation alleviates some barriers to dataset reusability, we encourage authors to include our modified Datasheet or equivalent material with their dataset.*

For some datasets we label as not understandable, the nature of the dataset is a contributing factor. Extremely large datasets compressed into a single downloadable file, datasets stored as machine-readable files such as `pkl` or `sql`, or datasets consisting of machine-readable content such as hashes or smart contract addresses require a level of effort to understand that other datasets do not. While these features of a dataset may not be avoidable, authors can support the reusability of their dataset by providing a well-formatted and documented sample of their dataset. A sample allows dataset users to understand if the dataset is appropriate for their work before investing extensive time and resources into the dataset.

**Recommendation 2.** *Datasets that cannot be fully accessed easily should be accompanied with a sample of data to help dataset users understand if the data is appropriate for their work.*

Data sharing for the purpose of reuse involves supporting future dataset users as much as possible. The extra time and resources that go into providing quality datasets for the research community will save resources for the community over time by limiting the creation of redundant datasets. The benefit of quality datasets for Computer Science research has already been demonstrated with the creation of ImageNet in

2009. ImageNet is a large-scale image database that changed the way machine learning models are designed because Deng et al. [41], [42] understood that quality models require quality data. The effort needed to provide quality datasets for reuse is not insignificant, and therefore should further be incentivized by the research community by allowing dataset creation alone to be a sufficient contribution for publication.

## 7.2. Qualitative Codebooks

When sharing a qualitative codebook, there are considerations authors should make to maximize utility. During our analysis, a primary factor in a codebook being labeled as not understandable is sharing codes with no descriptions or abbreviated codes without explanation. In other cases, descriptions are insufficient to be independently understood, often requiring the context of interview questions or the manuscript. In contrast, strong, independently understandable codebooks provide descriptions that appropriately describe how to identify a code, with the addition of a quote as an example for further clarification.

We recommend that codebooks provide codes, strong descriptions of the codes, and quotes to support how the code is used in practice. Additionally, sharing parent codes, or mapping questions to codes can help provide context. This allows for research to be reproduced and validated, or even reusable for future User Study papers with a similar context across varying populations.

Second, we recommend that all studies provide either participant counts for each code and when appropriate, participant-by-participant codebook mapping. Participant-level mappings may not be feasible for every study due to risks of deanonymization and Institutional Review Board (IRB) guidelines. Providing numbers within codebook data allows these resources to be used in future work to form hypotheses for tangential studies. For example, a study looking at security concerns with incarcerated populations might look toward other technology studies with the same population to form preliminary hypotheses, with the relative prevalence of codes aiding study design.

Because every aspect of a user study requires approval by an IRB, authors must decide what information they will publicly share before the data is collected. As a result, we encourage authors be intentional about their data-sharing goals during the study design phase and consider the value that sharing qualitative data provides to the research community. To support our recommendations, we provide codebook templates on our website and an example codebook in Appendix B.

**Recommendation 3.** *User Study papers should provide a well structured and descriptive codebook to provide context for their own results and support future research in similar areas.*



## 8. Limitations and Discussion

To the best of our ability, we attempt to minimize biases and identify limitations in this work. We acknowledge that there are still several limitations, biases, and avenues for future work, which we outline in this section.

We aim to systematically select papers using the methodology outlined in Section 3 and confirm the papers through consensus between two reviewers. We also require consensus between two raters for subjective aspects of our coding process such as the degree of detail in documentation and the understandability of datasets. To enhance objectivity where possible, we record specific observations from each dataset (e.g., repository missing a description for dataset fields, errors in unzipping files).

Our analysis focuses on the accessibility of datasets for reuse, rather than reproducibility, and we discuss accessible datasets within this context. We recognize that some authors may have intended their datasets solely for reproducibility. However, due to limited discussion in the Security community on the difference between reproducibility and reusability, we do not attempt to make this distinction.

This work centers on datasets accessible to independent researchers without additional assistance from the authors and whose access information is included in the original publications. This approach excludes datasets that are available upon request and those initially developed for publication but made publicly accessible later. Reaching out to authors for access to datasets or investigating datasets promised but not directly linked in publications could provide further insights into data sharing practices, which we leave as a potential direction for future research.

