Forensic Science International: Digital Investigation 52 (2025) 301867

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

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

DFRWS EU 2025 - Selected Papers from the 12th Annual Digital Forensics Research Conference Europe

Preserving meaning of evidence from evolving systems

Hannes Spichiger [a,*], Frank Adelstein [b]

[a] Hochschule Luzern Informatik, HSLU I, Suurstoffi 1, Rotkreuz, CH-6343, Switzerland
[b] Hexordia, PO Box 155, Bridgeport, NY, 13030, USA

ARTICLE INFO

Keywords:
Data preservation
Reference data
Data uncertainty
Evolving systems
Distributed systems

ABSTRACT

Preservation is generally considered as the step in the forensic process that stops evidence from decaying. In this paper, we argue that the traditional scope of preservation in digital forensic science, focused on the trace, is not sufficient to ensure the stop of decay in the context of evolving systems. Instead, insufficiently preserved reference material may lead to the loss of meaning, resulting in an overall increase of uncertainty in the presented evidence. An expanded definition of Preservation and a definition of Reference Data are proposed. We present suggestions for future avenues of research of ways to preserve reference data in order to avoid a loss of meaning of the trace data.

1. Introduction

Forensic experts have a privileged role within the justice system. They are asked to give meaning to technical and scientific evidence presented in front of court, with the aim to elucidate events relevant for the case at hand. In particular in relation to digital evidence, this role is essential to help decision makers understand the significance of presented data. This step is however not without its challenges. The experts aim to give meaning to those traces as close to what actually happened as possible (Pollitt et al., 2018; Roux et al., 2022). To do so, they rely on reference data, defined here as follows:

Definition. **Reference Data** is data used as a standard to classify or interpret trace data acquired from digital devices.

An equivalent concept exists in most other domains of Forensic Science where it is called reference, standard, or print. Reference data is created through experiments conducted in comparable conditions, either by the experts themselves or published by other practitioners and researchers. In case of missing or outdated data, this can lead to erroneous conclusions. The meaning that is assigned to the data might not be representative of the past events.

In particular with evolving systems, any system whose functionality changes over time, assigning meaning can pose a major challenge. If the reference data was not created on the same version as the traces, the original meaning may not be understood by the examiner. The more a system changes, the more parts it has that may influence its behaviour, the more susceptible to change its functionality is, the more traces are affected by this issue. Given the current trends in technology (cloud services, continuous improvement of applications, distributed systems), obtaining relevant reference data may become a core challenge regarding digital evidence in the years to come.

Little prior work exists addressing this issue: In the context of cell site analysis, it is general protocol to conduct measurements as close to the event as possible in order to reduce the risk of the network changing in between (Hoy, 2015). An extensive discussion of potential changes in cell site measurement data can be found in (Tart, 2022). Some discussion exists regarding changes in gun and shoe soles, impacting the possibilities of firearm (Bonfanti and De Kinder, 1999) and shoe mark analysis (Davis and Keeley, 2000; Stauffer, 2000). Effects of weathering are known for paint (Jost et al., 2016; van der Pal et al., 2016). A recent analysis of a group of active paints showed that their chemical makeup changes over time, requiring timely gathering of reference material (Pintilie et al., 2024). However, to our awareness, this aspect has not been discussed in a more generalised manner.

Research exists in the area of ground truth data, but most of it focuses on the artificial data generation and its related challenges, and how to collect, organise, and archive artefacts from investigators for real systems (Abt and Baier, 2014; Breitinger and Jotterand, 2023; Garfinkel et al., 2009; Grajeda et al., 2017; Horsman, 2024). Some papers (Breitinger and Jotterand, 2023; Horsman, 2024) mention that the software version should be noted but do not discuss highly dynamic, opaque server side code. Little prior work exists addressing the issue of the implications and potential impact of distributed systems with rapidly evolving server-side code, in which the changing functionality of

* Corresponding author.
E-mail address: hannes.spichiger@hslu.ch (H. Spichiger).

https://doi.org/10.1016/j.fsidi.2025.301867

Available online 24 March 2025

the back-end system could make the meaning attributed to reference or test data obsolete or even misleading.

In this article, we aim to highlight the importance of reference data in DFS practice. A focus is placed on how changing systems require timely intervention, as it may become impossible to recreate a suitable reference environment later. This article does not aim to provide practical guidance, as there are currently no solutions and recommendations are likely to vary for different types of traces. However, we will suggest avenues where such solutions may be found.

The remainder of the paper is structured as follows: After the introduction in Section 1, the process of evidence preservation is presented in Section 2. Section 3 describes the impacts of evolving systems. Considerations regarding the gathering of reference material in digital forensic science are presented in Section 4. Section 5 presents paths forward before a conclusion is reached in Section 6.

2. Evidence preservation

Preservation is a core part of the Forensic process and often specified as an independent stage in models describing this process. For example, the OSAC Framework for Harmonizing Forensic Science Practices and Digital/Multimedia Evidence defines preservation as *protect traces from alteration (e.g., isolating them from surrounding environment), collect traces in a manner that changes as little as possible, and evidence management activities such as storing evidential items* (Pollitt et al., 2018).

The core aim of Preservation, as very explicitly stated in the above definition, is to maintain the state of the traces at the time of their discovery as well as possible. This is a primary requirement for Forensic work, as alterations may impact the later conclusion. As such, this aspect is also at the centre of many other definitions of preservation (cf. Adams, 2012; Casey, 2011 p.187 ff., p.245 ff.).

However, there is another essential aspect of this phase: It also aims to ensure the availability of the traces for later in the analysis. Measures taken during this phase should aim at optimising the chances of success during later stages, as approaches that guarantee even perfect conservation are completely useless if access to the traces is no longer possible afterwards. Definitions of preservation do generally imply this aspect. For example, the evidence management aspects of the OSAC definition above clearly have the later availability as an aim although it is not explicitly stated.

