Chapter 1

ON DETERMINING THE AGE OF QUESTIONED DIGITAL DOCUMENTS

Martin Olivier

Abstract The determination of the age of a document is often of paramount importance when questioned documents are examined. In a physical setting, such age determination may be based on static and/or dynamic principles. Dynamic principles are based on the manner in which documents change over time. Such aging has no equivalent in the digital domain. The static methods are based on the introduction date of ink, paper and other writing materials. Digital documents are based on artefacts that were often introduced on known dates; the static methods to date document therefore have digital equivalents. The paper introduces two principles that are used as a basis for determining the age of a document. The paper uses these principles to determine a temporal creation window for a document, in a manner that makes it possible to speak about the confidence in such a window. In those cases where a creation window can be shown to pre-date or postdate a purported creation time or other critical event, it follows that the document cannot be authentic.

Keywords: Questioned digital document examination, forensic digital document examination, digital document age

1. Introduction

The forensic discipline of questioned documents is a well-established field that has proven its value over many years. The discipline uses various techniques to answer a range of questions about documents about which questions arise in a legal matter. Questions may exist about the authenticity of a document, the tool(s) used to create the documents and authorship (for example, based on the signature that may be present on a document). A question that exists in its own right, or as an aid to answer other questions, is the question about the age of a document.

The age of a document is a relevant question in a variety of cases. A collectible antique document has to date back to its purported time of creation to be authentic. A discrepancy is an example where the document enters the legal fray; if authenticity is challenged after the sale of such a document, civil and/or criminal proceedings may follow.

The age of a document is also often relevant in financial matters. A common example is that of a contested will: an authentic will could not have been executed after the person's demise. Typical financial documents, such as receipts, invoices and promissory notes are all examples of documents where creation times may be and are questioned.

Digital documents often contain one or more embedded creation dates. Such dates are often recorded as metadata in the document, and by the file system on which the document may be stored. The reasons why one cannot rely on such metadata are well-known in the forensics community. However, given its importance, the issue is revisited in section 1.2 below.

The age of physical documents is determined using two distinct approaches. The static approach identifies inks, stationary and other materials used to create a document. This is combined with knowledge about the time of introduction (and use) of such materials. The document has to postdate the introduction of materials used. The dynamic approach considers the changes that physical material undergoes over time. Changes include fading, oxidation and other forms of physical decay. If the rate of decay is known, the age of the document can be determined (noting that decay is influenced by a range of variables). Digital documents are not subject to physical decay. Hence the age of a digital document has to be determined using a static approach that considers the artefacts used to create a document.

This paper systematically develops a static approach to forensically determine the age of digital documents. In order to justify forensic conclusions, the paper introduces two principles that it uses as a basis to reach conclusions. The *principle of inclusion* and *principle of replication* are introduced and expounded upon later in the paper.

Kapoor et al [29] highlight that age determination of a (physical) document is one of the more challenging facets of forensic document examination. The constraints of the digital sphere (such as a lack of physical degradation) imply that digital document age determination will also be challenging. However, a documented process promises to be useful in those situations where it leads to a justifiable conclusion.

The paper is structured as follows. Section 1.2 revisits the lack of trust in metadata. Section 1.3 considers a few examples of mostly physical documents where the creation dates were important. Section 1.4 considers the essence of how the age of a physical document is determined and

then applies the relevant parts to determination of the age of a digital document. Sections 1.5 and 1.6 describe the principles of, respectively, replication and inclusion that are used to determine the creation window of a digital document in section 1.7. Section 1.8 concludes the paper.

2. Metadata

As noted in the introduction, various forms of metadata pertaining to the age of a document may be associated with the document. Embedded metadata often indicates when the document was created. The entry

```
/CreationDate(D:20120122074131-08'00'
```

in a PDF document indicates the time and date when and time zone where the document was (purportedly) created. The file system on which a document is stored typically keeps a record of when the document was created as well as the times at which the document was modified or accessed last. Depending on context, other metadata may be available.

The value of metadata for forensic examinations has been contested over the years. In an early paper on digital forensics Buchholz and Spafford [15], for example, express the wish that systems would capture more metadata, given its utility in the (then new) field. However, they also lament the ease with which metadata can often be modified. They suggest a future where some metadata (such as file creation dates) cannot be modified — a feature that "should be present on any file system." It often remains trivial to modify many instances of metadata — as already illustrated by the cited paper by Buchholz and Spafford.

Various techniques can be used (in some cases) to increase confidence in metadata, and in particular, in timestamp metadata. One popular technique is to correlate various instances of metadata [30, 35]. Often a lack of correlation is used to detect tampering [35]. In contrast, in what follows in this paper, specific instances of correlation will be used to infer correctness of timestamps.

The lack of trust in timestamp metadata is illustrated by the number of papers that focus on the detection of such tampering [35, 31, 38, 40]. As noted, some such papers are based on a lack of correlation. A different strategy is the use of machine learning to detect anomalies, despite serious concerns about the use of AI when forensic proof is required [33]. Yet another popular approach is one that stores data in a tamper-proof environment (for example, by using blockchain technology), which solves the problem, but only if one is able to convince a reliable party to store all such values in the tamper-proof container before it becomes relevant in a legal matter.

The age of a document usually becomes contested some time after creation of the document; often it would a document that was moved between systems using various document sharing mechanisms. Hence, operating system metadata would be irrelevant. A robust approach that is not subject to natural changes or simple tampering is required.

The paper deems a document to be a sequence of bytes that would lead to a common understanding in the digital sphere. Such a sequence of bytes would usually form a file of some specific type. Examples range from typical documents created by office suites to executable programs. An important consequence of this definition that that saving a document in a new format (such as saving a `.doc` document as a `.docx` document) would create a new document. This requirement is necessary for the static analysis approach followed in the paper. It is possible to consider similarities between, say a document stored in a `.doc` format and another document stored in a `docx` format ans explore the possibility that one of the documents is simply a conversion from the other. The current paper does not explore such comparisons.

Differences between various metadata dates in a document (or a difference between a date in the content and a metadata date) may be of value to dispute the authenticity of a document, but the certainty with which the document may be deemed to be forged depends on the nature of the document and the differences in times. Some documents may include multiple creation times in their metadata. As an example, a photo captured on a camera that geotags the photo typically stores the capture time based on the camera's internal clock, as well as based on the clock of the GPS satellites used to determine the location where the photo was taken. A forger *may* fail to change all metadata and such an inconsistency *may* provide the forensic document examiner with a data point to determine that a questioned document cannot be authentic; however, other explanations for such differences need to be ruled out before any conclusions can be reached.