Our data collection was conducted before new requirements were introduced in the Security community's submission process regarding research artifacts. The *CCS 2024* [12] and *USENIX Security 2025* [11] Calls for Papers (CFPs) introduced new mandates around artifact availability. Additionally, AECs for both venues require that artifacts be hosted on platforms that ensure permanent accessibility, recommending institutional repositories or public repositories such as Zenodo, FigShare, Dryad, or Software Heritage. Notably, these guidelines explicitly prohibit hosting artifacts on GitHub or project websites.

These new requirements are likely to significantly reshape data sharing practices in the Security community, as our findings indicate that GitHub has been the primary hosting platform over the past five years. In our analysis, Zenodo was used to host 6 datasets, FigShare hosted 1, and neither Dryad nor Software Heritage were used. We chose to host our dataset from this work on the Open Science Framework (OSF) repository, which was used by 10 datasets in our work, due to the Center for Open Science's commitment to maintaining data on OSF for at least 50 years [43]. Additionally, OSF enhances dataset accessibility by allowing users to preview data directly on the platform, a feature not available on all public repositories. Future research will be needed to assess how these new requirements fully influence data sharing practices within the community.

## 9. Related Work

The requirements around data sharing practices in health and science journals have led to research examining the availability of datasets in these fields. Researchers have examined the frequency of data availability statements in publications that generate datasets [8] and assessed the comprehensiveness and reusability of publicly available datasets, evaluating factors such as documentation, format, and accessibility [8], [44], [45]. A significant area of interest has also been the reproducibility of findings, particularly the challenges posed by incomplete publicly available data [46]. Additionally, for datasets that are not accessible, researchers directly engaged with authors to measure their responsiveness to requests for data [44], [47], [48], [49].

Although data sharing practices in the Computer Security community have not received the same level of attention as in other fields, some works have addressed specific concerns about data availability in Security. Zheng et al. [50] surveyed 965 Computer Security and Computer Measurement publications from 2012-2016 to identify how many use existing datasets versus creating new datasets; they also checked to see if the links to the new datasets were still live. They found that only 6% of the surveyed papers created and publicly shared a dataset. Cremer et al. [51] gathered information on cybersecurity datasets from journal publications and discussed how publicly available cybersecurity data affects the insurance industry's management of cyber risks. Finally, Olszewski et al. [52] reviewed the availability of datasets used to train and test machine learning systems published in Computer Security conferences from 2013-2022 and measured how the availability of these datasets affects the reproducibility of machine learning research. Our work focuses exclusively on new datasets created for purposes beyond machine learning and provides the first comprehensive analysis of their accessibility and reuse.

## 10. Conclusion

Data sharing specifically for reuse involves significant effort as it requires dataset creators to provide enough documentation to allow future users to comprehend the entire context of the data. Therefore, researchers should be intentional when sharing data, both in how they share and in terms of their goals for sharing. To understand the state of data sharing practices in the Computer Security community as well as the level of data reuse, we perform the first longitudinal study of publicly available datasets. Through our analysis of 948 papers that create a dataset as a contribution, we find that most papers do not provide a statement about the availability of their data. Furthermore, of the 265 accessible datasets, the practices for hosting and documenting are inconsistent, causing some datasets not to be understandable. We also find that the overall reuse of accessible datasets is minimal, especially for specific dataset types (e.g., Internet traffic datasets or qualitative codebooks). For authors who wish to make their datasets reusable, we

provide recommendations for dataset documentation, formatting practices, and the structure of qualitative codebooks. While this work represents the first step toward understanding the state of data sharing and reuse in Computer Security, the adoption of recommendations and best practices could move the Security community to better incorporate existing datasets to the benefit of all researchers.

## Acknowledgments

The authors thank our anonymous reviewers and our shepherd for their valuable comments and suggestions. This work was supported in part by the National Science Foundation under grants CNS-2446321 and CNS-2206950.