For digital objects, questions of preservation often arise in relation to the current state of the preserved device. Particularly with current Smartphones, it is generally considered essential to keep the devices running, as storage encryption stops successful extraction without password in many cases, thus reducing the later availability of the traces. On running PCs, gathering a RAM-dump may be the only way to have access to some of the information on the device (Hausknecht et al., 2015). Other devices may not have permanent storage altogether, essentially destroying the data when shut down. In most cases, devices are also isolated from the network and manipulation of the device is kept to a minimum to stop external influences from altering the data on the device (Ayers et al., 2014). During preservation, the overall situation has to be considered, potentially overriding general rules. For example, if wiping of a PC is ongoing, it is often a better call to shut the device down to stop the wiping (e.g., by removing the power source), than to attempt a RAM recovery beforehand.

Often considered as the ultimate way to ensure preservation is data extraction. Once data is no longer just on the evidentiary device, but a copy is on a secured system, the immutability of this data can be guaranteed through more managerial means, such as calculating and documenting hashes, ensuring traceability through access control and documenting measures taken. Once the data is copied, time is generally considered to be no longer an issue, at least for practical considerations (Casey, 2011 Ch. 7.6).

3. Impact of evolving systems on evidence

We use the term *evolving systems* to describe highly dynamic distributed systems where components of both the server and client-side change often, which can impact the stability of reference data. In this section, we discuss how modern distributed systems impact trace and reference data, how this data can change, or its meaning can "decay" over time, and how this introduces uncertainty into both trace data and reference data.

3.1. The modern distributed system, trace data, and reference data

For over a decade, we have had ubiquitous networking, where end users have relatively powerful computers (laptops, desktops, and smartphones) connected via an ever-present high-speed network (cellular, WiFi, or wired) to vast arrays of servers that provide large-scale storage and massive computation power. The growth of cloud computing allows small companies the ability to rent scalable fully deployed servers without having to deal with the costs of setting up, running, protecting, and maintaining their own servers. This combination of powerful end devices for users (Smartphones) and an affordable way to build up computing power on the back end has led to the modern ecosystem of applications that users run on their phones. The client/server model has been around for several decades, but these recent innovations have had a subtle impact on some of the fundamental tenets of digital forensics, specifically evidence preservation and reference data, as defined in the introduction.

For digital forensics, reference data that provides meaning to its corresponding trace data can be gathered at the same time as the trace data. It is possible to get it at a later date through testing a similar system, but the investigator needs to ensure that the test system has no significant differences that might impact the results. We discuss testing in Section 3.3.

An implicit assumption is the reference data, just like the trace data, does not change. Reference data is assumed to not change or if it does, it changes in a slow, predictable or well documented way. In the modern digital ecosystem, this is not always the case.

3.2. Evolving Systems–Change over time: Same evidence, different meaning

On a conceptual level, digital evidence can change in the same ways as its physical counterparts, for example by decay, adding material, or altering it. Investigators use cryptographic hashes to show that the data has not changed since acquisition.

An underlying assumption of this approach is that if preserved evidence has not changed, then its meaning must not have changed either. But that is not always true in an *evolving system*, a highly distributed environment where the remote server runs programs that follow a rapid prototyping/continuous delivery software development paradigm, relying on third party libraries to provide functions that call functions in other libraries that call functions in other libraries, and so on. With continuous delivery, developers can update publicly facing software libraries rapidly– code on a production server could change daily. Additionally, due to geography-specific features, phased rollouts and A/B-testing, two users using the same app at the same time, may observe different behaviour. Because of all of the layers of dependencies, even though any given change might be small, over time significant functionality can change.

Obviously, this would not change the data the forensic investigator preserved and kept in an isolated system, but the way the server uses and interprets that data *can* change. This could be when the underlying API calls change, or it could be caused by subtle changes in deep layers of the software libraries the server users.

How the backend server interprets data might change, and the code for it might only be seen on the server side. For example, the server could

provide the front-end client with a list of IDs of other users. Those could be users with whom they have played a shared game, or simply a list of users on the server who are available to play a game. The protocol for requesting and retrieving that data might not have changed, but the way that data was interpreted and used could have changed and there would be no indication.

3.3. Testing and presumed meaning

With a continuous integration approach, the version of the backend will regularly change, and when that happens, libraries it uses might change versions too. For small systems, like a mobile game published by a small company, there might be no history published publicly on the changes of each version of the server code, which might be split across many programs following the microservice paradigm. The front-end code on the mobile device likely will not have the version of the server code (or all of the services it provides), especially if the API itself has not changed. Given that backlogs for forensic analysis can be months or even years, it may be virtually impossible to determine what, if anything, has changed on the server side since the data was acquired. And more importantly, there is no way to determine if the server is now interpreting data differently. In the example above, the client stores a list of IDs, but has no additional context information, other than perhaps "UserID" as the label for the list. Without visibility into the server, a forensic examiner would not be able to determine what that list represents, i.e., people with whom the user had communicated or merely people who happened to be logged onto the same server at the same time.

Another foundational element of digital forensic investigations is that testing can be used to validate the interpretation of data (NIST, 2017). A laptop computer with Windows that has been in secure storage for a year will have the same file system on it as it did when it was seized, and the meaning of the timestamps on its file system will not have changed, even if Microsoft has released an OS update since then.

But in the distributed systems approach, there is no practical way to seize the server and preserve it for a year. Any validation tests that an examiner does will be using the server's interpretation of the data with respect to when the validations tests are run, which might not be the same as when the data was acquired. Determining if anything significant has changed on the server in the interim might not be possible. If the application is not widely popular, it is likely that little documentation exists for it or its change history.

