ETSI TS 103 643 V1.2.1 (2022-01)

Techniques for assurance of digital material used in legal proceedings

Reference
RTS/CYBER-0079

Keywords
information assurance

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Specification (TS) has been produced by ETSI Technical Committee Cyber Security (CYBER).

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The present document defines a process of receiving, transforming and outputting material that can be assured digitally. The process is called the "Digital Evidence Bag" (DEB). The present document identifies the ways that a DEB can be used to provide assurance of material used in legal proceedings. Specifically, the assurance of the material is not dependent on the process having been carried out by a qualified or trained human expert.

The present document is designed to be used in situations where a risk assessment of the handling of digital material has identified that extra assurance of the integrity, provenance, continuity and validity of the digital data is required.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at https://docbox.etsi.org/Reference/.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

[1] IETF RFC 4122: "A Universally Unique IDentifier (UUID) URN Namespace".

[2] ETSI TS 103 307: "CYBER; Security Aspects for LI and RD Interfaces".

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

[i.1] ISO/IEC 17025: "General requirements for the competence of testing and calibration laboratories".

[i.2] Lives and Opinions of Eminent Philosophers, Diogenes Laërtius (c. 225 CE).

[i.3] Navigation and Nautical Astronomy, James Inman (1835).

[i.4] ISO 8601: "Date and time -- Representations for information interchange".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

case-specific input material: input material for a Digital Evidence Bag that is specific to the particular investigation or case

Digital Evidence Bag (DEB): process of storing digital evidence which can be assured digitally

Purely Digital Transformation (PDT): transformation in which a repeatable, deterministic, pre-specified, fail-safe, well-defined digital function is performed on entirely digital data

NOTE: See clause 5.3.1 for more information.

reference input material: relevant material (if any) which is used to support the case-specific input material by adding context or background

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

ADT	Assured Digital Transformation
B-DEB	Basic Digital Evidence Bag
DEB	Digital Evidence Bag
DEB+H	Digital Evidence Bag with Hashing
DEB+HIA	Digital Evidence Bag with Hashing and Input Assurance
DEB+IA	Digital Evidence Bag with Input Assurance
DIPCV	Data Integrity, Provenance, Continuity and Validity
GDPR	General Data Protection Regulation
LED	Law Enforcement Directive
PDT	Purely Digital Transformation

4 Basic principles

4.1 Summary

The present document gives a definition for a "Digital Evidence Bag", which is a process for storing and transforming digital material. Annex A provides an informative description of when this process is intended to be used.

The present document defines and specifies requirements for the following types of Digital Evidence Bag:

1) A Basic Digital Evidence Bag (B-DEB) (see clause 5).

2) A Digital Evidence Bag with Hashing (DEB+H) (see clause 6).

3) A Digital Evidence Bag with Input Assurance (DEB+IA) (see clause 6).

4) A Digital Evidence Bag with Hashing and Input Assurance (DEB+HIA) (see clause 6).

Annex F provides recommendations for testing a DEB.

NOTE: A Digital Evidence Bag with Digital Signature would also meet many of the same goals as the Digital Evidence Bag with Hashing and Input Assurance and this is being considered for a future version.

5 Definition of a Basic Digital Evidence Bag

5.1 Reference model

The model for a Basic Digital Evidence Bag is as shown in Figure 1.

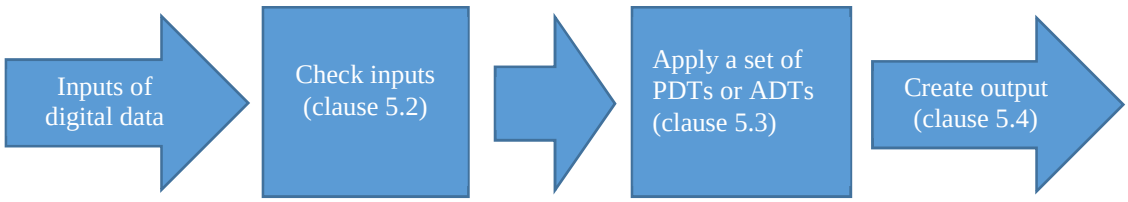

Figure 1: Model for Basic Digital Evidence Bag

5.2 Inputs of digital data

5.2.1 Nature of inputs

There are two types of input material: case-specific input material and reference input material (as defined in clause 3).