Time and datestamps in metadata play a different role in 'autopsies' or other forms of even reconstruction than they do in the case of questioned documents. Often relative time is more important than absolute time. Such use of timestamps is outside the scope of the current paper.

3. Document age used as evidence

The age of documents were at the centre of a number of well-known cases. This section briefly introduces a case where the age of a signature on a will was instrumental to prove the will a forgery. Then it moves to the well-known case of the forged Hitler diaries. Finally, another

questioned will case is considered where the will in question was sent by email and eventually deemed to be acceptable by the court.

3.1 Patel vs Patel, EWHC 133 (CD)

Around early 2015 a certain Girish Patel petitioned the England and Wales High Court (Chancery Division) to accept a will of his late mother that was dated 23 June 2005. His mother passed away in 2011 and a will dated 18 June 1986 had earlier been accepted as her last will and testament, and was now contested [41]. It was common cause that the mother's signature on the newer document was authentic. However, expert testimony highlighted a number of temporal inconsistencies: (1) People's signatures change over time and the version of her signature on the new document differed from other known copies of her signature that dated from 2005 — when the new document was purportedly signed. It matched versions of her signature that dated from before 2005. (2) The ink used for her signature exhibited much more signs of ageing than the ink used for the witnesses' signatures. (3) Ink particles from the printer was found on her signature that suggested that the signature already existed when the document was printed. None of these observations have equivalent notions in the digital domain. Observation 3 above was challenged and the court did not base its decision on it.

Given that this was a civil matter, the case was decided on the balance of probabilities — an issue that is raised repeatedly in the judgment. The court ruled against the petitioner and determined that the 2005 document was not authentic. In a subsequent case Girish Patel was incarcerated on the grounds of perjury [42]. Later Girish Patel was also found guilty of forgery in a criminal case. Neither of these later cases (that required proof beyond reasonable doubt) required further examination of the forged will.

3.2 The Hitler diaries

The Hitler diaries were purportedly written by Adolf Hitler, but were, in fact, forged by a Konrad Kujau in the early 1980s [24].

One of the first concrete forensic contradictions raised, was temporal: polyester was found in the binding of one of the diaries; polyester manufacturing commenced a number of years after Hitler's death. Chemical analysis of the ink used in the diaries eventually proved that the content was written one to two years before the diaries were offered for sale.

After a lengthy police investigation, a confession by Kunjau and an extended trial, Kunjau was sentenced to serve time in prison. Various

other parties who were involved in the scandal were also sanctioned — including incarceration in one case.

Finding polyester in the binding of the diaries is an example of static analysis approach for which digital equivalents will be explored in the remainder of this paper.

3.3 Van der Merwe v Master of the High Court and another [2010] ZASCA 99

In 2004 a certain John Henry Munnik van Schalkwyk executed a last will in favour of an animal welfare society. At some later stage he and a friend, Hendrik van der Merwe, decided to mutually bequeath their entire estates to one another. They had been friends over a long period, and neither had any remaining family. On 26 July 2007 Van Schalkwyk sent an email with a will attached, bequeathing his estate to his friend, and asked the friend whether the will correctly expressed their agreement. John van Shalkwyk passed away only a few months later, without ever signing the will. The Act that governs such wills in South Africa prescribes all the formalities that have to be met before a document constitutes a valid will. It also allows a court the discretion to declare a 'will' valid even if all formalities have not been met; a person's last wishes should not be ignored simply because of some formality — as long as they were indeed the deceased's last wishes.

The friend of the deceased approached the South Gauteng High Court in Johannesburg with the request that the emailed document should be declared a valid will. The respondents were the Master of the High Court and the Society for the Prevention of Cruelty to Animals (SPCA). The Master responded as the party who decides whether to accept a will or not. The SPCA acted as a respondent since it was the original beneficiary in the 2004 will. Both parties simply indicated that they would abide by the decision of the court. The high court decided that accepting the document would "open the floodgates for any person to submit any document…as a Will of a testator" [32, par 10]. It accordingly dismissed the application.

On appeal, the Supreme Court of Appeals (SCA) reached the conclusion that it was indeed the wish of the testator to leave his estate to his friend. One aspect they considered in support of this conclusion was the fact that a copy of the email was still present on the computer of the deceased; no further analysis of the document was conducted. The court decided that the emailed will was to be deemed acceptable and directed the Master "to accept the [emailed] document executed by the deceased

during 2007 … as the will of John Henry Munnik van Schalkwyk for the purposes of the Administration of Estates Act 66 of 1965" [32, para. 19].

4. Determining the age of a document: Static and dynamic approaches

While digital documents do not age physically, the 'material' (file formats, fonts, and so on), may be used to determine a boundary for the age of a document in those cases where the introduction of such formats to the market is known. This is not dissimilar to static methods use to bound the age of physical documents. The paper will explore methods used to determine the age of physical documents and then turn its attention to digital documents.

4.1 Physical documents

Already as early as 1910 textbooks [34] discussed the reasons to determine the age of questioned documents, as well as methods to determine such an age. The reasons include detection of forgeries when a document's purported age does not agree with its physical age. This mismatch between actual and purported age was deemed to be important in a range of contexts, including "bankruptcy and settlements of estates [where] the important question may arise whether certain entries or memoranda in books of account were actually made in due course of business on the dates they bear, or whether they were made at a subsequent period for the fraudulent purpose … of showing a certain result at the date of settlement" [34, p.432].

"The materials that make up a [traditional] document may tell something of the earliest possible date of preparation" [25, p.273]. When text is written or typed on stationary where the date when the production of the stationary is known, this may be useful. The ink may be based on a type of fluid or gel that embeds pigments that were only introduced into writing instruments at some known date, which may be useful. An option to determine the age of a document is to use physical changes that occur over time. Paper and ink, for example, oxidise over time; the degree of oxidation may be useful. Obviously these physical indicators of age have no proper equivalent in the digital sphere. However, the notion of *earliest possible date of preparation* is useful; an equivalent notion of *earliest creation time* (ECT) will be used when digital documents are discussed in subsequent sections of this paper.