Another example of how the meaning of data can change over time is the Windows Registry. Registry keys are created such that their value controls some behaviour of the system. Over time new keys are added to the registry that will override the meaning of the old keys. For example, if the MDMWinsOverGP key is present and set to 1, in cases where equivalent MDM and GP policies exist, the MDM policy will be used and the GP policy will be blocked (Microsoft, 2024). Looking at the value of one key is not enough. The investigator must validate that changing the key's value affects the system as expected. And clearly the test system used to establish the reference data must run a version of the OS as close to the OS version on the system containing the evidence. Testing a server's current behaviour does not guarantee that that was the behaviour a year ago when the device was seized. So even though the data has been properly preserved and can be shown not to have changed, and the investigator performs tests on how to interpret the meaning, the presumed meaning can be invalid with respect to the data on the device from a year ago.

Investigators must validate that their tools work, and the more critical a piece of evidence is, the greater the importance of testing that the tool operates properly on that specific type of data. Generally, labs and investigators test their tools to make sure that they work, but a typical forensic tool processes an enormous amount of data and produces a large number of results based on that data.

It is difficult to test that every reference data or function of a tool is accurate, and investigators have limited time for such activities. And in fact, the tool might be accurate, but the OS might have inaccurately labelled some data.

Before we can address the problems of reference data, we must first consider the uncertainty associated with evidence.

3.4. The illusion of determinism and sources of uncertainty

Forensic investigators seek answers and try to determine what happened in a given situation. The reality is that there are no absolutes, uncertainties exist in every part of a case, and "proving" a case means that a consensus exists that the probability is high that the evidence can be interpreted in a certain way and the doubt (uncertainties) are low (Cf. Casey, 2002). In US law, the standard of proof might be "a preponderance of evidence" or "beyond a reasonable doubt". The court system was designed to deal with uncertainties. Traditional (non-digital) forensic analysis includes potential uncertainty for scientific techniques. The mindset that "computers are deterministic and binary, so we must be able to come up with a simple, ground truth yes or no answer" can mislead digital forensics investigators. But given the complexity of all of the layers and incomplete information, an absolute yes or no is generally not feasible.

Operating under the illusion of determinism pushes the investigator towards a black and white world, where likely evidence and scenarios are considered to be absolute truth, and ones that are less likely are considered to be flawed and incorrect, and thus abandoned. This view sets the investigators up to fall prey to various cognitive biases in their investigations, such as selection bias, and seek ways to justify their conclusions, rather than using the data to guide them to their conclusions, whether it is to their liking or not (Sunde and Dror, 2019).

We need to move beyond the deterministic view and seek to understand and mitigate the sources of uncertainty where possible.

Similar to Locard's Exchange Principle (Locard, 1920), no event of importance happens in a vacuum. There will be multiple sources of evidence present in multiple modalities. This can be digital evidence, traditional evidence, witnesses, and more. And similarly, within a single modern computer or connected device, at any time many events are happening concurrently, leaving traces in many sources including local and remote log files.

Relatively straightforward cases may have a low uncertainty. For example, an investigator might find thousands of CSAM files on a suspect's computer with no evidence of any malicious programs on the computer or in its memory. Additionally, there might be shoeboxes of DVDs with labels in the suspect's handwriting categorising the material including download dates.

Other cases may have several pieces of linked evidence that have significant uncertainty. For example, a home computer could be shared by several household members, have multiple user profiles, yet one profile is shared by multiple members, and there is evidence that one or more viruses existed on the computer. Without additional evidence, data from that computer would have a significant uncertainty associated with it since it could be difficult to determine who used the computer at any time, including actions caused by the potential virus.

During an investigation, evaluating multiple, competing hypotheses can help reduce uncertainty by eliminating alternate explanations or showing that they are highly unlikely, which can reduce the uncertainty of the remaining hypothesis. This approach also helps mitigate potential cognitive biases (Sunde and Dror, 2019 p. 106). For this, the gathering of relevant reference material is essential.

4. Reference material in digital forensic science

Whilst it is never possible to eliminate uncertainty from the process, we can take measures to reduce it. In this section, some possibilities in relation to issues caused by distributed systems are discussed. These possibilities do in no way aim for completeness.

4.1. Gathering of reference material

As discussed in the previous section, the significance of data cannot simply be derived from evidentiary data. In unknown applications, the values of fields and flags need to be understood through testing. Suppose an unknown app had a database containing a column entitled “read” and containing a binary flag. That by itself does not tell whether the field actually indicates the read-status of the table’s message, what causes the value in the field to change and what state corresponds to what status. Such reference material is relatively easy to gather: On a test device, in an environment as close to the one in which the trace was generated, data is generated with the application of interest. What factors need to be identical for the environment to be considered sufficiently close will depend on the studied traces and will certainly be a topic of research to come. The undertaken interactions with the device are documented, and the data is analysed to understand what actions are possible to cause data structured in the way of the evidentiary data. Such reference data creates an understanding on how the studied application creates and stores data, reducing uncertainty resulting from missing understanding of the studied data. This data also has the advantage that it can be published, allowing other practitioners faced with the same application to reuse this data.

Other data may be case specific and cannot be reused between cases. It has now been well documented that location-data and its accuracy is location-dependent (Ryser et al., 2024; Yoo et al., 2020). It is therefore not possible to infer about the accuracy of location-related traces without conducting measurements on site. The process to conduct such surveys is described in (Spichiger, 2023). Other examples of case specific reference data include PRNU (Geradts et al., 2001) or health data (Van Zandwijk, 2022; Van Zandwijk and Boztas, 2019, 2023) These measurements will not allow eliminating uncertainty on the conclusions; however, they allow understanding the scope of the uncertainty.