EXAMPLE: Examples of reference input material are maps or publicly available reference data.

Basic DEBs shall follow the specifications for input material as listed in clauses 5.2.2 to 5.2.4.

5.2.2 A unique identifier for each input

5.2.2.1 Identifiers for case-specific input material

For a Basic DEB, each input of case-specific input material (see clause 3) shall have a unique identifier attached to it. One of the two following approaches shall be used:

1) The identifier shall consist of:

   a) an identifier supplied by the originating organization; and

   b) a unique identifier for the originating organization. A globally-unique identifier shall be created for the originating organization, using a combination of a nationally-unique identifier together with a country code.

2) The identifier shall be a randomly chosen globally unique identifier as defined in IETF RFC 4122 [1].

Each piece of case-specific input material should include where relevant an identifier of a request that prompted the generation of the input.

5.2.2.2 Identifiers for reference input material

For a Basic DEB, the reference input material (see clause 3) should also have an identifier to make it clear where it came from, and should also identify the time it was collected if that is significantly different from the time the material is being submitted to the DEB.

5.2.3 Time and location information for input material

The time information in a Basic DEB shall consist of the following:

- All time information as supplied by the originating organization. The input material should contain time and date information, including indication of the time zone, for the point at which the data was generated or created (or for the period over which the data was generated).

- DEB Entry Time: A timestamp shall be added to indicate the time and date, including indication of the time zone, the data was received at the Digital Evidence Bag.

In the case that the time of creation is clearly indicated by the originating organization, it shall be checked that the DEB Entry Time is after the time of creation of the material.

The location of collection of information should be included where the point of collection is not necessarily fixed to one place and is relevant to the value of the material collected.

EXAMPLE: A contract has been placed with a laboratory to provide information, and the contract includes a statement of the formats in which the data will be provided.

5.2.4 Format of input material

The format for each input file to a Basic DEB should be known or clear (i.e. known via a communication in advance of sending the data, or clear from the evidence file itself). Each input file should be checked syntactically for data formats where there are suitable automated checks.

EXAMPLE: If data is submitted in XML and the XML schema is known and agreed, then each input file is checked against the schema.

5.3 Applying Purely Digital Transformations and Assured Digital Transformations

5.3.1 Definition of a Purely Digital Transformation

A Purely Digital Transformation (PDT) is one in which a repeatable, deterministic, pre-specified, fail-safe, well-defined digital function is performed on entirely digital data.

Specifically:

- It is repeatable in that if the step is performed again by a different computer or operator, in a different environment, in a different country or at a different time, the outcome is always the same.
- It is deterministic in that the same inputs to the process always give the same outputs, which is not dependent on the training or skill level of an operator.
- It is pre-specified in that the full details of the process are known to all relevant parties in advance and (ideally but not essentially) the details are published.
- It is fail-safe in that its failure modes are easily distinguishable from successful outcomes (in particular, that a failure mode looks very different from a successful output with no records in it).
- It is well-defined in that the version numbering is present and accurate and that the formatting is clear and specified in all places.
- It is digital in that its input and output are digital.

5.3.2 Use of PDT in a Digital Evidence Bag

Within the Digital Evidence Bag, one or more PDTs (as defined in clause 5.3.1) may be applied.

For each transformation, the DEB shall check:

- That the formatting and definition of input files is clear.
- That any standards referred to have a correct version number and are designed for the purpose in question.
- That the input(s) each has an identifier for the material in question.

The DEB shall record:

- The time and date that the transformation took place, ensuring time zone is clear.
- A unique identifier to the output.
- An identification of the process that took place and an identifier to the entity that performed it.

The recommendations in Annex E should be followed when handling the types of data listed in Annex E. Examples of PDT are given in Annex D.

5.3.3 Definition of an Assured Digital Transformation

An Assured Digital Transformation (ADT) is one in which a documented, fail-safe task is performed on entirely digital data by a suitably qualified person.