## References

- [1] C. S. Marcum and R. Donohue, “New Guidance to Ensure Federally Funded Research Data Equitably Benefits All of America,” May 2022.
- [2] “ACM IMC 2016 - Call for Papers,” <https://conferences.sigcomm.org/imc/2016/cfp.html>, 2016.
- [3] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, “Datasheets for datasets,” *Communications of the ACM*, 2021.
- [4] L. Koesten, P. Vougiouklis, E. Simperl, and P. Groth, “Dataset Reuse: Toward Translating Principles to Practice,” *Patterns*, 2020.
- [5] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, 2016.
- [6] “Joint Data Archiving Policy (JDAP),” *DRYAD*, 2020.
- [7] S. A. Sloman, “Opening editorial: The changing face of Cognition,” *Cognition*, 2015.
- [8] T. E. Hardwicke et al., “Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition,” *Cognition Royal Society Open Science*, 2018.
- [9] PLOS, “Data availability,” 2019. [Online]. Available: <https://journals.plos.org/plosone/s/data-availability>
- [10] S. Krishnamurthi, “About artifact evaluation.” [Online]. Available: <https://artifact-eval.org/about.html>
- [11] “USENIX security ’25 call for papers,” 2024. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity25/call-for-papers>
- [12] “CCS ’24 call for papers,” 2024. [Online]. Available: <https://www.sigsac.org/ccs/CCS2024/call-for/call-for-papers.html>
- [13] M. Pushkarna, A. Zaldivar, and O. Kjartansson, “Data cards: Purposeful and transparent dataset documentation for responsible ai,” in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.
- [14] A. Goodman, A. Pepe, A. W. Blocker, C. L. Borgman, K. Cranmer, M. Crosas, R. Di Stefano, Y. Gil, P. Groth, M. Hedstrom, D. W. Hogg, V. Kashyap, A. Mahabal, A. Siemiginowska, and A. Slavkovic, “Ten Simple Rules for the Care and Feeding of Scientific Data,” *PLoS Computational Biology*, 2014.
- [15] N. McDonald, S. Schoenebeck, and A. Forte, “Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, 2019.
- [16] K. Roberts, A. Dowell, and J.-B. Nie, “Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development,” *BMC Medical Research Methodology*, 2019.
- [17] S. D. L. C. for Informatics, “dblp: computer science bibliography,” 1993. [Online]. Available: <https://dblp.uni-trier.de/>
- [18] “dataprep,” 2020. [Online]. Available: [https://docs.dataprep.ai/user\\_guide/connector/dblp.html](https://docs.dataprep.ai/user_guide/connector/dblp.html)
- [19] “Wayback machine,” 1996. [Online]. Available: <https://web.archive.org>
- [20] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen, “Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation,” in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2019.
- [21] R. S. Raman, P. Shenoy, K. Kohls, and R. Ensafi, “Censored Planet: An Internet-wide, longitudinal censorship observatory,” in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2020.
- [22] V. Nair, W. Guo, J. Mattern, R. Wang, J. F. O’Brien, L. Rosenberg, and D. Song, “Unique identification of 50,000+ virtual reality users from head & hand motion data,” in *Proceedings of the USENIX Security Symposium*, 2023.
- [23] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, “The dataset nutrition label: A framework to drive higher data quality standards,” 2018. [Online]. Available: <https://arxiv.org/abs/1805.03677>
- [24] N. Perry, M. Srivastava, D. Kumar, and D. Boneh, “Do users write more insecure code with AI assistants?” in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2023.
- [25] J. Bernd, R. Abu-Salma, J. Choy, and A. Friik, “Balancing power dynamics in smart homes: nannies’ perspectives on how cameras reflect and affect relationships,” in *Proceedings of the USENIX Symposium on Usable Privacy and Security (SOUPS)*, 2022.
- [26] S. Stephenson, M. Almansoori, P. Emami-Naeini, and R. Chatterjee, ““it’s the equivalent of feeling like you’re in jail”: Lessons from firsthand and secondhand accounts of IoT-enabled intimate partner abuse,” in *Proceedings of the USENIX Security Symposium*, 2023.
- [27] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*. John Wiley & Sons, 1973, pp. 1991–1992.
- [28] R. A. Fisher, “The logic of inductive inference,” *Journal of the Royal Statistical Society*, 1935.
- [29] N. Vasavada, “Fisher’s test for exact count data calculator, with follow-up chi-squared test,” 2016. [Online]. Available: <https://astatsa.com/FisherTest/>
- [30] A. M. Mood, *Introduction to the Theory of Statistics*. McGraw-Hill, 1974.
- [31] D. J. Benjamin and J. O. Berger, “Three Recommendations for Improving the Use of  $p$ -Values,” *The American Statistician*, 2019.
- [32] Z. Durumeric, E. Wustrow, and J. A. Halderman, “ZMap: Fast Internet-wide Scanning and its Security Applications,” in *Proceedings of the USENIX Security Symposium*, 2013.
- [33] Z. Durumeric, D. Adrian, P. Stephens, E. Wustrow, and J. A. Halderman, “Ten Years of ZMap,” in *Proceedings of the Internet Measurement Conference (IMC)*, 2024.
- [34] A. Narayanan, E. Ramadan, R. Mehta, X. Hu, Q. Liu, R. A. K. Fezeu, U. K. Dayalan, S. Verma, P. Ji, T. Li, F. Qian, and Z.-L. Zhang, “Lumos5g: Mapping and predicting commercial mmwave 5g throughput,” in *Proceedings of the Internet Measurement Conference (IMC)*, 2020.
- [35] E. M. Redmiles, N. Warford, A. Jayanti, A. Koneru, S. Kross, M. Morales, R. Stevens, and M. L. Mazurek, “A comprehensive quality evaluation of security and privacy advice on the web,” in *Proceedings of the USENIX Security Symposium*, 2020.
- [36] C. Faklaris, L. Dabbish, and J. I. Hong, “A self-report measure of end-user security attitudes (sa-6),” in *Proceedings of the USENIX Symposium on Usable Privacy and Security (SOUPS)*, 2019.