With the exception of applications documented by researchers, as far as the authors are aware, there is currently no widespread practice among digital forensic practitioners to systematically gather reference data. According to AppBrain, over 1.5 Million Apps are currently available on Google Play (AppBrain, 2024), only a fraction of which have ever been documented from a Forensic perspective. An example of this can be seen in Commonwealth vs Arrington (SJC, 2024), where admissibility of iPhone Frequent Location was discussed in front of the Supreme Court for the County of Suffolk (Massachusetts, US). In this case, the state’s expert witness conducted experiments regarding the reliability of the questioned traces only in preparation for the trial, despite the facts taking place in 2015. Ultimately, the court ruled the Frequent Location evidence as inadmissible, among other things due to insufficient testing.

Beyond admissibility reasons, this oversight is fundamentally dangerous: Without any reference data, the value of a trace cannot be assessed and attempting to do so is akin to guesswork. It is somewhat surprising that this is not called out more frequently, as no one would ever allow a specialist in dactyloscopy to express an opinion on whether a suspect is at the source of a finger mark without looking at the prints of that person. Just because someone has a lot of experience with studying mobile applications does not mean that they know how the application of interest in the case at hand works.

The way the issues we discuss could manifest is in limiting what conclusions could be drawn from certain evidence or changing what once might have been primary evidence into supporting evidence.

Examiners that follow standard practices would not make assertions based on evidence that has a high uncertainty. As stated in a NIST report on digital investigation techniques, "Digital investigation techniques are based on established computer science methods and when used appropriately are considered reliable" (Lyle et al., 2022).

The National Software Reference Library (NIST, 2016) is a collection of sets of hashes of software. The software includes files distributed with standard operating system releases as well as well-known malicious programs. The hashes are used to quickly find or eliminate files from consideration during or after acquisition. Whilst helpful, it cannot be complete, as new software (both malicious and benign) is continuously released. Similarly, digital forensic software that interprets certain file formats (e.g., file systems, browser caches, email, etc.), will suffer the same limitations. Therefore, investigators must be able to create their own reference information, and to do that they must create known test data either using or as input for the new software. NIST’s OSAC Digital Evidence Subcommittee Task Group on Dataset Development created Guidelines for Dataset Development provides guidance on how to create data sets to support testing in a reliable, repeatablOSAC, 2022).

It is the authors experience that practitioners will sometimes leverage engineering experience for an argument about how good naming and programming practices ensure the proper workings of code and the sensible labelling of values. We cannot assume this to be the case. In particular with lesser-known applications, the code might not be written by the most skilled programmers and even those make errors. The more obscure the observed functionality is, the higher the risk of it not quite behaving the way we expect it to, and therefore also the risk of misinterpretation of the trace.

Whilst this problem is not new, the increase in the rate of change of analysed systems can only make this problem worse. Digital forensic practices in a law enforcement context used to be heavily focused on data recovery. However, cases where the intricacies of digital systems are of relevance have become more and more frequent. With rising digital literacy among lawyers, and awareness for potential issues, it is more often required to be able to explain how a system exactly was working where previously methods and techniques would not be challenged. For this, the gathering of relevant reference material is crucial.

4.2. Stability of reference environment

As discussed in Section 3, studied environments change. This also impacts the creation of reference data at a later time, as the system of interest may no longer exist. Whilst it can be challenging to find devices of a specific make and model, and deploy software in the required version, it is generally something that is feasible, thanks to the human tendency to collect things. This is no longer true with distributed systems, as remote services are entirely out of control of the user.

This volatility of the environment has a degrading effect on the assigned meaning of the gathered traces. As stated before, without relevant reference material, it is not possible to assess the significance of a trace. Failing to gather reference material sufficiently early may result in a complete loss of value for these traces.

The aspect of time in preservation is not per se new. It is well known as a factor for the trace, probably best illustrated by (Kind, 1994) as a

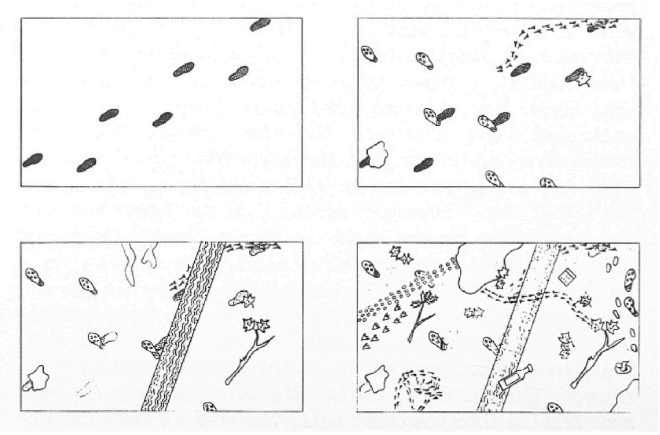

Fig. 1. Reproduction of the visualisation by (Kind, 1994) showing trace degradation. Author authorisation was impossible to obtain.

series of shoe marks in the mud (cf. Fig. 1). With passing time, other people pass by, cars crossover, leaves, twigs and trash are left in the vicinity. These later events cause the scene to be polluted, degrading and ultimately destroying the trace. With digital traces, we see this for example in system logs, which generally have some form of retention limit, either through age or log size, or with erased files that run the risk of being fully deleted by overwriting or garbage collection mechanisms (Joshi and Hubbard, 2016). An extreme example of this are RAM captures, which often contain inconsistencies due to the RAM changing during the capture process, potentially leading to inconsistencies within the capture itself (Ottmann et al., 2023; Rzepka et al., 2024).

For reference material, this aspect is far less discussed, although not new. In the same situation as described by Kind, the securing of the shoe for reference material is equally important as the securing of the trace, as continued use of the shoe will alter its sole, potentially to the point making a meaningful comparison impossible (Girod et al., 2008).