Specifically:

- The task is documented, in that there is a document listing how the task is to be performed, which was available to the person performing the task and is available to be referenced during legal proceedings. The document lists the inputs required and, where appropriate, the format of each input. The areas that require human judgement or skill are explicitly stated in the documentation. If the task needs suitable facilities, this is stated in the documentation. If the task needs suitable software, this is stated in the documentation.

NOTE 1: The present document does not define what suitable facilities or suitable software are.

NOTE 2: The present document records what was required and what was used so that the court can decide whether it was suitable, perhaps by referring to other standards in this area e.g. ISO/IEC 17025 [i.1].

- The task is fail-safe, in that the documentation lists the common failure modes, showing how failure modes are easily distinguishable from successful outcomes (in particular, that a failure mode looks very different from a successful output with no records in it).
- The task is digital in that its input and output are digital.
- The person is qualified for the task, in that they have suitable recorded experience or qualifications to cover the areas of the task that have been identified as requiring human judgement or skill.

NOTE 3: The present document does not specify what qualifications would count as suitable.

NOTE 4: The purpose of the present document is to define how to create a record of the task that was performed, including details of the person who performed the task including their qualifications. The content of the record can allow a court to establish whether it considered the qualification suitable for the task, perhaps by referring to other standards in this area.

5.3.4 Use of ADT in a Digital Evidence Bag

Within the Digital Evidence Bag, one or more ADTs (as defined in clause 5.3.3) may be applied.

For each transformation, the DEB shall:

- Check formatting of input files is clear (in line with the documentation).
- Check that the input(s) each has an identifier.

The DEB shall record:

- The time and date that the transformation took place, ensuring time zone is clear.
- A unique identifier for the output.
- An identification of the process that took place, giving a link to the documentation that describes the required human judgement, facilities and software (described in clause 5.3.3).
- A unique way to identify the person who performed it. Where possible, this should be done using an identifier for an organization (e.g. the person's employer) together with a unique ID within that organization. This requirement may also be met using a country code plus a unique identifier for a person within a country.
- The qualification of the person who performed, as described in clause 5.3.3. No unrelated qualifications shall be added.
- The location of performing the task, if the task needed suitable facilities (as stated in the documentation).

- The software used with unique name and version number, if the task needed suitable software (as stated in the documentation).

5.4 Details for creating the output

The Basic DEB output file shall contain the following information:

- List of all input files, including identifiers (as defined in clause 5.2.2) and the DEB entry time (clause 5.2.3).
- For each transformation that was applied, a list of the details for that transformation from clause 5.3.2 (for PDTs) or 5.3.4 (for ADTs).
- The software name and version that was used.

NOTE: There can also be requirements for the material in the Digital Evidence Bag to be deleted in a complete and assured way. These requirements are out of scope of the present document, though it is noted that a number of the techniques in the present document (e.g. list of all processes that have been applied, identification of inputs and outputs of each stage) can help to demonstrate a list of material which needs to be deleted.

6 Definition of other Digital Evidence Bags

6.1 Introduction

Clause 6 specifies the following types of Digital Evidence Bag:

- DEB+H (with Hashing).
- DEB+IA (with Input Assurance).
- DEB+HIA (with Hashing and Input Assurance).

NOTE: A Digital Evidence Bag with Digital Signature would also meet many of the same goals as the Digital Evidence Bag with Hashing and Input Assurance and this is being considered for a future version.

Each of the definitions builds on the definition of a Basic DEB in clause.

Clause A.3 explains when it can be appropriate to choose each of the different types of DEB.

6.2 Definition of a DEB+H

6.2.1 General

A Digital Evidence Bag with Hashing (DEB+H) shall meet the specification of a Basic DEB (clause 5). In addition, it shall use hashing as specified in clause 6.2.2.

NOTE: A Digital Evidence Bag with Hashing would typically be used in situations where assurance was required that material had not been changed from the point at which it was submitted to the DEB (and potentially earlier than this, depending on when the hash was taken) through to the point it was used in court. See clause A.3 for more information.

6.2.2 Use of hashing in a DEB+H

A DEB+H shall create a single input file with all input material. The input material shall include all identifiers as defined in clause 5.2.2 and the DEB entry time as defined in clause 5.2.3.

