Guidelines for Dataset Development

Prepared for

The Organization of Scientific Area Committees for Forensic Science (OSAC)

Prepared by:

OSAC Digital Evidence Subcommittee Task Group on Dataset Development

Date October 2022

OSAC
Organization of Scientific Area
Committees for Forensic Science

Guidelines for Dataset Development V2

Introduction

A dataset is a collection of files or data (obtained from a digital device) created to have a desired set of attributes and known content. In digital forensics, these datasets can be used for a variety of quality assurance purposes, including but not limited to: the testing and evaluation of tools and methods, training and performance monitoring of practitioners, and verification of artifacts or analysis findings. The quality and robustness of a test dataset is generally dependent upon the development documentation. A well-documented dataset facilitates more rigorous testing and reliable results.

The entire digital forensics community benefits when a large number and variety of quality test datasets are available. However, dataset creation can be cumbersome and time-consuming, and little guidance or standard practice documentation exists to support those wishing to generate test datasets.

This paper attempts to provide that guidance, in order to advise and assist interested parties in their efforts to produce high-quality, well-documented datasets, whether for individual, local use or for shared use among the broader digital forensic community. For those intending to utilize or evaluate datasets created by other individuals or organizations, see Use of Existing Datasets.

Scope

The purpose of this document is to provide guidance and resources for the development and use of datasets in digital forensics. The intended audience is potential developers and users of the datasets, to include: practitioners, tool/method developers, researchers, instructors, and students.

This document does not attempt to address all scenarios or purposes for which dataset development may be needed, nor does it define requirements for individual use cases. Dataset developers are encouraged to utilize the documentation guidance and template provided in this document to improve consistency and quality among datasets used in digital forensics.

The steps and considerations outlined in this document were drafted with a view to optimizing demands on personnel time and other resources while creating rigorous and detailed documentation consistent with the intended usage of the generated dataset.

Keywords

Digital forensics, datasets, quality assurance, testing, evaluation, verification

Use Cases for Datasets

A test dataset is a means for measuring how well a *forensic tool* accomplishes a task, a *forensic practitioner* accomplishes a task, or what artifacts are created in a given *software and hardware environment* for specific actions:

• Forensic Tool: Evaluating the tool's ability to do a specific task such as acquiring data without modifying the original data, computing a cryptographic hash, navigating a file system, finding and extracting a specific artifact, or other tasks. This type of dataset can often reveal the limitations of a tool in accomplishing a task.

• Forensic Practitioner: Evaluating how well a forensic practitioner can accomplish an assigned task. These datasets can be used for training (including as an aid in training non-technical stakeholders on review of data), educational use, capture the flag (CTF) type competitions, competency tests, and proficiency tests.

• Software and Hardware Environment: Examining the artifacts created by an operating system (OS) and applications running in a given hardware environment. This type of dataset has several uses. Examples include: investigating particular artifacts for the development of an extraction method for an artifact and the development of background information about an artifact for court testimony, research to support tool development.

Developers of these datasets should have a good understanding of tool limitations and the underlying device technologies so that a rich set of behaviors can be revealed by the dataset.

Dataset Development

The process of developing datasets consists of several phases: Planning, Setup/Preparation, Data Population, and Acquisition.

Careful documentation of each of these phases will result in a highly versatile and useful dataset.

This document can be used to provide instruction and templates for both inhouse datasets and for those designed for broader distribution. The template in Appendix A may be all that some users need.

Documentation

Beginning in the planning phase, thorough documentation of the dataset and dataset development process is critical to reliable and effective use of the data. This documentation should include:

- A description of the dataset, its intended use, and known limitations.
- A description of the relevant technical environment in which the dataset was created, as appropriate to its intended use, including:
    - The manufacturer, model, versions, and security patch level of the involved device(s), operating system(s), and software application(s).
    - Software installation methods (e.g., store vs. direct install vs. sideload install) and timestamps
    - Device and account configurations, permissions, and other settings including the presence of pre-existing data or conditions

> The level of detail should be appropriate to the type and intended use of the dataset.
>
> For example, a dataset to verify one specific artifact may not need much documentation while a dataset to support development of new tool capabilities may require extensive documentation.