Entire books have been written just based on the age of ink. *Advances in the Forensic Analysis and Dating of Writing Ink* [14] is an example of such a book. It discusses, amongst others, various fountain pen inks

introduced in the 1880s and 1940s, various ball-point inks introduced from 1939 to 1968, fibre pen inks introduced in the mid 1960s and gel-based inks introduced in the mid to late 1980s. Many tests "are based on solvent extraction and changes in dye concentration" [14, p.7], but such tests are, amongst others, not suitable for gel-based inks. Such methods (and tests that are based on oxidation) provide not only an earliest possible date of production, but specific time time period during which the ink was applied to the paper. A number of factors impact the duration of the time 'window' during which it could have been created, as well as impact on the certainty with which a conclusion about age may be made. Clearly, no equivalent exists in the digital domain.

A distinction is made between static, dynamic and supplementary approaches to determine the age of a document [29, 16, 17]. Static approaches are based on the introduction dates of types of paper, types of ink and/or other components used to produce a document. Dynamic approaches depend on the changes of physical material (such as paper or ink) over time. Supplementary approaches include methods such as radiocarbon dating and subjecting (parts of) the document to conditions that enable the examiner to study the reaction of the document to such conditions. It also includes chronology, where it may be possible to arrange documents that share an origin in sequence. This may, for example, be two letters from a pad where impressions from handwriting on one page can link another page to its original place in the pad. This enables the examiner to establish relative ages of related documents.

The notion of chronology, as expressed by Kapoor et al, corresponds broadly with a remark by Hilton: "Another technique [to help determine the age of a physical document] involves comparison with known undisputed material, that is, by comparing it with a series of documents prepared with the same typewriter or pen and ink, or paper, for example, and fitting the document in question into such an established framework" [25, p.273]. In the digital context, email is one example of documents where earlier messages often leaves 'impressions' in later documents (in the form of quoted earlier messages included in a new message). This may enable the examiner to establish a *relative order* between such messages. However, email may also be a document type that can be dated more accurately than most other digital document types — as will be discussed later in this paper.

4.2 Digital documents

Not much research has been conducted on methods to determine the age of digital documents.

Spennemann and Spennemann [43] use the revision save identifier (RSID) in MS Word documents to determine the order of versions of a document. Each version that is saved retains the RSID of the earlier version(s) and adds a new RSID to the list. This enables one to determine the 'genealogy' of such documents, including instances where more than one document stem from a common shared ancestor, but where their histories are different since the occurrence of the common ancestor.

The disc allocation strategy used by a file system determines the manner in which the fragments that constitute a file are spread across the disc. In some cases it is possible to validate the date stamp of a file by comparing the date of a (recovered) file with files that have been allocated to neighbouring clusters on the disc [8]. This is simpler with older file systems, such as the FAT file system. A number of papers have been published that consider the temporal behaviour of NTFS drives — see [20, 22, 9] as examples.

Casey [19] considers how digital activities occur in time windows that he refers to as *strata*. Events that occur during a stratum need to be seen in context; while useful conclusions are often possible, room for misinterpretation exists. A newly created file may retain the file system data (in particular, the creation time) of the old file. Casey's paper notes a case where saving an MS Word file (using *Save as* with the old name of the file) causes the file to retain its creation date on the file system, but stores the *Save as* time as the creation time in the file itself.

Ho, Kao and Wu [26] derived a number of rules about the impact of managing files in the cloud has on their time stamps (compared to their timestamps on systems such as FAT, NTFS and Ext4). None of these papers speak to the age of a document, though. Their primary goals are event reconstruction or deleted file recovery.

5. The principle of replication

> *When multiple independent authoritative sources agree about an event, it follows that details about the event have been recorded (and reported) correctly.*

To explore such a principle, the notions of *multiple*, *independent* and *authoritative* have to be explored.

In this context we will deem a source to be *authoritative* if it records details about some action it is responsible to handle. The context in which the principle will be used in the paper is that of email. A mail transfer agent (MTA) that forwards an email on its route to its destination is the authoritative source about what transpired when the email was forwarded. Although being deemed authoritative, it may be mistaken about some details — the specific concern in the current paper is

that of time. If the clock of the MTA is wrong, it may report the time of the event incorrectly.

The second criterion used in the principle of replication is the requirement that the observation must be supported by *multiple* 'witnesses'. It seems obvious that an increase in the number of sources from which an observation may be confirmed increases the confidence with which the observation may be accepted as correct, simply because the odds that such agreement can be ascribed to chance decreases. However, if two sources of an observation used a common third party to make the observation, the common third party may be a better explanation of agreement than relying on the odds of two witnesses being equally wrong. Recall that the domain in which the principle that will be applied below is that of forwarding emails. If two MTAs share a clock both would report the event to have happened at the time indicated by the clock — which may reflect time incorrectly. However, if each MTA has its own clock, the odds that they would both would reflect the same incorrect time are small. In general, the greater the number of witnesses that report the same observation (for which no explanation, such as a shared clock, can be found), the greater the odds that the observation is indeed correct.

Observations reported by multiple witnesses that obtain formation from a shared source (in terms of some specific attribute, such as time) effectively count as a single observation in terms of this principle. Suppose an email is sent from an organisation A where several MTAs handle the email. However, all MTAs effectively share a clock. Then all these observations combined have the same weight that a single observation would have had. Suppose this email now arrives at an MTA at organisation B, which uses a clock not linked to A. This forms a second observation; if it agrees with the observations at A, chance does not explain this correlation. Suppose the email is now handled by several MTAs at B, where all the MTAs at B share a clock. If these additional MTAs agree, confidence in the observations does not increase. Effectively, the MTAs at A form one witness and the MTAs at B form a second witness.

It is necessary to reflect on what it means for two witnesses to 'agree'. Often processes will take some time to complete. Each MTA may forward the email to the next MTA after some very short delay (but longer delays do occur). It is suggested that it is possible to empirically observe email forwarding times and determine expected variance. Most emails will be forwarded within some limited time period. An appropriate meaning for *most* should be determined empirically, but it is possible that, say, the forwarding time for 99% of emails are clustered together with about 1% being outliers. Whatever percentage of times are within

such a cluster may then be deemed to be 'normal' and assumed to be 'normal' variance, while outliers may be deemed as a sign that the times reported differ — and that no conclusion should be based on them. Empirical testing is required to determine whether email forwarding does exhibit the hypothesised behaviour.

The final requirement used by the principle of replication is *independence*. This has already been encountered above where it was determined that multiple witnesses that 'depend' on the same source, effectively count as a single source. Witnesses A and B are independent in terms of a specific observation if A did not obtain the information from B, B did not obtain the information from A and A and B did not both obtain the information from some shared origin.