EXAMPLE: A number of pieces of input material could be turned into a single file through use of zip.

The single input file shall be hashed using the algorithms from ETSI TS 103 307 [2], clause A.3.4.

NOTE 1: This requires the use of two (or more) algorithms, allowing for new algorithms to be added and older ones to be removed without dependence on a single algorithm (it facilitates crypto-agility).

The hash shall be stored in a manner which provides assurance that it cannot be changed over time. The underlying goal is that the assurance is strong enough to satisfy the purposes identified in the risk assessment (see clause A.3). Assurance shall be provided using one or both of the following mechanisms:

1) Through examining the physical, cryptographic or procedural controls that are in place at the storage site (e.g. examples are given for assuring storage at the Communications Service Providers in ETSI TS 103 307 [2]).

2) Through storage of hashes at a dedicated function which uses publication of material (along with clear cryptographic controls) to demonstrate that hashes are unchanged. A function as specified in Annex G may be used to perform the hash assurance role.

When assurance is required, it is sufficient to check that the hash of the input material matches the hash that was stored and assured as described above. Further examples are given on this point in ETSI TS 103 307 [2] in clause A.3.

NOTE 2: There are advantages to assuring that a hash is unchanged (rather than demonstrating that the material itself is unchanged): the material can contain details that need to be kept confidential i.e. this is a technique that assists in the preservation of privacy.

Hashes may also be created at other stages of the process. If used, such hashes shall be in addition to and completely independent of the hash described in this clause. They shall follow the same processes as the hash of the single input file (i.e. shall use the algorithms described in this clause, shall be stored as described in this clause and may be checked as described in this clause).

If ADTs are used in the DEB, then it is noted that the processing inside the DEB is not necessarily completely deterministic and repeatable (as it contains an element of human judgement). Therefore consideration should be given to creating an additional hash covering the information in the DEB output file as defined in clause 5.4.

6.3 Definition of a DEB+IA

A DEB+IA shall follow the specification in clause 5 for the definition of a Basic DEB.

Additionally, the DEB+IA shall identify the delivery protocol or technique which was used to input the material to the DEB. The DEB+IA shall log any credentials that were provided or were necessary in order to use the specified delivery protocol or technique. The output file shall contain the credentials used.

EXAMPLE 1: If the material was delivered over a secured private network, then there is assurance that the material originated from someone who had access to the secured private network. The credentials would be the log-on details (e.g. username and time) that were used to access the network.

EXAMPLE 2: If material is delivered to a specific physical or network location (e.g. a specific folder) then there is assurance that the material originated from someone who had permission to access that location. The credentials would be the log-on details that were given which enabled the user to access the location.

NOTE: Situations would need to be assessed on a case-by-case basis to see whether the protocol was appropriate for meeting the risks identified (see clause A.3).

6.4 Definition of a DEB+HIA

A DEB+HIA shall follow the specifications in both clauses 6.2 and 6.3 in addition to the specification of a Basic DEB in clause 5.

Clause A.3 explains when it can be appropriate to each of the different types of DEB.

Annex A (informative): Context

A.1 Purpose of the present document

The present document considers the assurance of handling and processing of material which is potentially to be used in legal proceedings (often referred to as assurance of evidence in court). Material goes through various steps prior to its presentation in legal proceedings; this is sometimes called the chain of evidence.

The present document notes that, in some situations, it is possible and beneficial to provide digital assurance (i.e. using digital and/or cryptographic techniques) to a set of processes or transformations of the material. The present document provides a specification for a process (called a Digital Evidence Bag or DEB) which can be used for handling digital material.

The processes outlined in the present document do not (in general) provide assurances about whether the material was correct or accurate at the point it was originally collected or created. The goal of the present document is to provide assurance of the subsequent processes of handling, storing, transforming and presenting the material.

A.2 Void

A.3 Choosing which type of DEB to use

It is not essential to use a DEB for all digital evidence. The purpose of the present document is to provide a set of tools (DEB or DEB+H,or DEB+HIA) which are chosen where additional assurance is deemed to be appropriate for digital material. In determining whether to use a DEB (or which type of DEB to use), a risk assessment should be made. The risk assessment would determine and analyse the potential challenges to the material when it is presented in court, and to highlight those which are likely to result in a significant and realistic challenge. The risk assessment should consider those aspects of Data Integrity, Continuity, Provenance and Validity as described in Annex C. Once the risks are understood, it can be seen whether the DEB (or DEB+H or DEB+HIA) would be able to mitigate the risks, which determines whether (or which) DEB it is appropriate to use.