    - Paired or synced devices (and their history)
    - Networking information such as network names or domains, Wi-Fi, cellular carriers
    - User Account information
    - Time zone and other locality information
- A description of tools or automated processes used including those that might affect the data such as the use of emulated devices
- Contemporaneous records of all steps taken in the creation of the dataset including timestamped user actions and, where appropriate, photos or screenshots of important key actions or information, such as:
    - Power events
    - Lock/unlock events
    - Settings modifications
    - Login/out events
    - Device/application/network access credentials (e.g., password/PIN)
    - Connection of external devices and networks including the date and time of connection(s) such as USB, Wi-Fi, Bluetooth
    - Sent and received communications
    - Sent and received data (e.g., files, media)
    - Storage events (e.g., file creation/modification/deletion)
    - Location-related events
    - Physical activity-related events (walking, running, sleeping, heart rate, etc.)
    - Operating system updates and patches
- A description of the methods used to extract or acquire the dataset from the host devices or accounts. When the method uses a proprietary acquisition utility, the description should include the name and version of the utility
- For sustained datasets that change over time: a history of modifications and events for the dataset
- A plan for the maintenance of sustained datasets (*see* Maintenance of Datasets)

When feasible or available, this data should be recorded during the planning, setup, and preparation phases prior to the beginning of dataset generation. Detailed preparation documentation, including a list of planned activity/interactions, facilitates an easier and more efficient data population process. Using a pre-configured worksheet can help guide this preparation and standardize the collection of this information (*see* Appendix A).

Planning

Planning is critical to creating robust datasets. This section presents key considerations. Depending on the use case for the dataset, many of them will not be relevant. Dataset generation and development is driven by the intended use of the dataset. Properly planning the steps involved before beginning dataset creation will help ensure the needed data points and artifacts are generated and available for subsequent review or testing. Planning should lead to a written description of both the steps needed to create the dataset and a test plan, including the procedures as step-actions will assist in achieving reliable, repeatable results.

Background Research

There are many sources of information that can be used to help plan creation of the dataset such as developer guides, user guides, app store descriptions and documentation, FAQs, privacy policies, online support resources, and discussion forums. In cases where additional information about the target dataset environment is needed, pre-development experimentation and testing may provide insights valuable for the creation of desired artifacts. This is often an iterative and recursive process, with new information gleaned supporting the refinement of the processes in the test plan.

Advance research of applications and application capabilities is recommended. Applications often take advantage of features available in one version of a particular operating system which are not available on different versions. The availability of these features may be dependent on the underlying hardware as well. The availability of features or capabilities of applications may vary among versions.

Permissions also significantly impact the type of information and functionality an application can access, and if inconsistent, even the same version of the same application may end up storing different types of information. Determine the permissions requirements of the application(s) of interest and configure as closely as possible in the dataset application utilized in development of the reference dataset. For example, if an application is denied access to the device location in one use case and is granted access in another, an examiner may find different artifacts in the dataset for that app, even if the same version and same test activities are utilized.

There are multiple questions to ask during planning, and particular care when the intent is to support analysis findings.

- What is the version of the OS/application?
- What are the permissions?
    - How do the permissions affect application/OS functionality?
    - Can 3rd party apps/OS access device location, camera, microphone, device storage?
    - What is the application's privacy policy
- Which Location Storage was selected (i.e., microSD card, cloud)?

There are other application settings of interest. An application may be configured to store information locally on a removable (i.e., microSD card) or non-removable storage device, or on remote servers. An application's settings may also include access to geolocation, microphone, SMS service, or contacts. Device settings such as Wi-Fi, Bluetooth, cellular, and time zone source may also be a consideration.

Environment Selection

Hardware, firmware, and software selection can have significant impacts on the availability of created data. For example, user data for many mobile applications is only available via a full file system extraction. Selecting a device for which the developer has this capability is necessary for accessing the created data. It might be prudent to perform a test extraction.

Hardware plays a critical role in test datasets in that it can affect the amount and content of data. Ensure the hardware being used to generate test data is as close as possible to the original. Applications may exhibit certain behaviors on a device due to a particular piece of hardware, and not exhibit that same behavior on another device lacking said hardware. OS platforms also play a critical role.

Datasets can be populated either on a real device or in emulation. The preference is always to test in the same environment as the evidence when applicable. The ideal situation would include the same configurations to include hardware, firmware, and software versions; however, it is not always possible to create the exact same environment. When the testing environment differs from the evidence environment, it should be documented. Since different environments may result in different data outputs, documentation should focus on those elements most likely to impact the findings.