Again, referring to the domain of application later in this paper, each MTA independently records the time an email is processed by the MTA if the MTAs do not share clocks. Hence, they are independent 'witnesses' about the time the email was processed. In contrast, the fact that the email was sent is not independently recorded by each MTA. Every MTA 'learns' about the email from the MTA from which it receives the email. A malicious actor who is able to introduce a forged email at an appropriate MTA may cause the subsequent MTAs on the route to also record that such an email was sent.

The fact that the various MTAs should not rely on the same source has at least two implications in this context. If MTAs share a clock (or derive their time from the same NTP server), they may not be deemed to be independent. Given the fact that NTP servers often obtain times from other, more authoritative, NTP servers makes independence a topic that requires extensive reflection. The second implication is that the same party should not be able to gain administrative access to two MTAs for them to be deemed independent. In principle, the better the governance at at least one of the MTAs, the less likely it becomes that an arbitrary malicious actor would have gained access to it. This also requires extensive reflection.

Rather than exploring these implications in detail, it is known that most emails are currently handled by one of a handful of email service providers. It will be argued in section 1.7 below that it is practical to subject such services to regular intermittent tests that compare the times provided by those servers to one another, which, as will be argued is a sound basis to accept independence. We therefore do not explore the minutiae to confirm independence between less common services further in the current paper.

The principle of replication was influenced by Casey's certainty scale [18]. Casey's certainty scale also uses notions of replication, protection

against tampering and independence to assign increasing levels of confidence to conclusions. Some of Casey's three indicators (independence, protection and replication) are, in principle, independent and should therefore lead to a partial ordering of certainty, rather than the fully ordered scale he presents. It is also not entirely clear that independence and protection are not two sides of the same coin. Moreover, Casey's scale is intended for conclusions, whereas the principle presented here deals with certainty about *observations*. A full critique of Casey's certainty scale is beyond the scope of the present paper. It is sufficient to both acknowledge the inspiration drawn from Casey's scale and to note that it does not meet the requirements of the current paper. Hence the paper cannot simply have used Casey's certainty scale, but had to put forward a new principle.

6. The principle of inclusion

The primary principle to be used in this part of the paper is the observation that whenever some artefact x is used to create some artefact y, x must have been created before y could have been created. In what follows, the notation $x \sqsubseteq y$ will be used to indicate that x was used to create y. Stated somewhat differently, the notation indicates that x in some sense forms part of y.

Suppose $c(x)$ indicates the creation time and date of some artefact x, then the observation above may be expressed as follows:

$$x \sqsubseteq y \implies c(x) \leq c(y) \tag{1}$$

Equality (that is $x \sqsubseteq y \implies c(x) = c(y)$) may occur due to the fact that digital time is discrete — see the discussion by Cohen [21].

In what follows the notions of *creation time* and *creation date* will be used almost interchangeably. Rather than using a specific date or time the discussion will evolve towards a period of time that will be referred to as a *creation window*.

The term *artefact* above was used to refer to anything that came to exist at some time. The notion of artefacts includes digital documents; the principle above is therefore useful to determine the relative age of two documents if the one document includes the other. An example that will be used extensively below is where an email (which is obviously deemed to be a digital document) includes or embeds another document as an attachment. Clearly, a composite document has to include the document used in a form that leaves no doubt that the used document existed at the time the composite document was created. An obvious example of a case where this does not apply is where an HTML file x links to some other HTML file y. It is possible replace the referenced

document y with a different document also called y, without changing the referencing ('composite') document at all. Hence (in the case of documents, at least) $x \sqsubseteq y$ can only be said when (if possible at all) modification of x would require y to change too.

However, the relationship between x and y may be indirect. Where y contains a sufficiently strong cryptographic hash of x, the condition that x forms a part of y is met, even though x is not included or embedded in y: If x is modified it would be necessary to modify y to maintain the relationship between x and y. Modification of x without modifying y would have been possible if another document x' could be found, where x and x' have identical cryptographic hashes. For sufficiently strong cryptographic hashes, this is not possible [37].

6.1 Format specifications

One example of using an artefact to create a digital document is the use of the specification of the document type. Consider PDF as an example. The first PDF specification was released in 1993 [11]. Hence a PDF document cannot predate 1993. Several versions of the PDF standard have been released over the years, as summarised in table 1. In principle, a document that is represented using a specific version of PDF could not have been created prior to the definition of that version.

A challenge based the data provided in table 1 is the fact that the precision is limited to an entire year, which may not be ideal. In some instances (if not all) it may be possible to determine a more precise date. As an example, the publication date of the PDF 1.1 standard is 1 March 1996 [12]. However, it is obvious that such a standard does not come into existence within a day: Many people would have worked on the standard. Where a standard is linked to some specific product, programmers, alpha testers and beta testers may all have worked on code in which the standard was used. Importantly (for a relatively complex standard) such people typically would have had access to tools (in the form of prototypes, work-in-progress, release candidates and, typically, final code of the product) before the standard would have been 'frozen' and published. This was clearly a more significant issue in the case of PDF 1.1. The specification for PDF 1.1 [12] was only published in 1996. However, PDF 1.1 was the "the native file format of the Adobe Acrobat 2.0 family of products" [12, p.3], which was already released in 1994 [6].

At the same time, new standards often need time to be adopted in the market. Many end-users will only start using the new standard when they upgrade their software to versions that support the new standard. Hence, even when a standard is adopted on a specific day (or even year),

Version	Release date	
1.0	1993	[11]
1.1	1994*	[12]
1.2	1996	[13]
1.3	2000	[1]
1.4	2001	[2]
1.5	2003	[3]
1.6	2004	[4]
1.7	2008	[5]
2.0	2017†	[27]
2.0	2020†	[28]

*The standard for PDF 1.1 was only published in 1996; however, that standard notes that PDF 1.1 was used in Acrobat 2.0. Acrobat 2.0 was released in 1994 [6].

† A standard for PDF 2.0 was published in 2017 [27]. That standard was revised in 2020 [28].

Table 1. Years in which PDF specifications were released

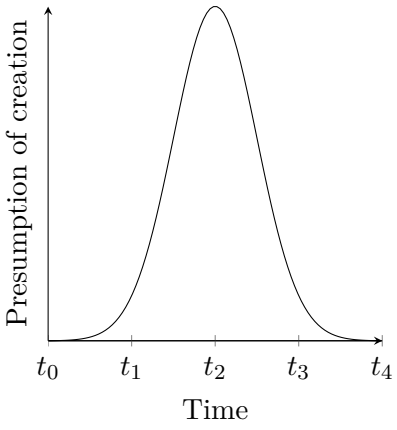

Figure 1. Creation window of a standard

the assumption should be that the standard became available over a period of time. Figure 1 illustrates this.