A.4 Void

Annex B (informative): Examples

B.1 Introduction

This clause contains examples which illustrate concepts from the main body of the present document. These are illustrations but do not change or supersede the normative text which they are supporting.

B.2 Examples of transformations which are not PDT

This clause contains details to illustrate the definition of a Purely Digital Transformation (see clause 5.3.1).

The following characteristics would demonstrate that a transformation did not meet the definition in clause 5.3.1 i.e. that the transformation is not purely digital:

- Something which requires human skill or judgement.
- Something in which an untrained user could create apparently correct answers which differed from those that a trained user would create:
  - It is acceptable for a PDT to need a basic level of training e.g. which button to press to start the process.
- Something which requires calibration or expert maintenance in order to avoid errors:
  - It is acceptable for purely digital techniques to need updates e.g. refresh of algorithms.
- A transformation which uses an analogue measurement as part of the transformation.
- A transformation where the same inputs could get a different output depending on external factors e.g. the weather:
  - It is acceptable for a PDT to report an error or fail to complete in some situations e.g. a power cut, provided these situations do not result in an incorrect but apparently successful output.

B.3 Examples regarding accuracy and completeness of input material

The accuracy and completeness of the input data is not part of the normative requirements of the present document.

The following examples can be helpful to aid consideration of this issue:

- In some cases, material originates from data captured by or measurements taken by members of law enforcement e.g. at a crime scene. In these cases there is control over the chain of evidence from the point the data is created i.e. through the chain of evidence it is straightforward to be aware of the training, expertise and oversight of those involved.
- In other cases, the data comes from a third-party system e.g. call records retained by a Communications Service Provider. In these cases, it would be possible to examine existing assurances about the accuracy and completeness of the data. For example, if the data came from billing systems, then it is likely that there are existing standards and processes which can give assurances about the accuracy and completeness of billing records. These can be useful in providing assurance about the material.
- In many examples there will be theoretical risks regarding the accuracy and completeness of the input material. These are considered in light of the risk assessment outlined in clause A.3, noting that it is not necessary to discount or ignore material because of an unlikely or theoretical risk that it is incorrect.

B.4 Example of linking to physical evidence

It is possible to make a link from physical evidence to digital identifiers through the use of sprays/gels/liquids containing a substance which is uniquely linkable to a specific identifier e.g. a binary number. An over-simplified example would be a paint which consisted of up to 8 specific colours - the presence or absence of each colour would give a 0 or 1 in a given position. The paint could be applied to a particular physical object, and therefore that object would be associated with that binary number.

Such an identification scheme is potentially useful in situations where a link is desired from material in the Digital Evidence Bag to physical evidence.

Annex C (informative): Data Integrity, Provenance, Continuity and Validity

C.1 Introduction

The risk assessment (clause A.3) considers assurance of Data Integrity, Provenance, Continuity and Validity (DIPCV) and other concerns as appropriate. This annex lists those aspects of DIPCV which are most relevant to Digital Evidence Bag process.

C.2 Integrity

The Digital Evidence Bag is relevant to data integrity in terms of providing assurance that the output material has only been changed in accordance with the transformations that are specified, and that the input material has not been altered since it was input.

C.3 Provenance

The Digital Evidence Bag is relevant to data provenance in terms of providing assurance that the origin of the input material was logged and any credentials provided were noted (along with how those credentials were checked), and a check that the input material was correctly formatted and identified. It provides confidence that any filtering, searching or matching that took place was performed with a consideration of common misunderstandings (see Annex E), and was performed in line with well-specified practices.

The Digital Evidence Bag does not provide assurance about whether the input data was accurate when it was input (see clause B.3).