Some reasons to opt for emulation include:

- Quicker processing
- Lower cost than hardware
- More accessible environments
- Ability to snapshot a state to use as a base for different scenarios

Development Considerations

If required, evaluate the need for allowing, inhibiting, or spoofing geolocation. Consider personal location privacy. Be aware of locations where data is populated. Passive locations can still contribute geolocation information. To protect the location of home or work locations, ensure that devices are off and in a faraday enclosure for transport and storage.

If required, identify a method to provide anonymized or unattributed Internet access.

Consider the need for (non-IT related) resources required for the creation of online accounts such as identity documentation like a driver’s license, credit card, or non-VOIP phone number.

The development of the test plan steps and their order of operation are very important. The order and timing of these steps often impacts the creation or availability of artifacts. Step actions should be sufficiently separated in time (e.g., one minute apart) so event timestamps are easily distinguishable in logs and other metadata.

There can be variance in the precision of timestamps of different artifacts. Different artifacts may also be created at specific intervals.

Due to the variance in precision of timestamps and intervals, it is a best practice to execute all actions 1 minute apart to be able to more easily understand resultant data and which artifacts correlate to specific actions. Performing all actions one minute apart simplifies documentation and leads to clearer understanding of the resultant data.

Assess the advantages and disadvantages of automated (i.e., scripted) vs. manual data (content) creation considering accuracy, speed, and repeatability, vs. time to create the script and ability to customize if changes are needed.

Address in advance the ability to access and acquire the created data. When data is created in an application, service, or hardware, consider how the data will be extracted as some devices are harder to acquire data from than others. This will vary with the dataset and whether the created data is stored on a local device or remotely and whether raw access is available via API or a provider’s portal (e.g., Google takeout).

Many service providers offer a portal (e.g., Google takeout), but be aware that there may be delays. Assess the differences in provider-delivered account contents vs. acquiring contents via self-archive utilities, application programming interfaces, or other extraction utilities. (See *Dataset Acquisition*.)

Setup/Preparation

Prior to beginning the creation of the dataset, the developer should review the plan and prepare the subject devices for seeding. This preparation may include:

- Sterilizing media by wiping, formatting, or device resetting as appropriate
- Installing or configuring operating systems or utilities to allow access (e.g., flashing ROMs to mobile devices, rooting "jailbreaking," installing monitoring utilities)
    - Note: some mobile applications will refuse to run when they see su (superuser) is present, or, on Android, the bootloader is unlocked
- Disabling operating system and application updates to ensure the test environment doesn't change during dataset development and acquisition.
- Setting up the lab environment (hardware or virtual) for network and internet access as needed
- Performing test extractions from subject devices to confirm device support and artifact availability
- Performing baseline extractions to support change analysis and facilitate exclusion of known artifacts during dataset review

Dataset Creation

Account Profile Creation

As used in this document, the term "Profile" means a digital persona, real or fictitious, representing a single identity and including all associated attributes. Some use cases may require the use of real persons' identities. This presents additional concerns that should be well understood by the developer prior to dataset creation.

Creating a profile requires documentation of several key items. It is critical to document all accounts and credentials and record when they are created. Access credentials including passwords and multifactor keys should be stored securely.

A sample User Account Information section is available in Appendix A: Dataset Development Documentation Template.

There are websites for creating fictitious data. Several are listed in the table below. These sources are provided for information; this is not an endorsement that the services are appropriate or that all the content is fictitious.

Website	Purpose	URL
Fake Name Generator (a)	Names, passwords, and biographical data	http://www.fakenamegenerator.com
Fake Name Generator (b)	Names, passwords, and biographical data	https://name-fake.com/
This Person Does Not Exist - Random Face Generator	Profile Pictures	https://thispersondoesnotexist.com/

There are several tips for successful profiles

- Fictitious profiles should only talk to other fictitious profiles. When communicating with other fictitious profiles make sure they have the same standard
- Do not connect with people known in real life
- Do not use personal Subscriber Identity Modules (SIM), phone number, etc.
- Try not to create evidence outside of use case locations - store a device in a radio frequency (RF) shielded enclosure when going to and from non-case locations. Before heading home from the data population location, place the device in Airplane Mode and turn-off.
- Use a prepaid credit card for app purchases and do not use personal card information
- Set up two-factor authentication (2FA) for all accounts to prevent public hacks/takeovers after dataset public release.
- Create an email with an account that allows for anonymity and doesn't require a phone number or 2FA to start (i.e., *Protonmail*). Then use that email address as the second factor for other accounts.