Note that figure 1 is not intended as a probability density function. The fact when an artefact, such as a standard, was created cannot be expressed as a probability. The uncertainty rather speaks to knowledge of and access to such a standard. Initially $(t_0 - t_1)$ few people have knowledge of the standard to be developed. At some stage (t_1) the

standard is published. Over time $(t_1 - t_2)$ the market 'adopts' it and it is supported by an increasing number of tools. At some time (t_2) the standard may be deemed to be available to everybody who wants to use it. From this point onwards $(t_2 - t_3)$ the number of new adopters decline. Eventually (t_3 and beyond) there is little reason to believe that any tool that uses the standard was created because its developers only recently learnt about the standard.

The figure should therefore be interpreted as follows: Suppose a document is based on a standard that was created in the window as depicted in figure 1. If nothing is known about those responsible for creating the questioned document, it could have been created as early as t_0. Where there is reason to accept (either as fact or under a version of events offered by a litigant) that those who created the document did not have anything to do with the development of the standard, it becomes possible to accept t_1 as the earliest date at which the document could have been created. In fact, given that the majority of users of a standard would not have been involved with the creation of the standard, it is possible to use the later t_2 as the first possible time on which the document could have been created; however, such an assumption reduces the certainty of the conclusion.

While attributes of individuals should not normally have an impact on digital forensic examinations, there seems to be issues of access here that merit consideration.

Note that no inference should be drawn from the scale used on the graph in figure 1: The x-axis scale have equidistant labels, but the assumption cannot be that intervals are equal. The increase and decrease in the confidence that the standard was created at some time may be asymmetric — as depicted in figure 2. In fact, the entire depiction of the creation time serves merely as a placeholder that may, in fact, take any shape over a period of time.

When the law of inclusion (equation 1) was introduced, it was noted that the case for equality would be revisited. The discussion of PDF standards enables us to justify equality based on precision: The standard for PDF 1.2 was published in 1996 [13]. It is entirely plausible that some documents based on version 1.2 were created in 1996. This extends to the more generic case where the creation date of a standard should be deemed to be some period in time, and a document's creation date may well overlap with the creation date of the standard on which it is based. Such a possible overlap now becomes an issue that needs further consideration. Overlap will be discussed below.

Another challenge already touched on, is the manner in which the version of a PDF document is to be determined. It is known that a

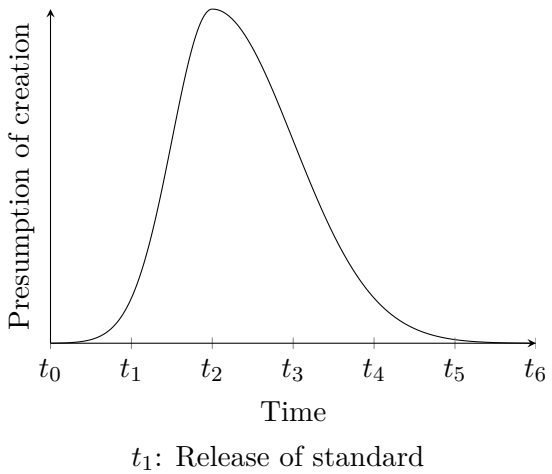

Figure 2. No relationship between the increase and decrease in adoption rate is implied

PDF document contains its version in the very first line. A typical PDF document may start as follows:

```
%PDF-1.6
```

This is true even for version 1.0; according to its standard the "current version is 1.0, for which the first line of a PDF file is `%PDF-1.0`" [11, p.42]. However, PDF standards are backwards compatible. The standard for version 1.2, for example, states that "1.0-conforming files and 1.1-conforming files are also 1.2-conforming files" [13, p.62]. Note that, starting with "PDF 1.4, the version in the file header can be overridden by the **Version** entry in the document's catalog dictionary (located via the **Root** entry in the file's trailer …)" [2, p.63]. The version entry may therefore provide a more accurate (or later) indication of the PDF version of the standard used in (parts of) a PDF document than indicated by the document's header.

While the version number in the header and/or trailer of a PDF document indicates the version that the creator of the purports to use, the elements used in the created document may have all already been defined by some earlier version of the PDF standard. The observed version number may therefore not be supported by the use of new elements in the indicated version. It is easy for a fraudster to predict a future version number; it is much harder for a fraudster to use functionality from a future version. In fact, this version number is metadata that is subject to the concerns expressed in Section 1.2. However, a document that claims

to be represented in some version and that uses a feature introduced in that version provides internal support about the correctness of the version. The examiner should weigh the credence of the version depending on the observations that support a claim that a specific version was used, and adjust the confidence in conclusions accordingly.

Similar comments apply to many other document formats to a greater or lesser extent than PDF. In some cases a document standard is not proposed in any public manner, and introduced to the public by the fact that some software starts using that (proprietary) format. Compatible software may then depend on reverse-engineering of the format, or the format my be published long after the software that introduced the format was made available. In such cases, the software should be deemed to be the original defining document of the format. Some formats may continually be amended or extended by software houses — especially where the format makes it possible to include information that would be ignored by tools that are not aware of such extensions. Consider HTML tags introduced by various browsers (and, at least, conceivably supported by HTML authoring tools). Netscape was only one example of a browser that introduced proprietary tags, and `<blink>` is one such example [10]. In such cases the creation time of the tool may again be more useful to determine aspects of the age of the document than the creation date of the document standard would be.

6.2 Character codes

Some encoding is always used to represent text in any document that includes text. The principle of inclusion may therefore be applied to help determine the age of a document. The ISO/IEC 8859 series of standards is one example of character encodings that were developed over a period of time (1987 to 2001, Parts 1 to 16). It should be noted that these standards are based on ASCII (and are backwards compatible with ASCII). The ASCII standard was first introduced in 1963 [7]; hence a document may use '1963' ASCII, but claim to be encoded in ISO/IEC 8859-15 (which corresponds to IANA character encoding ISO-8859-1, which an examiner would likely find in a MIME description of character content). This again raises the question: Does the examiner rely on an annotation that the document was encoded using `iso-8859-1`? Clearly the meaningful use of at least some characters in that standard that were not part of ASCII would increase the confidence with which the examiner can conclude that the document was created after the adoption of the standard. To exacerbate matters, various parts of the ISO/IEC 8859 series are based on earlier standards proposed by other bodies (such as

ECMA). In some cases the earlier standards have been modified before being accepted as an ISO/IEC standard; this usually involves very few characters and, unless those characters are used in some meaningful manner, use of the earlier standard cannot be distinguished from use of the later standard, unless it is identified by name.