With digital traces, reference material may become hard to obtain through a range of reasons. In particular if the element of interest is produced by a remote service, changes on this service may have an impact on its behaviour. Software updates may change basically any aspect of a system. Registry keys may be reassigned for a different function. Hardware updates may result in changed response times. IP-addresses and URLs may be reassigned, and the content shown by web servers may change. Changes in clock synchronisation systems may cause timestamps to behave differently. All these changes could lead to a different interpretation of found traces and therefore negatively impact their potential evidentiary value.

As this degradation of value can be understood as a failure in preservation, we here suggest an extension of the definition of the preservation by (Pollitt et al., 2018):

Definition. **Preservation**: protect traces from alteration (e.g., isolating them from surrounding environment), collect traces in a manner that changes as little as possible, and evidence management activities such as storing evidential items. **This includes gathering relevant supplemental information about the traces such as metadata, reference data, and context.**

This additional data is essential for the later use of the traces as evidence in front of a court and assuring its existence in the context of preservation allows investigators to reduce uncertainty in their expert's conclusion. Metadata is information about trace data, like timestamps on files and is commonly used in Digital Forensics. Context is information about the investigation, which could include the examiner's ID and the versions of the software they used, or in Physical Forensics, whether a body was moved, or perhaps rolled over and examined by paramedics, before police arrived. In the context of the OSAC framework, the recording of context and metadata is considered to be part of a distinct forensic activity: Documentation (Pollitt et al., 2018). These aspects are integrated here in Preservation as well, as failure to record them may have an impact on the value of the evidence as discussed below. As the focus of this paper is on reference data, context and metadata are out of scope and not further explored here.

Fig. 2 illustrates the problem of trace and reference degradation in a conceptual way. On the X-axis, the degradation of the trace is shown. This corresponds to Kind's visualisation: The further along the axis, the stronger is the degradation of the trace. The Y-axis tracks how strongly the reference object or environment has changed, leading to less valuable reference material. Point A shows the situation at the creation of the trace: both the trace and the reference material are in perfect condition. This ideal state is impossible to attain, due to the delay between the creation of the trace and the intervention on the scene, visualised on the graph by a grey zone of impossibility. With the progression of time, the point for any potential piece of evidence will start moving away from this position. This does not have to be in a linear manner. It may be stable in one dimension for some time. For example, a powered off device will for most intents and purposes not change significantly,

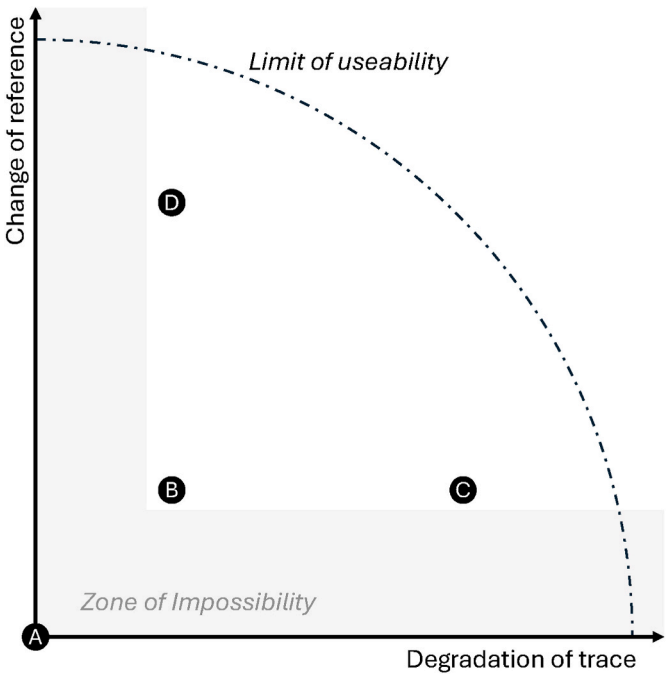

Fig. 2. Conceptualisation of the influence of trace and reference degradation on the potential value of the trace as evidence.

resulting in little to no movement on the X-axis. Similarly, some events may cause a rapid elongation, for example, an online service shutting down, causing rapid movement along the Y-axis as it will no longer be possible to study the behaviour of this service in a controlled setting.

The distance to the origin is representative of the maximal information and certainty on that information that can be gained from any piece of evidence situated at this position on the graph. As is quite evident, this is dependent on both the quality of the reference material and the quality of the trace. Arguably, there will be a point at which a threshold is reached, at which the remaining value of the evidence is so low that using it in court does not provide any meaningful value anymore. This limitation is shown by the dashed line.

In this visualisation, preservation as defined above consists in stopping the point from moving any further on the graph. Point B is close to what is practically possible to obtain: in close proximity to the grey zone, both the trace and the reference material are in a state relatively close to the moment the trace was created. If both the trace and reference material are gathered at this moment, the potential value as evidence will be as high as possible. Points C and D represent situations in which only one of the two aspects was properly preserved. The trace in the case of point C and the reference material in point D. They serve to illustrate again the necessity of having both with good quality, as their distance to the origin is the same despite them representing quite different situations.

As an example, let us consider the clock of a mobile phone. It is standard practice to document the time of the device and note any deviations from the actual time when seized (point B). This comparison of the phone time with a reference clock is a gathering of reference data. It is done to infer whether the clock at the moment of the creation of traces (point A) was correctly set. Whilst this does not prove that this was the case (as the clock may have gotten deregulated or synchronised in between A and B) it is the best we can do, as we had no access to the device at an earlier point in time. Failure to document this at time B may increase the uncertainty about timing, as deregulation may take place between seizure (point B) and clock documentation (point D). In a case handled by one of the authors, a device sitting in evidence (charging and isolated from the network) demonstrated a significant deviation in time. As the device had been documented to be in sync at time of seizure, it

was possible to state that this drift took place after seizure and had therefore no impact on the timing of the traces. This statement would not have been possible without the documentation at time B and would have had an impact on the degree of certainty possible to achieve from the traces recovered from the device.