- [37] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, “(un)informed consent: Studying gdpr consent notices in the field,” in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2019.
- [38] Y. LeCun, C. Cortes, and C. J. Burges, “Gradient-based learning applied to document recognition,” in *Proceedings of the IEEE*, 1998.
- [39] A. Krizhevsky, “Learning Multiple Layers of Features from Tiny Images,” 2009.
- [40] S. Layton, T. Tucker, D. Olszewski, K. Warren, K. Butler, and P. Traynor, “Sok: The good, the bad, and the unbalanced: Measuring structural limitations of deepfake media datasets,” in *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2024.
- [41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings for the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [42] D. Gershgorn, “The data that transformed AI research—and possibly the world,” 2017. [Online]. Available: <https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world>
- [43] C. for Open Science, “Manage and share your research with osf - an easy, integrated platform.” [Online]. Available: <https://www.cos.io/products/osf>
- [44] A. F. Magee, M. R. May, and B. R. Moore, “The Dawn of Open Access to Phylogenetic Data,” *PLOS ONE*, 2014.
- [45] D. G. Roche, L. E. Kruuk, R. Lanfear, and S. A. Binning, “Public data archiving in ecology and evolution: How well are we doing?” *PLOS Biology*, 2015.
- [46] P. Obels, D. Lakens, N. A. Coles, J. Gottfried, and S. A. Green, “Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology,” *Advances in Methods and Practices in Psychological Science*, 2020.
- [47] N. Milia, A. Congiu, P. Anagnostou, F. Montinaro, M. Capocasa, E. Sanna, and G. D. Bisol, “Mine, Yours, Ours? Sharing Data on Human Genetic Variation,” *PLoS ONE*, 2012.
- [48] L. Tedersoo, R. Küngas, E. Oras, K. Köster, H. Eenmaa, Ä. Leijen, M. Pedaste, M. Raju, A. Astapova, H. Lukner, K. Kogermann, and T. Sepp, “Data sharing practices and data availability upon request differ across scientific disciplines,” *Scientific Data*, 2021.
- [49] J. M. Wicherts, D. Borsboom, J. Kats, and D. Molenaar, “The poor availability of psychological research data for reanalysis,” *APA PsycNet*, 2006.
- [50] M. Zheng, H. Robbins, Z. Chai, P. Thapa, and T. Moore, “Cybersecurity research datasets: Taxonomy and empirical analysis,” in *Proceedings of the USENIX Workshop on Cyber Security Experimentation and Test (CSET)*, 2018.
- [51] F. Cremer, B. Sheehan, M. Fortmann, A. N. Kia, M. Mullins, F. Murphy, and S. Materne, “Cyber risk and cybersecurity: A systematic review of data availability,” *The Geneva Papers on Risk and Insurance - Issues and Practice*, 2022.
- [52] D. Olszewski, A. Lu, C. Stillman, K. Warren, C. Kitroser, A. Pascual, D. Ukirde, K. Butler, and P. Traynor, ““Get in Researchers; We’re Measuring Reproducibility”: A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences,” in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2023.

platforms used by the accessible datasets in our analysis. Some datasets are hosted across multiple platforms. Institution repo refers to private repositories run by an institution such as an university or research group.