There may be resource limitations to creating data. Some applications may require vetting that may involve additional resources to obtain including credit card information, driver's license, or an invitation from an existing user of the application or service. Certain applications may require cell service on a device or a phone number. In other words, it might be imperative to create accounts before generation of the dataset itself. Many applications and services might require 2FA, or other verification methods that might require pre-planning and logistics.

It is important to ensure that profiles do not expose the Personal Identifiable Information (PII) of real people. Profiles should only talk to other profiles and not to real people.

Data Population

After properly researching and planning the creation of the dataset and taking the appropriate preparatory steps detailed above, dataset creation can begin by following the developed plan while contemporaneously documenting all steps and actions. Use another device to capture photos of the data population activity. Depending on the use case, screenshots documenting data population actions may also be useful. Deviations from the plan and explanations for doing so should be thoroughly documented.

A sample Test Activities section is available in Appendix A: Dataset Development Documentation Template.

Profile Enrichment

In some use cases, population of profile data with additional content including photos, videos, and other media may add value to the dataset. It is often useful to use content created by others rather than developer-created content. Developers should be cautious when using such data to avoid violating the content creators' intellectual property rights. Appropriate options include obtaining media from sites that categorize content by license type, e.g., Creative Commons licenses or other acceptable use provisions, and purchasing or licensing commercially available stock photos and video.

Dataset Acquisition

Once the creation of the dataset is complete, the data must be acquired. Different types of acquisitions are supported by a wide variety of tools and methods, and the type of acquisition obtained can significantly impact the content of the recovered data. For example, a full file system acquisition may include third-party data that is absent from a logical acquisition.

Research may be needed to ensure that the target data will be included in the planned acquisition type, and dataset developers should strive to obtain an acquisition sufficiently comprehensive to ensure the dataset can be used as intended. Continuing with the previous example, if the intended purpose of the dataset is to test a tool's stated ability to parse messages from a privacy-oriented, communications application, the developer should ensure their chosen acquisition tool supports a full file system of the test device. Sometimes encryption keys required to parse content are in alternative data locations.

Data within the dataset may be volatile and multiple acquisitions or acquisition types may be needed to obtain the needed information. For example, a database may have a daily cleanup routine, and time to acquisition will determine the amount and integrity of the data artifacts recovered. A secondary acquisition may be valuable in determining the type of volatile records and the period of time of availability for recovery.

There are other reasons for creating multiple acquisitions. This may include comparing changes from the baseline to a later state, or to compare results from different acquisition methods. There are tools that can be used to evaluate changes between different versions. This can be useful in understanding locations of traces created from specific actions.

Furthermore, the tool and/or method used and the type(s) of acquisition(s) must be documented so this information can be reported to other users making use of the dataset, if applicable, since there may be limitations on the utility of the dataset due to the acquisition/extraction type(s).

Be sure to document information related to the extraction. A sample Acquisition Information section is available in Appendix A: Dataset Development Documentation Template.

Pre-Release Activities

Prior to releasing a dataset, a basic quality control check should be performed including a check to ensure that private information – **potentially dataset author information –** is not released. The following items should be considered:

If data or metadata could disclose account information, ensure the following:

- Account passwords have been changed prior to release
- Both the original and current password are documented.
- Financial vehicles (pre-paid cards, etc.) are inaccessible or hold no balance

General quality control includes:

- Review of data for inadvertent PII inclusion
- Review of data for inadvertent questionable material
- Regression testing if this is a dynamic dataset (reference test plans documented in *Planning*)
- Look for inadvertent inclusion of information about the creator of the dataset or participants, such as locations of people's homes or offices.

Despite following best practices, it is possible that PII or other material may be released. Contact information should be provided with the dataset. This enables a user to notify the dataset creator(s) of the presence of PII or other questionable content. The dataset may then need to be modified to remove or replace the information in the raw image and re-release the dataset.

Use of Existing Datasets

When considering the use of existing datasets, the following overall issues should first be addressed:

- Does the dataset meet the needs of the use case?
- How can confidence in someone else's dataset be established?
- Is the set created data, or does it contain real data which has all sensitive data anonymized/redacted?
- Must permission be obtained or citation given?
- Are there other precluding factors?

Meeting the Needs of the Use Case

To evaluate whether the dataset meets the needs of the use case, there are several factors to assess. To begin, the design of the dataset should be understandable, with sufficient documentation for the purpose of the use case. The documentation should enable the user to determine the dataset content, and the pertinence to the intended use. The *Documentation* section of this guide may be used to review other datasets. Any limitations need to be identified, to understand the boundaries of the usefulness of the dataset.