4.3. Prioritisation

As with the gathering of traces, it is never possible to gather reference material for everything. The sheer number of existing applications and services makes this impossible. Consequently, it is necessary to prioritise and select what reference material to gather. As with any triage process, these decisions rely on risk management considerations, reducing as far as possible the potential to lose relevant evidence. The following is a non-exhausting list of factors likely to impact these decisions.

4.3.1. How well is the service known and documented?

The more frequently used a type of trace is by the forensic community, the more likely it is that its functionality is known and that changes in its behaviour are observed and documented. Care should be taken that the generally accepted behaviour is based on observation and not just assumption.

Additionally, even well-known applications have more obscure features. If a feature is only rarely used, or it was used in a non-conventional way in the case at hand, it may be necessary to gather reference data despite the app's notoriety.

4.3.2. How volatile is the system?

Environments change at different paces and how difficult it is to reconstruct the environment may change and is going to impact whether it is reasonable to gather reference material in advance. For example, a URL pointing to a clear web site is generally going to have some persistence and its state is likely to be documented by secondary services such as web archives. In contrast, dark web addresses change more frequently and are less likely to be documented elsewhere. Gathering reference material to show where such an address was pointing needs therefore be handled with a different urgency.

4.3.3. How likely is this trace going to be used as evidence?

Not all gathered data will find its way in front of a court. Material unlikely to be used in court is equally unlikely to require corresponding reference material. This is of course somewhat self-fulfilling, as the absence of reference material negatively impacts the evidentiary value and therefore lowers the likelihood of it being used as evidence. Practitioners should be aware that they reduce their own possibilities going into trial by making this selection early on.

4.3.4. How much uncertainty is acceptable?

Whether or not the remaining uncertainty is acceptable for a court will heavily depend on the jurisdiction and the judicial culture. Over time, for a specific court system, experience should establish itself, what is acceptable and what not.

4.3.5. Would exclusion of this piece of evidence massively impact the case?

In any criminal trial, prosecutors are going to present the best evidence they have. What, and how much, evidence this is, varies heavily. As a result, the impact of a particular piece of evidence not being admitted also changes from case to case, and different degrees of measures should be taken to ensure the robustness of the used evidence. If it is likely that digital traces are crucial for the case, as much potential reference material should be gathered as reasonably possible.

As will be discussed in the next section, the development of structured processes to assist practitioners with prioritisation should be a focus of future research.

5. Future research and tools needed

We have described how the problem of uncertainty in reference data can reduce the usability of well-preserved data. There are reference data sets, such as the Computer Forensic Reference DataSet (CFReDS) that was created and is maintained by NIST, the National Institute of Science and Technology (https://cfreds.nist.gov/), but it relies on contributions for its content. Whilst it can be very useful, it is impossible to archive snapshots of all software, libraries, and web sites continuously. Investigators must take an active role in either gathering appropriate reference material or reducing uncertainty in other ways. Below, we present several approaches that could help with the problem of getting timely reference data, including methods to prioritise the order in which an investigator's limited time should be spent on getting reference data. All of these will require research, development, and testing to determine the validity and benefit of these approaches.

We envision several research efforts described below that could yield tools and methods to help preserve reference data and manage uncertainty. It is likely that other fields, such as archival science which seeks indefinite, long-term preservation of data (Dietrich and Adelstein, 2015) and knowledge engineering, may have answers to some of the problems that may be transferrable to DFS. As with all research endeavours, projects in the proposed directions should study the feasibility of the approach.

5.1. Create a tool or service that allows investigators to gather server-side information automatically

Such a tool could be used to document the API being used, its version number, and how the input and output values relate to each other, and other information that could be used to determine if anything has changed since the evidence was initially acquired. Or if an IP address or domain name were significant, the tool could watch for changes to the DNS data and document them so investigators could know if the same server was likely to still exist at the time of the analysis of the preserved traces. An investigator would need to submit the relevant information to the program, such as the web API URI, the IP number, or domain name, and possibly credentials. The program would gather the essential information and periodically check for any changes. Code templates could test or analyse the DNS records or the API and how the input and output map to each other.

5.2. Create a public database of known applications and artefacts that could be queried and accept public submissions

This would be a "VirusTotal + Archive.org for Forensics Artefacts", allowing investigators to submit artefacts or APIs to the system where it would be preserved and automatically watched for changes. This would be part of the Preservation process. If, because of backlogs, the analysis is done, say, a year or so later, the database could be queried to see if the reference artefacts of interest have changed, such as the API on a server or configuration files for an application, or essential functionality of a web browser.

This could be used to reduce some sources of uncertainty–specifically if there was any configuration or semantic change that might impact the evidence. If such a change was present, then the examiner would need to assess the risk this uncertainty posed to the validity of the data. If the changes would not impact the interpretation of the data, for example, an email that is displayed in a slightly different font, then no further action would be required. Otherwise, more corroborating or supporting evidence from different sources would be needed.

In addition, this tool could be useful for teaching digital forensics by allowing training material to be valid for longer than when any single component of the exercise changes versions without requiring updating.

Research would be needed to determine what sort of information sources should be stored, how the information would be gathered, the

ontology and taxonomies to use, and how the system could be extensible, since it might include APIs, files, DNS records, and other sources that have very different properties, representations, and methods to acquire them. In addition, this would need to be a non-commercial undertaking that would require a funding source. The Artefact Genome Project AGP (https://agp.newhaven.edu/about/start/) is very similar to this suggestion, although it currently is not sufficiently complete to fulfil the here described need.