Unicode has become widely adopted and forms a potentially useful tool to help determine the age of documents. New characters are added annually to the standard, with Version 16.0.0 due for release in 2024. A new version typically includes support for more (often ancient) scripts, new characters and new emojis. The version of Unicode used in a document is usually not explicitly indicated; hence the examiner has to rely on characters that were introduced by a specific version to identify the version. Unfortunately the set of new characters added annually are typically specialised characters (such as those from ancient scripts) and unlikely to occur in most documents. The introduction of new currency symbols is a noteworthy case for documents that include financial references. The Euro sign was, for example, introduced in Unicode 2.1 (May 1998); the BitCoin sign was introduced in Unicode 10.0 (June 2017); the new Indian Rupee sign was introduced in Unicode 6.0 (October 2010); while a sign exists for Ethereum, it has not been added to Unicode yet, and may be added in future. Such currency symbols regularly occur in specific documents (such as contracts), and may be useful in those, admittedly isolated, cases where they occur.

Similarly, new emojis that are added may be very useful in those contexts where they occur. As one (arguably atypical) example, consider the microbe emoji introduced in Unicode 11.0 (June 2018). This emoji was used extensively in various chat and other (mostly informal) communication channels during the Covid-19 pandemic.

ISO formally began working on an international character set in 1989 and published the first draft of Unicode in 1990. The standard (ISO/IEC 10646-1:1993) was published in 1993. Work on an efficient encoding scheme for locations that primarily use the Latin character set began soon after work on Unicode started. UTF-8 (Unicode Transformation Format – 8-bit) was created in 1992, presented at a conference in 1993 [39] and included in the 1993 Unicode standard. Locations that primarily used other scripts followed suit. Shift JIS was designed to concisely encode (primarily) Japanese script; it was standardised in 1997.

Whenever a document is encoded using a character code or a special representation for that character code, creation of the character code or the representation format must predate the document (as described in the principle of inclusion).

7. The creation window of a document

The creation process of a document may be such that it leaves sufficient traces to enable the examiner to determine its creation window. Email is an example of such documents and discussed in some detail below. Secondly, from the principle of inclusion, it is known that, if a document x is included in a document y, document x was created prior to document y. If the creation window of y is known, it established the latest creation time (LCT) of document x. Email is regularly used to transmit other documents (as attachments) and therefore forms one specific case where this principle can be used.

7.1 Email

An email is typically created on a source computer, transferred via (possibly independent) mail transfer agents (MTAs) and eventually delivered in the recipient's mailbox. Even in cases where a cloud service is used to create, send and receive email, multiple independent mail transfer agents may still be involved and maintain (independent) logs of when an email was transmitted by one MTA and was received by another. It is possible that the logs of the various parties that participate in email transmission may not all agree. It is, for example, entirely possible that an email is created on a computer, while the computer is not connected to the network. Once the computer connects to the network, the email may be transmitted to the first MTA en route. It is also possible that some delay occurs when one MTA forwards an email to the next MTA, due to congestion or other network-related conditions. In many cases examination of the email header will indicate the points on the path where the email was delayed and this would often suggest a reason that would explain the delay. In any case, the creation date of a document has been assumed to be a period of time above and such a period caters for clocks that do not entirely agree.

The challenge caused by incorrect clocks is well-known, but arguably has become less of an issue in the always-connected world in which most emails are sent. Moreover, a few email services are involved in a large percentage of email sent globally. While it is hard to precisely determine the market share of services such as Google Mail, Outlook and Proton Mail (and which metrics to use to determine market share), it is clear that such services handle a large percentage of emails. Numbers provided by various sources on the Internet tend to estimate Google's share to be in the range of 27% to (significantly) more. The exact market share is not important for our purposes; the fact that these services are popular is obvious. When email is sent between two or more parties, the

likelihood that the email would be handled by a major email provider increases. We may hypothesise that such big email providers tend to record the time they handle an email quite accurately. This hypothesis is simple to test on a continual basis, given that few email services are truly major handlers of email: With an empirical record of correctness (including variance) of time and date stamps by a major email provider, the examiner is in a better position to make claims about the date the email was processed by that provider. Where a reliable clock is not available to test accuracy, testing agreement between clocks by pairs of major providers (as per the principle of replication) is a viable alternative. Regular testing may be augmented using (archived) reports of network events that may have an impact on processing of email.

In fact, major Google Mail delays are rare enough to be reported on by the news media [36]. Google maintains a dashboard of incidents that they investigated [23]. In 2023 only nine incidents related to Gmail are listed. Four listed incidents involved a delayed processing of email: On 1 December 2023 an incident lasting 1 hour, 32 minutes caused global delays. One day earlier, on 30 November, an incident lasting 4 hours, 45 minutes caused global delays. On 5 May 2023 an incident lasting 56 minutes caused delays to for some customers in Spain and Morocco. On 26 January 2023 an incident lasting 1 hour, 48 minutes caused delays in some specific cases that were initially deemed to have been delayed mailing list processing. This clearly demonstrates that major delays are rare. Care should, however, still be exercised when dealing with the delay of a specific email. In any case, a delay would typically cause various times of email processing to disagree and hence make it difficult to determine the age of the email; therefore it is not likely to result in wrong conclusions being reached. Agreement of timestamps is more important for the examiner.

In the case of older emails, MTA logs are unlikely to be available. In many such cases the time at which the email was purportedly sent would be common cause (since it is recorded in the body of the email). It is also realistic to assume that the email, as received, with headers from MTAs will be available. In such a case, where the times agree, the times may be accepted with a high degree of confidence (given independence). Confidence may be increased where the email conversation in which the document was exchanged is available for examination.

Various other online services (such as cloud storage and instant messaging) may also shed light on the time at which a document already existed. We do not canvass those services in the current paper.