File Type	Security	Measurement	User Study	Total
compressed file	49	20	5	74
csv	28	18	12	58
txt	25	8	3	36
json/l	17	7	2	26
pdf	2	-	21	23
xlsx	8	-	2	10
pkl	5	3	1	9
html	3	-	2	5
webpage	3	1	-	4
tsv	3	-	1	4
sql	2	2	-	4
log	1	2	1	4
png	2	1	-	3
js	1	-	1	2
pcap	-	2	-	2
pem	-	2	-	2
docx	-	-	2	2
apk	1	-	-	1
avro	1	-	-	1
bin	1	-	-	1
C programs	1	-	-	1
core	1	-	-	1
db	1	-	-	1
list	1	-	-	1
map	1	-	-	1
supp	1	-	-	1
tex	1	-	-	1
wav	1	-	-	1
dat	-	1	-	1
graph	-	1	-	1
graphml	-	1	-	1
har	-	1	-	1
jl	-	1	-	1
keylog	-	1	-	1
mkv	-	1	-	1
ods	-	1	-	1
sha512	-	1	-	1
wireless data files	-	1	-	1
ftt	-	-	1	1
sav	-	-	1	1

TABLE 4: File types

## Appendix A.

Table 4 shows the file types used by the accessible datasets in our analysis. Some datasets are shared as multiple file types. This table does not include the data that is shared in the Appendix of the paper. Table 5 shows the hosting

Platform	Security	Measurement	User Study	Total
Github	57	22	16	95
Appendix	6	2	67	75
Institution repo	15	11	10	36
Google	23	7	5	35
Project website	10	2	-	12
OSF	2	-	8	10
Zenodo	6	-	-	6
Dropbox	3	-	-	3
Sharepoint	3	-	-	3
FigShare	1	-	-	1
Hugging Face	1	-	-	1
Internet Archive	1	-	-	1
Amazon AWS	-	1	-	1
Box	-	1	-	1
Mega	-	1	-	1
4TU	-	-	1	1
Bitbucket	-	-	1	1
PasteBin	-	-	1	1
Sciebo	-	-	1	1

TABLE 5: Hosting platforms

## Appendix B.

Figure 8 is a sample codebook created using one of our codebook templates. The codebook includes codes, strong descriptions of the codes, quotes to provide examples, codes mapped to participants and counts. Additional templates and code to generate similar codebooks are provided on our website.

Code	Description	Quote	P1 P2 P3 P4 P5 P6 P7 P8 P9	Count
sec_communication	When discussing security concerns; a participant mentioned lack of communication	"I think we would have responded to that incident much faster if we had better communication"	● ○ ● ● ○ ● ○ ○ ●	5
sec_education	When discussing security breaches; a participant mentioned education as a source	"I don't think that we would have been breached if our team was more informed of threats"	● ○ ● ○ ● ● ○ ● ○	5
sec_impede	Security was mentioned as impeding progress	"I want to implement security but it will slow down production"	● ○ ● ○ ○ ● ● ○ ○	4
phish_attack	Phishing attacks were mentioned as a concern	"I just hope my employees don't get phished"	● ○ ○ ● ○ ● ○ ○ ●	4
phish_education	In the context of phishing; education was identified as a need	"I don't think anyone has ever taught us how to detect a phishing email"	● ● ○ ● ○ ○ ○ ○ ●	4
phish_awareness	Raising awareness about phishing risks was discussed	"We need to make everyone more aware of phishing threats"	○ ● ● ○ ● ● ○ ○ ●	5
phish_prevention	Phishing prevention strategies were emphasized as important	"Implementing stronger anti-phishing measures is a priority for us"	● ○ ● ○ ○ ● ○ ● ●	5