The application of the dataset may have different evaluation criteria depending on whether the dataset itself is a source, or it is used for restoring data to a test device. Verification and compatibility considerations may need to be addressed. If the dataset is to be a resource for training exercises and tool testing, it should document the known content for the pertinent data. If the dataset is to be a resource for tool development, it should contain the necessary test scenarios for the use case.

Be aware that an existing dataset may not be suitable for accomplishing the goal of the use case. This may depend on whether the dataset was created with the intent of public or private use.

Establish Confidence Level for an Existing Dataset

It is the responsibility of the consumer to establish a level of confidence in someone else's dataset which suits the use case. Contributing to the evaluation are the quality of the documentation, assessments by peers, the reputation of the producer, testing of the dataset, and the robustness of the dataset.

The quality of the documentation may be judged using several criteria. There may be a basic level of recorded actions, or there may be detailed logs and records of actions, settings, and attributes. High quality documentation aids the user in understanding system level details. Detailed documentation is not needed for all areas for all uses. Detailed *documentation* may include important information such as:

- Settings; e.g., device settings; namely, Wi-Fi on/off, cellular on/off, and location settings may affect granularity of location based data
- Environment; e.g., hardware differences in Apple devices result in a different format for unique identifiers of the extraction
- Application level permissions
- Key data elements needed for the intended use of the dataset
- Notifications and popup messages; these can be hard to document
- Key Actions; e.g., logon/logoff, device attachment, service activation
- Timestamps
    - Potential clustering issues - too clustered if timestamp granularity is potentially too close
- Geolocation referencing, granularity

Peer assessments may be used to determine who else has used this dataset. Any papers, blogs, or other reviews could bring insight into the quality and applicability of the dataset.

The reputation of the producer may influence the level of confidence the consumer has in the dataset. Personal reputation or organizational reputation may also be taken into account.

Timeliness of the dataset may aid in determining confidence; if dated, recent options may be preferable, or datasets from a timeframe suitable to the use case may be preferred.

Testing can add to the confidence level, if a subset of expected results is obtained when the dataset is examined. It may be possible for some of the attributes of the dataset to be verified in order to give confidence in the rest. That said, independent validation and verification of both the dataset as a whole and the artifacts contained within might be necessary throughout the lifecycle of the dataset.

For some situations, multiple scenarios may be needed to understand the behavior of a system. User behavior, system behavior, and unknown factors can interact and impact the overall system behavior. A consumer should watch for causation and correlation. Just because the use case reacts in a fashion once, does not mean it will always work that way. There may be a plethora of combinations of device settings, application settings, OS settings, with user activity, and with service provider settings. The consumer should consider any scenario testing with one dataset versus considering how else can the scenario happen. If the expectation is that the consumer may recreate the exact setup that dataset simulates when testing the meaning of an artifact and its meaning. For fast changing tech (e.g., social media applications), this is particularly important. The tester should be on guard against cognitive bias.

The consumer should not use a general dataset to address a specific situation that it wasn't designed to address, unless the dataset's documentation is detailed enough to determine the specific situation was thoroughly addressed.

Created Data vs. Real Data

There are times when it may be that real data may be preferable. If real data is used, there are additional considerations:

- Potential for contraband to be present in the data (i.e., Child Sexual Abuse Material [CSAM)
- Privacy issues with PII and potential General Data Protection Regulation (GDPR) violations
- Useful for volume limit testing. Get to thresholds where things would break (e.g., 5 years of data)
- Uncover variables not considered in a manufactured dataset
- May help remove bias of test data author

Sometimes the only available data sources for testing purposes contain real data, for example warrant returns. If creating a public dataset to synthesize data from an evidentiary source, it may be possible to create synthetic data. Creating synthetic datasets requires that you already know what to expect, but allows you to utilize a dataset without PII or other information from real cases.

Permission and Citation

Datasets may be available under free or open licenses, which allows unlimited reuse. The consumer should perform due diligence to ensure the dataset license type is known. If use case results are to be shared, the consumer should determine if the initial creator desires dataset citation when it is used, especially in formal publications. Publications based on the use case should use Digital Object Identifiers (DOI) for datasets.

It is expected that the community will be good stewards of datasets that they are utilizing. It is expected that dataset users will not attempt to collect any associated cloud-based data by utilizing account tokens or credentials recovered from the dataset. Instead, contact the owner of the dataset to request permission to access cloud-based data. Additionally, to prevent the spillage of PII, do not make social media connections to real datasets. See *Maintenance of datasets* and *Community Involvement*.