The period over which such an online document was created will be depicted in a similar manner to the depiction of creation periods used

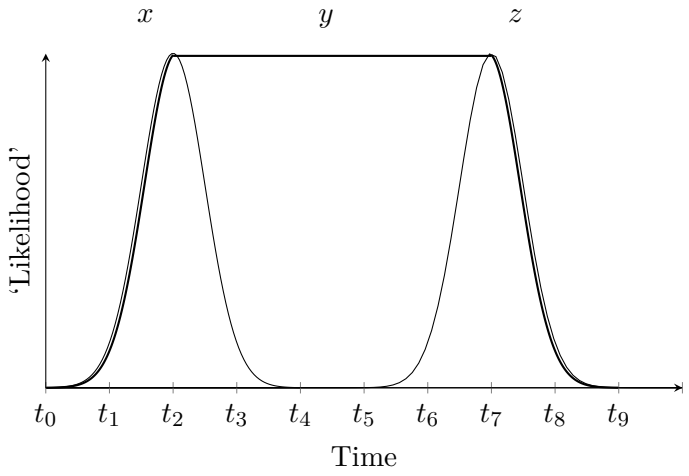

Figure 3. Creation window of y (t_0–t_9) from ECT (t_0–t_4) to LCT (t_5–t_9)

earlier. The interval covered is expected to be much shorter, with more specific boundaries from the earliest to the latest point the document could have been created. Email is also open to further examination given the partial ordering of messages, replies and forwarded emails in a conversation. The principle used is again one of inclusion, where a reply or forwarded email includes the earlier email. Such analysis is not considered further in the current paper.

7.2 Document inclusion to determine an LCT

It is currently rather typical to send important documents (such as contracts) as attachments to emails. For this reason there is a large genre of documents where email may be of particular importance to determine the age of the documents.

Figure 3 illustrates the case where $x \sqsubset y \sqsubset z$ for artefacts x, y and z. In this example x may be some PDF standard, y may be a document that uses that PDF standard and z may be an email to which y has been attached. The two creation times in figure 3 represent the earliest creation time (ECT) and latest creation time (LCT) of document y: the ECT is represented by the curve from t_0 to t_4, while the LCT is represented by the curve from t_5 to t_9.

The curve in figure 3 that 'connects' the ECT to the LCT represents the window when y was created. It spans the graph from t_0 to t_9. The edges of the creation window graph has been drawn to coincide with the

left edge of the creation window of x and the right edge of the creation window of z; the graph has been offset slightly from those edges to show that there are two graphs following the same path. Note again that the graphs do not purport to present a probability density function that indicates the probability that y was created at some time t. (If the graphs contained in figure 3 were probability density functions, the area underneath all three graphs would have been equal to 1.) There is no basis to suggest that the edges of either the creation windows of x or z depict an accurate probability. In addition, given the specific scenario, the width of the creation window of x is expected to be significantly wider than the creation window of z.

The graph should be read as follows: There is a time period at the left of the creation window of y ($t_0 - t_2$, or earlier) where unusual circumstances would have made creation of y possible; similarly, there is a time period at the right of the creation window ($t_7 - t_9$, or later), where unusual conditions would have made the creation of y possible. The flat line at the top of the creation window ($t_2 - t_7$) indicates that this period was the 'most likely' period period during which y was created in the sense that no special conditions need to be assumed. The line is flat to indicate that the available evidence does not indicate any time in that period as a more likely point at which the document y was created compared to other times. However, this does not mean that the probability that the document was created at any time over that period is equal: Given a specific artefact the notion of probability is meaningless. Such a specific artefact was created at some specific time and probability cannot be used to calculate that time.

Often a question about the creation time of a document will not be about the exact time at which the document was created, but about the relative ordering of events. To return to the question about a last will and testament, the question may be whether that document was created before the individual died (or was incapacitated). If the creation time clearly precedes or follows the event of interest, the question can be answered without further consideration. When the creation window overlaps with the event of interest in one of the areas where the 'likelihood' that the document was created is not at a maximum, the question to be asked is again not about probabilities, but about whether the special conditions that would have enabled creation during that period were present. With knowledge about the applicability of special conditions, it may be possible to reach a qualified conclusion about the relative order of document creation and an event of interest.

Documents often depend on the creation of more than one other artefact, such as a file format and character code used. In such cases, the

artefact created later would generally be preferred, because that provides a better estimate of the earliest creation time of the document. Generally, the earliest possible LCT would be preferred. However, special circumstances may apply that may not be based on the latest ECT or earliest LCT, for example when greater confidence may be ascribed to an ECT or LCT that does not lead to the shortest possible creation window. Trading confidence for precision should always be considered.

8. Conclusion

The paper discussed techniques that may be used to determine the age of digital documents. To do this, it introduced two principles: the principle of inclusion and the principle of replication. For most documents the creation window thus determined (between various ECTs and LCTs) may be rather wide. However, such a wide window may be useful — in particular where the window excludes a purported creation time of a document.

Several issues related to confidence with which conclusions could be reached were raised. They include the number of document attributes that confirm a version, the certainty with which the age of any containing documents may be determined, the extent to which certain information is replicated, and, potentially, attributes of the author of a document according to litigant's versions. This would naturally be a narrative certainty, rather than a statistical certainty, but statistical certainties are unattainable in most branches of forensic science. The approach presented provides a much clearer structure to express certainty, than is typically possible in digital forensics. Mechanisms to succinctly express such certainty require further research, though.

The ability to date documents has been shown to be useful in the legal context. As born-digital documents replace other documents, the ability to determine their age would remain useful. Further research is necessary to determine the extent to which the approach presented in this paper would meet that need. It is, obviously, possible to determine dates more precisely by, for example, using metadata. Care should, however, be taken not to adopt precision at the cost of confidence.

References

[1] Adobe Systems Incorporated. *PDF Reference second edition — Adobe Portable Document Format Version 1.3*. Addison-Wesley, 2000.

[2] Adobe Systems Incorporated. *PDF Reference third edition — Adobe Portable Document Format Version 1.4*. Addison-Wesley, 2001.

[3] Adobe Systems Incorporated. *PDF Reference third edition — Adobe Portable Document Format Version 1.5*. Addison-Wesley, 2003.

[4] Adobe Systems Incorporated. *PDF Reference fifth edition — Adobe Portable Document Format Version 1.6*. Addison-Wesley, 2004.

[5] Adobe Systems Incorporated. *PDF Reference sixth edition — Adobe Portable Document Format Version 1.7*. Adobe, 2008.

[6] A. Aotes. Navigating the PDF/A standard: A case study of theses in the university of Oxford's institutional repository. Masters thesis, University of Illinois at Urbana-Champaign, 2018.

[7] American standard code for information interchange. Standard X3.4-1963, American Standards Association, 1963.