5.3. Create a taxonomy to assess risk and urgency

A taxonomy or framework to assess risk and urgency inherent in digital forensic data would allow investigators to determine what data or processes have the highest uncertainty and offer some suggestions that might be able to reduce the uncertainty or mitigate their potential consequences. Such a taxonomy would likely need to be qualitative rather than quantitative, allowing investigators to characterise the uncertainty of data as low, medium, or high and assess the risk posed by the uncertainty of some data or collections of data. An analogy would be the aeroplane flight risk assessment tools presented in the FAA's Risk Assessment Handbook for pilots, that present the pilot with a short survey of yes/no questions about the plane, pilot, and environment, and associate a numeric value for each answer. The sum of the values map to a low, medium, serious, or high risk, and depending on the situation, certain choices or actions can be taken to reduce some of the risks to an acceptable level (Federal Aviation Administration, 2022).

5.4. Create a risk assessment tool to identify uncertainty and suggest ways to mitigate it

Creating a tool that could help identify the levels of uncertainty and risk among the significant evidence, and assess the uncertainty and risk associated with them, would also be useful to help investigators have a consistent way of managing uncertainty.

It could also alert the investigator if the levels of uncertainty suggest that the system might be behaving in an unconventional or unexpected way.

Once such a tool exists, the guidance it provides could be adopted into the standard operating procedures for acquiring and preserving evidence.

We are aware that the feasibility of some approaches suggested here is at the very least uncertain. It is very well possible that information sent by a server is not sufficient to establish the version of server-side code. As stated, projects collecting artefacts do exist, however they need to be hosted, maintained and alimented by someone, something that has previously proven challenging for the long term. The inherent complexity of the problem may render attempts to model risk factors and mitigation strategies challenging. Nevertheless, we do not want to dismiss these possibilities before they have been attempted as even partial solutions may improve the current situation.

6. Conclusion

In this article, we defined reference data for digital forensics, how it is used to analyse trace data, and associate meaning to it. We described potential problems if reference data is created late in the analysis process, especially when associated with evolving systems. Specifically, the longer the time gap between the creation of trace data and reference data, the greater the uncertainty of the associated meaning of the trace data. This led to an extended definition of data preservation that includes reference data.

We expect that evolving systems will increasingly become more dynamic, increasing the potential impact the uncertainty can have on cases. We concluded by suggesting a series of potential tools and research paths that may help reduce or mitigate uncertainty linked to the creation of reference data.

Acknowledgements

The authors would like to thank Elénore Ryser for her insightful and detailed remarks on an earlier draft of this paper and the constructive feedback and suggestions provided by the anonymous reviewers.

References

Abt, S., Baier, H., 2014. Are we missing labels? A study of the availability of ground-truth in network security research. In: 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS). Presented at the 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), pp. 40–55. https://doi.org/10.1109/BADGERS.2014.11.

Adams, R., 2012. The Advanced Data Acquisition Model (Adam): A Process Model for Digital Forensic Practice (Doctorate Thesis). Murdoch University.

AppBrain, 2024. Number of Android Applications on Google Play (Dec 2024) [WWW Document]. AppBrain. URL. https://www.appbrain.com/stats/number-of-android-apps, 12.12.24).

Ayers, R., Brothers, S., Jansen, W., 2014. Guidelines on Mobile Device Forensics (No. 800– 101 Rev. 1), NIST Special Publication (SP). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-101r1.

Bonfanti, M.S., De Kinder, J., 1999. The influence of the use of firearms on their characteristic marks. Associat. Firearms Tool Mark Examin. (AFTE) J. 31, 318–323.

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

Casey, E., 2011. In: Digital Evidence and Computer Crime, third ed. Academic Press, Cambridge, Massachusetts.

Casey, E., 2002. Error, uncertainty and loss in digital evidence. Int. J. Data Eng. 1.

Davis, R.J., Keeley, A., 2000. Feathering of footwear. Sci. Justice 40, 273–276. https://doi.org/10.1016/S1355-0306(00)71997-6.

Dietrich, D., Adelstein, F., 2015. Archival science, digital forensics, and new media art. Digital Investigation. In: The Proceedings of the Fifteenth Annual DFRWS Conference, vol. 14, pp. S137–S145. https://doi.org/10.1016/j.diin.2015.05.004.

Federal Aviation Administration, 2022. Risk Management Handbook (AA-H-8083-2a). United States Department of Transportation.

Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. In: Digital Investigation, the Proceedings of the Ninth Annual DFRWS Conference, vol. 6, pp. S2–S11. https://doi.org/10.1016/j.diin.2009.06.016.

Geradts, Z.J., Bijhold, J., Kieft, M., Kurosawa, K., Kuroki, K., Saitoh, N., 2001. Methods for identification of images acquired with digital cameras. In: Bramble, S.K., Carapezza, E.M., Rudin, L.I. (Eds.), Presented at the Enabling Technologies for Law Enforcement, Boston, MA, pp. 505–512. https://doi.org/10.1117/12.417569.

Girod, A., Champod, C., Ribaux, O., 2008. Traces de souliers. PPUR, Lausanne.

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

Hausknecht, K., Foit, D., Burić, J., 2015. RAM data significance in digital forensics. In: Presented at the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics. (MIPRO), pp. 1372–1375. https://doi.org/10.1109/MIPRO.2015.7160488.

Horsman, G., 2024. A template for creating and sharing ground truth data in digital forensics. J. Forensic Sci. 69, 1456–1466. https://doi.org/10.1111/1556-4029.15524.

Hoy, J., 2015. Forensic radio surveys for cell site analysis. In: Forensic Radio Survey Techniques for Cell Site Analysis. Wiley, pp. 1–2. https://doi.org/10.1002/9781118925768.ch1.