Other Precluding Factors

Existing datasets may use a version of an application, OS, or other environment which is inapplicable in the use case. There may be unexpected settings such as code page or character set support in the dataset. The spatial or temporal bounds of a dataset may not overlap with the needs of the use case. The consumer may need to weigh the cost of using an existing dataset having low confidence for the use case against the cost of creating a new, bespoke dataset.

Maintenance of Datasets

There are multiple reasons to maintain a dataset. Future usage of a dataset may be limited by the particular circumstances of the use case. When maintaining a dataset, it may be possible to then update the version of an Operating System or application to see how the update affects information from the previous operating system or application version, as well as how/where new artifacts are stored.

It is also important to maintain accessibility to old datasets; this includes test data, and accompanying operating systems, and applications no longer in common use. Older datasets are necessary not only for testing, but to support analysis and findings that may be the subject of testimony. For example, if an older version of an application or operating system is no longer available, it is quite possible that the version may exist in a dataset.

When planning for and preparing datasets, it is important to create a community of user accounts for “buildable” dataset development. As service providers attempt to remove suspected bot accounts, it becomes more difficult to sustain active profiles/accounts in multiple mediums. For example, *Facebook* may remove an account that is not regularly used or connected to a number of other accounts. To remedy this, having multiple profiles in regular communication with each other can prevent these accounts from being removed due to lack of activity or an irregular presence.

Another reason to maintain accounts is to be able to rapidly add other applications or artifacts to an already robust profile. A robust profile, used across multiple social media sites, is less likely to be seen as a suspected false account.

It is also important to periodically review the accounts to not only ensure persistence, but also to ensure that the accounts have not been altered inadvertently. Since the dataset may be shared, it is possible that credentials or tokens may allow for access to the account by others. Therefore, it is a best practice to change account credentials after publishing a dataset.

Community Involvement

There is an opportunity for the community to create shared usage datasets. This includes datasets that are standalone for specific purposes, as well as datasets where the profiles interact between multiple dataset projects in order to aid in dataset robustness and maintenance. For example, a dataset created for a CTF challenge may have profiles who have social media connections with a dataset created for public use. In order for this to be accomplished, there would need to be a set of agreed-upon standards. It is imperative that shared datasets uphold the same quality standards to ensure the dataset is not compromised in terms of containing contraband content, potential copyright issues, PII of real individuals, password maintenance, and other requirements as needed. An established ethics agreement between groups is encouraged.

There are several repositories currently in use by the community. This includes datasets hosted by the *National Institute of Standards and Technology* (NIST) as part of the *Computer Forensic Reference DataSet* (CFReDS) and *Digital Corpora*. Both repositories have the ability to host additional datasets, as well as their accompanying metadata and documentation.

There are multiple examples of robustly-populated, community datasets including the donations of Joshua Hickman's Android and iOS datasets to *Digital Corpora*, CTF challenge sets that have been publicly released, as well as university-created datasets such as the "Owl" datasets from *Marshall University*.

Limitations

Even robust and well-documented datasets cannot address all possible variations/variables.

References

Anobah, M., Saleem, S., and Popov, O. (2014). *Testing framework for mobile device forensics tools. Journal of Digital Forensics, Security and Law 9*(2): pp. 221-234.

Bhat, W., AlZahrani, A., and Wani, M. (2020). Can computer forensic tools be trusted in digital investigations? *Science & Justice*.

Brunty, J. (2011). Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner. https://www.researchgate.net/publication/320808735_Validation_of_Forensic_Tools_and_Software_A_Quick_Guide_for_the_Digital_Forensic_Examiner.

Computer Forensic Reference Dataset (n.d.). https://cfreds.nist.gov/.

DCAT-US Schema v1.1 (Project Open Data Metadata Schema) (n.d.). https://resources.data.gov/resources/dcat-us/#standard-metadata-vocabulary .

Digital Corpora. (n.d.). https://digitalcorpora.org/.

Guo, Y., Slay, J. (2010). Data Recovery Function Testing for Digital Forensic Tools. In: Chow, KP., Shenoi, S. (eds) Advances in Digital Forensics VI. DigitalForensics 2010. *IFIP Advances in Information and Communication Technology, vol 337*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15506-2_21.