Figure 8: Sample codebook

## Appendix C.

Table 6 below provides the full results of the *Datasheets for Datasets* analysis. The analysis covers the 10 most reused datasets in our collection, and two highly used ML datasets. The datasets include the following types:

- D1 - Internet Address List
- D2 - Cellular Traffic
- D3 - Malicious Content
- D4 - Cryptocurrency Activity
- D5 - IoT Device Scores
- D6 - Coding Prompts
- D7 - Binaries
- D8 - Code Samples
- D9 - Censorship Measurement
- D10 - Large-Scale Network Traffic
- D11 - MNIST
- D12 - CIFAR-10

		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
Motivation	For what purpose was the dataset created?												
	Who created the dataset?												
	Who funded the creation of the dataset?												
Composition	What do the instances that comprise the dataset represent?												
	How many instances are there in total (of each type, if appropriate)?												
	Does the dataset contain all possible instances or is it a sample of instances from a larger set?												
	What data does each instance consist of? "Raw" data or features?												
	Is there a label or target associated with each instance?												
	Is any information missing from individual instances?												
	Are relationships between individual instances made explicit?												
	Are there recommended data splits (e.g., training, development/validation, testing)?												
	Are there any errors, sources of noise, or redundancies in the dataset?												
	Is the dataset self-contained, or does it link to or otherwise rely on external resources?												
	Does the dataset contain data that might be considered confidential?												
	Does the dataset contain data that might be offensive, threatening, or might cause anxiety?												
Collection Process	How was the data associated with each instance acquired?												
	What mechanisms or procedures were used to collect the data?												
	If the dataset is a sample from a larger set, what was the sampling strategy?												
	Who was involved in the data collection process and how were they compensated?												
	Over what timeframe was the data collected?												
	Were any ethical review processes conducted (e.g., by an institutional review board)?												
Preprocessing	Was any preprocessing of the data done?												
	Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?												
	Is the software that was used to preprocess/clean/label the data available?												
Uses	Has the dataset been used for any tasks already?												
	Is there a repository that links to any or all papers or systems that use the dataset?												
	What (other) tasks could the dataset be used for?												
	Is there anything about the composition of the dataset that might impact future uses?												
	Are there tasks for which the dataset should not be used?												
Distribution	Will the dataset be distributed to third parties?												
	How will the dataset be distributed? Does the dataset have a (DOI)?												
	When will the dataset be distributed?												
	Will the dataset be distributed under a copyright or other IP license, and/or ToU?												
	Have any third parties imposed restrictions on the data associated with the instances?												
	Do any export controls apply to the dataset or to individual instances?												
Maintenance	Who will be supporting the dataset?												
	How can the owner/curator/manager of the dataset be contacted (e.g., email address)?												
	Is there an erratum?												
	Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?												
	Will older versions of the dataset continue to be maintained?												
	If others want to contribute to the dataset, is there a mechanism for them to do so?												

TABLE 6:  – Question Answered  – Question Not Answered  – Unclear Information  – Not applicable



## **Appendix D. Meta-Review**

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

### **D.1. Summary**

This paper presents a study on the practices of data sharing and reuse in the security and measurement communities. The work performs this through a pipeline that first evaluates (1) how many papers had datasets as a contribution, (2) how many papers had availability statements, (3) how many datasets were accessible, documented, and understandable, (4) how many of these datasets were reused, and lastly (5) how these datasets compare to datasheet standards.

### **D.2. Scientific Contributions**

- Independent Confirmation of Important Results with Limited Prior Research
- Addresses a Long-Known Issue
- Provides a Valuable Step Forward in an Established Field
- Provides a New Data Set For Public Use
- Establishes a New Research Direction
- Creates a New Tool to Enable Future Science

### **D.3. Reasons for Acceptance**

- 1) The paper addresses the long-known but understudied issue of data sharing practices, confirming that publicly accessible datasets are frequently not understandable, and that consistent data sharing and documentation practices do not exist. The paper also provides a valuable step forward by sharing insights and recommendations for future authors producing public datasets as research artifacts.
- 2) The paper provides a new dataset for public use.
- 3) The paper provides a new tool and methodology to further examine data sharing practices, which will foster innovative approaches and new best practices.