Joshi, B.R., Hubbard, R., 2016. Forensics analysis of solid state drive (SSD). In: Proceedings of 2016 Universal Technology Management Conference (UTMC). Presented at the Universal Technology Management Conference (UTMC), SDIWC, Minnesota.

Jost, C., Muehlethaler, C., Massonnet, G., 2016. Forensic aspects of the weathering and ageing of spray paints. Forensic Sci. Int. 258, 32–40. https://doi.org/10.1016/j.forsciint.2015.11.001.

Kind, S.S., 1994. Crime investigation and the criminal trial: a three chapter paradigm of evidence. J. Forensic Sci. Soc. 34, 155–164. https://doi.org/10.1016/S0015-7368(94)72908-X.

Locard, E., 1920. L'enquęte criminelle et les méthodes scientifiques, Bibliothčque de philosophie scientifique. E. Flammarion, Paris.

Lyle, J.R., Guttman, B., Butler, J.M., Sauerwein, K., Reed, C., Lloyd, C.E., 2022. Digital Investigation Techniques : a NIST Scientific Foundation Review (No. NIS T IR 8354). National Institute of Standards and Technology (U.S.), Gaithersburg, MD. https://doi.org/10.6028/NIST.IR.8354.

Microsoft, 2024. ControlPolicyConflict Policy CSP [WWW Document]. URL. https://learn.microsoft.com/en-us/windows/client-management/mdm/policy-csp-controlpolicyconflict, 12.12.24.

NIST, 2017. Computer Forensics Tool Testing Program (CFTT).

NIST, 2016. National Software Reference Library (NSRL) [WWW Document]. URL. https://www.nist.gov/itl/ssd/software-quality-group/national-software-reference-library-nsrl, 9.2.24.

OSAC, 2022. Guidelines for Dataset Development.

Ottmann, J., Breitinger, F., Freiling, F., 2023. An experimental assessment of inconsistencies in memory forensics. ACM Transact. Priv. Secur. 27. https://doi.org/10.1145/3628600.

Pintilie, E., Lecher, A., Doctor, E., 2024. Is Forensic Paint Analysis Affected by Environmentally Friendly Products? Lynn University Student Publications, Presentations, and Projects.

Pollitt, M., Casey, E., Jaquet-Chiffelle, D.-O., Gladyshev, P., 2018. A framework for harmonizing forensic science practices and digital/multimedia evidence. OSAC Technical Series. OSAC. https://www.nist.gov/system/files/documents/2018/01/10/osac_ts_0002.pdf.

Roux, C., Bucht, R., Crispino, F., De Forest, P., Lennard, C., Margot, P., Miranda, M.D., NicDaeid, N., Ribaux, O., Ross, A., Willis, S., 2022. The Sydney declaration – revisiting the essence of forensic science through its fundamental principles. Forensic Sci. Int. 332, 111182. https://doi.org/10.1016/j.forsciint.2022.111182.

Ryser, E., Spichiger, H., Jacquet-Chiffelle, D.-O., 2024. Geotagging accuracy in smartphone photography. Presented at the DFRWS APAC 2024. Brisbane.

Rzepka, L., Ottmann, J., Freiling, F., Baier, H., 2024. Causal inconsistencies are normal in Windows memory dumps (too). Digit. Threats: Res. Pract. https://doi.org/10.1145/3680293.

SJC, 2024. *Commonwealth V. Arrington*, 233 A.3d 910 (Pa. Super. Ct. 2020).

Spichiger, H., 2023. A likelihood ratio approach for the evaluation of single point device locations. In: Proceedings of the Tenth Annual DFRWS EU Conference, vol. 44, 301512. https://doi.org/10.1016/j.fsidi.2023.301512. Forensic Science International: Digital Investigation.

Stauffer, G., 2000. Modčle de Schallamach -Compréhension et description de ce phénomčne d'usure et exploitation de cette information dans le cadre de l'examen des traces de semelles, Séminaire de 4e année. Institut de Police Scientifique et de Criminologie. Université de Lausanne, Lausanne.

Sunde, N., Dror, I.E., 2019. Cognitive and human factors in digital forensics: problems, challenges, and the way forward. Digit. Invest. 29, 101–108. https://doi.org/10.1016/j.diin.2019.03.011.

Tart, M., 2022. Cell site analysis: changes to networks with time. Sci. Justice 62, 377–384. https://doi.org/10.1016/j.scijus.2022.04.001.

van der Pal, K.J., Sauzier, G., Maric, M., van Bronswijk, W., Pitts, K., Lewis, S.W., 2016. The effect of environmental degradation on the characterisation of automotive clear coats by infrared spectroscopy. Talanta 148, 715–720. https://doi.org/10.1016/j.talanta.2015.08.058.

Van Zandwijk, J.P., 2022. Have you been upstairs? On the accuracy of registrations of ascended and descended floors in iPhones. Presented at the EAFS 2022. Stockholm.

Van Zandwijk, J.P., Boztas, A., 2023. Digital traces and physical activities: opportunities, challenges and pitfalls. Sci. Justice 63, 369–375. https://doi.org/10.1016/j.scijus.2023.04.002.

Van Zandwijk, J.P., Boztas, A., 2019. The iPhone Health App from a forensic perspective: can steps and distances registered during walking and running be used as digital evidence? Digit. Invest. 28, S126–S133. https://doi.org/10.1016/j.diin.2019.01.021.

Yoo, E.-H., Roberts, J.E., Eum, Y., Shi, Y., 2020. Quality of hybrid location data drawn from GPS-enabled mobile phones: does it matter? Trans. GIS 24, 462–482. https://doi.org/10.1111/tgis.12612.