Horsman, G., & Lyle, J. R. (2021). Dataset construction challenges for digital forensics. Forensic Science International: *Digital Investigation*, 38. https://doi.org/10.1016/j.fsidi.2021.301264.

Horsman, G., (2019). Tool testing and reliability issues in the field of digital forensics. *Digital Investigation, (28)*: pp. 163-175.

Horsman, G. (2018). *"*I couldn't find it your honour, it mustn't be there!" – Tool errors, tool limitations and user error in digital forensics*. Science & Justice, 2018. 58*(6): p. 433-440.

Hyde, J. (2021) Methodology for Testing Forensic Hypothesis and Finding Truth. [Video Blog Post] Retrieved from https://www.eccu.edu/methodology-for-testing-forensic-hypothesis-and-finding-truth/.

Hyde, J. (2022) Creating Synthetic Test Data. [Blog Post] Retrieved from https://www.hexordia.com/blog-1-1/creating-synthetic-test-data.

Inoue, H., F. Adelstein, F., and Joyce, R. (2011). Visualization in testing a volatile memory forensic tool. *Digital Investigation*, (8): p. S42-S51.

Lyle, J., *Testing Disk Imaging Tools*, in *DFRWS*. 2002: Syracuse, NY.

Marshall, A. M. (2021). Digital forensic tool verification: An evaluation of options for establishing trustworthiness. *Forensic Science International: Digital Investigation, 38.* https://doi.org/10.1016/j.fsidi.2021.301181.

Marshall University (2019). Owl Dataset https://cfreds.nist.gov/all/DigitalCorpora/2019OwlScenario

Michel, M.; Pawlaszczyk, D.; Zimmermann, R. (2022). AutoPoDMobile—Semi-Automated Data Population Using Case-like Scenarios for Training and Validation in Mobile Forensics. *Forensic Sci*. 2, 302–320. https://doi.org/10.3390/forensicsci2020023.

Mohamed, A. F. A. L., Marrington, A., Iqbal, F., & Baggili, I. (2014). Testing the forensic soundness of forensic examination environments on bootable media. *Digital Investigation, 11(Supplement 2)*, S22–S29. https://doi.org/10.1016/j.diin.2014.05.015.

Pan, L., & Batten, L. M. (2009). Robust performance testing for digital forensic tools. *Digital Investigation, 6*(1), 71–81. https://doi.org/10.1016/j.diin.2009.02.003.

SWGDE. 2018. "Minimum requirements for testing tools used in digital and multimedia forensics." In Scientific Working Group on Digital Evidence, version 1.0. https://drive.google.com/file/d/1IId_kzWMH6NVf7edLiN8dJU-ocpOSjXh/view

Wilsdon, T., and Slay, J. (2006). Validation of forensic computing software utilizing black box testing techniques. *4th Australian Digital Forensics Conference*. Security Research Institute (SRI), Edith Cowan University.

Wundram, M., Freiling, F., and Moch, C. (2013). Anti-forensics: The next step in digital forensics tool testing, *IEEE Seventh International Conference on IT Security Incident Management and IT Forensics*, p. 83-97.

Xiaoyu Du, Hargreaves, C., Sheppard, J., & Scanlon, M. (2021). TraceGen: User activity emulation for digital forensic test image generation, Forensic Science International: *Digital Investigation, Volume 38, Supplement.* 301133, ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2021.301133.

Appendix A: Dataset Development Documentation Template

This resource is available for download from Dataset Development Documentation Template.docx

Intended usage(s) for dataset:

☐ Training exercise
☐ Capture the Flag exercise/test
☐ Verification of artifact/finding
☐ Other:
☐ Competency/Proficiency/Certification test
☐ Tool testing
☐ Product/Tool development

Development method:

☐ Hardware-Based/Physical Device
Make/Model:
OS/Version:
Device Identifier/Serial Number:
Phone Number (if applicable):
ICCID1:
ICCID2:
Applications tested:
Device date/time settings: Choose an item.

☐ Emulation
Emulation environment:
Version/Configuration:
Applications tested:

Date and time format used (include timezone): *YYYY/MM/DD HH:MM (24-hour UTC) Note that date formats are often confusing since there are many commonly used variations. The key point is to be clear what format is being used.*

User Account Information:

Platform/Service: Click or tap here to enter text.
User Account/Username: Click or tap here to enter text.
Password: Click or tap here to enter text.
Account Creation Date: Click or tap to enter a date.
Account Creation Time: Choose an item.

Repeat section as needed based on number of configured user accounts.