[8] A. Bahjat and J. Jones. File allocation chronology and its impact on digital forensics. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 612–018, Las Vegas, NV, USA, 2023. IEEE.

[9] A. A. Bahjat and J. Jones. Deleted file fragment dating by analysis of allocated neighbors. *Digital Investigation*, Volume 28(Supplement):S60–S67, Apr. 2019.

[10] R. Bangia. *Multimedia and Web Technology*. Firewall Media, 2004.

[11] T. Bienz and R. Cohn. *Portable Document Format Reference Manual*. Addison-Wesley, 1993.

[12] T. Bienz, R. Cohn, and J. R. Meehan. *Portable Document Format Reference Manual — Version 1.1*. Adobe Systems Incorporated, 1996.

[13] T. Bienz, R. Cohn, and J. R. Meehan. *Portable Document Format Reference Manual — Version 1.2*. Addison-Wesley, 1996.

[14] R. L. Brunelle and K. R. Crawford. *Advances in the Forensic Analysis and Dating of Writing Ink*. Charles C. Thomas, 2003.

[15] F. Buchholz and E. Spafford. On the role of file system metadata in digital forensics. *Digital Investigation*, 1(4):298–309, 2004.

[16] A. A. Cantu. A sketch of analytical methods for document dating. Part I. The static approach: determining age independent analytical profiles. *International Journal of Forensic Document Examiners*, 1(1):40–50, Jan/Feb 1995.

[17] A. A. Cantu. A sketch of analytical methods for document dating. Part II. the dynamic approach determining age dependent analytical profiles. *International Journal of Forensic Document Examiners*, 2(3):192–208, Jul/Sep 1996.

[18] E. Casey. *Digital Evidence and Computer Crime*. Elsevier, 2004.

[19] E. Casey. Digital stratigraphy: Contextual analysis of file system traces in forensic science. *Journal of Forensic Sciences*, 63(5):1383–1391, 2018.

[20] K. P. Chow, F. Y. W. Law, M. Y. K. Kwan, and P. K. Y. Lai. The rules of time on NTFS file system. In *Second International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE'07)*, pages 71–85, Bell Harbor, WA, USA, 2007. IEEE.

[21] F. Cohen. *Digital Forensic Evidence Examination*. Fred Cohen & Associates, fourth edition, 2013.

[22] X. Ding and H. Zou. Reliable time based forensics in NTFS. In *26th Annual Computer Security Applications Conference (ACSAC)*, Austin, Texas, USA, Dec. 2010. ACM.

[23] Google. Google Workspace Status Dashboard, Accessed on 14 Feb 2024. `https://www.google.com/appsstatus/dashboard/summary`.

[24] R. Harris. *Selling Hitler*. Faber & Faber, 1986.

[25] O. Hilton. *Scientific Examination of Questioned Documents, Revised Edition*. CRC-Press, 1992.

[26] S. M. Ho, D. Kao, and W.-Y. Wu. Following the breadcrumbs: Timestamp pattern identification for cloud forensics. *Digital Investigation*, 24:79–94, Mar. 2018.

[27] ISO 32000-2:2017 (PDF 2.0). International standard, ISO, 2017.

[28] ISO 32000-2:2020 (PDF 2.0). International standard, ISO, 2020.

[29] N. Kapoor, P. Sulke, R. K. Shukla, R. Kakad, P. Pardeshi, and A. Badiye. Forensic analytical approaches to the dating of documents: An overview. *Microchemical Journal*, 170:106722, 2021.

[30] R. Koen and M. S. Olivier. The use of file timestamps in digital forensics. In H. S. Venter, M. M. Eloff, J. H. P. Eloff, and L. Labuschagne, editors, *Proceedings of the ISSA 2008 Innovative Minds Conference*, Johannesburg, South Africa, July 2008. (Published electronically).

[31] A. Mohamed and C. Khalid. Detection of timestamps tampering in NTFS using machine learning. *Procedia Computer Science*, 160:778–784, 2019.

[32] M. S. Nvsa. Van der Merwe v Master of the High Court and another (605/09). Case [2010] ZASCA 99, Supreme Court of Appeals of South Africa, Sept. 2010.

[33] M. S. Olivier. Digital forensics and the big data deluge — some concerns based on Ramsey theory. In G. Peterson and S. Shenoi,

editors, *Advances in Digital Forensics XVI*, Advances in Information and Communication Technology, pages 3–23. Springer, 2020.

[34] A. S. Osborn. *Questioned Documents: A Study of Questioned Documents with an Outline of Methods by which the Facts May be Discovered and Shown*. Lawyers' Co-operative Publishing Company, 1910.

[35] D. Palmbach and F. Breitinger. Artifacts for detecting timestamp manipulation in NTFS on Windows and their reliability. *Forensic Science International: Digital Investigation*, 32:S1–S10, Apr. 2020.

[36] J. Peters. Gmail had two outages this week that delayed emails. *The Verge*, Dec. 2023. `https://www.theverge.com/2023/11/30/23982915/gmail-outage-issues-delaying-emails`.

[37] M. Peyravian, A. Roginsky, and A. Kshemkalyani. On probabilities of hash value matches. *Computers & Security*, 17(2):171–176, 1998.

[38] H. Pieterse, M. S. Olivier, and R. van Heerden. Playing hide-and-seek: Detecting the manipulation of Android timestamps. In H. S. Venter, M. Loock, M. Coetzee, M. M. Eloff, and S. Flowerday, editors, *Information Security for South Africa (ISSA)*. IEEE, Aug. 2015.

[39] R. Pike and K. Thompson. Hello World or Καλημέρα κοσμέ or こんにちは世界. In *USENIX Winter 1993 Conference Proceedings*, pages 43–50. Jan. 1993. San Diego, CA, USA.

[40] K. Rani and C. Sharma. Tampering detection of distributed databases using blockchain technology. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Noida, India, 2019. IEEE.

[41] A. Simmonds. Patel v Patel. Case [2017] EWHCA, 133 (Ch), England & Wales High Court (Chancery Division), 2017.

[42] M. Smith. Yashwant Dahyabhai Patel v Girish Dahyabhai Patel and Others. Case [2017] EWHC 1588 (Ch), England & Wales High Court (Chancery Division), 2017.

[43] D. H. R. Spennemann and R. J. Spennemann. Establishing genealogies of born digital content: The suitability of revision identifier (RSID) numbers in MS Word for forensic enquiry. *Publications*, 11(3), Sept. 2023.