Application Information:
Application: Click or tap here to enter text.
Version: Click or tap here to enter text.
Installation Date: Click or tap to enter a date.
Installation Time: Click or tap here to enter text.
Username: Click or tap here to enter text.
Password: Click or tap here to enter text.
Permissions:

Default:
Requested by app:
Granted by user:
Notes:

Repeat section as needed based on number of installed applications.

Operating System (Device/Emulated System):

Name:
Version:
Installation Date:
Installation Time:
Username:
Passcode/Password:
Administrator/Root Privilege? ☐ Yes ☐ No

Operating System (Host, if using emulation):

Name:
Version:
Installation Date:
Installation Time:
Username:
Passcode/Password:
Administrator/Root Privilege? ☐ Yes ☐ No

Test Activities:

Test Procedure:

Include brief paragraph or list outlining test prep, setup, activities, and extraction info.

Detailed log: *Add table rows as needed*

Date	Time	Application	Action	Content/Details/Location

Acquisition Information: (if applicable)

Acquisition Tool: *Include tool name and version number*

Acquisition Date: Click or tap to enter a date.

Acquisition Time:

Acquisition Method/Type: Choose an item.

Acquisition File Name(s):

Acquisition Notes:

Hash Value(s): *List hash type and value for each extraction file*

Filename: Click or tap here to enter text.

Hash Type: Choose an item. Hash Value: Click or tap here to enter text.

Known Limitations: *Include brief paragraph or list outlining limitations re: dataset creation or usage. Might want to include some typical examples – like variability among app/OS versions, underlying hardware, etc. How was the test data populated – manual vs automation? Hardware vs emulated caveats, etc.*

Appendix B: Example Use Cases

This section contains several sample use cases categorized based on the designations in the Use cases for datasets section of the document.

Forensic Tool

- Product/Tool development - When building and developing digital forensics tools, either Open Source or Closed Source, test sets are often used to both determine data storage and validate the functionality of tools. During product development new datasets are often generated as part of the initial development process as part of the research of new artifacts. Tools sets can test the function of a specific artifact or the functionality of the tool to ensure that data is not altered by the tool. These datasets can be part of the Software Development Life Cycle (SDLC) to ensure that tool functionality and capabilities are maintained as new features and versions are released. Product development may require samples (knowns) to develop tools that find, parse, or otherwise act upon the data. In some cases, the data may be a system or module that has exhibited some behavior or activity rather than contains a known artifact. Forensic datasets are used throughout the SDLC: initial design, during product development tests, final product testing and later regression testing. Every time an operating system or application updates, the dataset may need to be updated. This may include new features and forward and backward compatibility. New and legacy artifacts can have different characteristics. Some tools report version compatibility tested by the developer and allow for user feedback (community). It's good to be able to tell users what specific versions have been tested. Example that comes to mind: Facebook changed database files/structure, need to add support for new versions while retaining support for old versions. When testing tools it is important to remember that a difference between results of different versions of the same tool may be representative of a change in support rather than a tool failure.

Forensic Practitioner

- Competency/Proficiency/Certification tests - Intended to test examiner's knowledge more so than the functionality/reliability of the tool(s) used. Well-documented test dataset development allows for test criteria/expected responses to be established prior to test administration so that they can ensure the dataset includes artifacts that will not be automatically parsed by commercial forensic tools (ensure the trainee has the opportunity to develop skills in manually parsing data).
- Learning/knowledge competitions / Capture-the-Flag challenges - Intended to test the examiner's knowledge beyond how to use a given set of tools. Problems are often developed to force the user to show a deeper set of skills.

Software and Hardware Environment

- Verification of artifacts/interpretations - Artifacts are developed but the interpretation may change over time. There is a constant need to find and understand new artifacts and to ensure that old artifacts retain their meaning.
- Peer-review of artifact research - During academic review of parsed artifacts, author or reviewer generated datasets are an essential part of a detailed review of the stated data storage.
- Case specific requests - A variety of requests can be developed relating to a specific case that require testing. These requests for information (RFIs) can come from a variety of sources.

- Application deconstruction and reverse engineering - This process can be used to find deeper understanding as to why a particular trace is created. Dataset should be as similar as possible to source data from which the artifact in question was recovered
- Testify - Occasionally, there is a need to verify information to support testimony or a finding in a report. This is needed when there is an important finding based on an artifact whose meaning is not well established. In these situations, there is a need for a dataset with known content similar to the evidence material. This can be used to show that the technique used can reach the known correct answer.