C.4 Continuity

The Digital Evidence Bag is relevant to data continuity in terms of providing assurance that the material was changed only in line with the transformations that are listed in the output meta-data i.e. while the material was in the Digital Evidence Bag there is a clear, assured set of changes that took place. The present document can provide assurance that the output is repeatable and deterministic (see clause 5.3.1), without the requirement for human expertise or training.

EXAMPLE: Material in the Digital Evidence Bag is labelled with unique identifiers to demonstrate the path of the material.

C.5 Validity

The Digital Evidence Bag is relevant to data validity in terms of providing assurance that it is possible to check that the transformations that were applied to the data were fit-for-purpose. It means that the input file was checked to have been in accordance with a specified format. It reduces the possibilities for misunderstandings or errors took place in handling the data.

The present document does not check that the input data was "correct" in any sense, which can be dependent on the expertise of the person who created or requested it.

C.6 Other considerations

It is also important to meet, where appropriate, data protection legislation and considerations, though the details of these are out of scope of the present document. For example, some legislations (such as under the LED and the GDPR) contain requirements for the confidentiality of the information. The Opinion of the European Data Protection Board is noted to be relevant.

Annex D (informative): Examples of functions for performing purely digital transformations

D.1 Introduction

This annex gives examples of PDT functions which can be used in a DEB provided they are implemented in a way which meets the criteria in clause 5 They are intended to be helpful illustrations but this is not a guarantee that these functions are always Purely Digital Transformations.

D.2 Finding items in common between two or more lists

- Given two lists of parameters, create a list of parameters which are present in both the lists. Given more than two lists of parameters, create a list of parameters which are present in all lists (or in a specified number of them).
- Given a list of entries in a database, find all entries which match on a given parameter.

D.3 Filtering of a list of items based on a criterion

Given a list of items, remove all those that do not meet the specified criterion.

A tolerance can be specified i.e. specify whether elements will be accepted if they are within a given tolerance of the criterion.

EXAMPLE 1: A time range has an extra tolerance of 2 seconds in case of incorrect clocks (note that this example is not appropriate in all situations).

EXAMPLE 2: This can also include cropping of a photo or extracting a section of audio or video material. If human experience was used to determine which portion was relevant (i.e. that other parts were not pertinent) then that stage of the process would be categorized as an ADT (see clause 5.3.3). However, the function of removing part of the data on the basis of a specified filter (e.g. "keep only the first 20 seconds") is a Purely Digital Transformation (PDT).

D.4 Adding additional data from reference material

- The input consists of the starting data file, together with the reference input material which is being used to add additional information to the starting data file.
- An identifier is picked which is used in the starting data file.
- Then, for each record in the starting data file, the identifier for that record is searched for in the reference input material. If the identifier is found, the data from the corresponding record in the reference input material can be added to the record in the starting data file.

D.5 Presentation of material

Overlay a set of results on a map, provided each result has a single geographical location.

If results are present that do not have a location, they can be logged in a way which makes it clear that the map is only complete when considered with the additional records. A location is not be assumed or guessed.

D.6 Change of formatting or codec

Change the format of a file provided the criteria from clause 5 are met.

EXAMPLE: Changing the codec for audio or video files.

Annex E (informative): Considerations when handling certain data types

E.1 Introduction

This annex lists considerations for handling transformations of different types of data. These are not intended to be prescriptive, and in many cases they will not be appropriate or relevant for the circumstances. For each data type, these items should be considered by the creator of a DEB when handling material of that data type.

E.2 Phone numbers

When comparing phone numbers, the creator of a DEB should:

- Consider how international prefixes will be handled (specifically, consider whether two phone numbers will be deemed the same if they are identical except that one has no country code and the other has the country code in which the search is taking place).

E.3 Names

When comparing names, the creator of a DEB should consider:

- Whether titles, prefixes ("de", "de la", "van") or suffixes ("Esq.") are being considered.
- Whether middle names or initials are considered.
- Whether names should be handled as case-insensitive (in general they should).
- Whether spaces or hyphens can be ignored.

E.4 Addresses

When comparing postal addresses:

- The DEB creator should consider that many combinations of abbreviations are possible (St. or Street. Rd. or Road) and many different orderings of components are possible.
- The DEB creator should bear in mind that it is usually clearest to start from postcodes or zip codes. They are not usually case sensitive. The DEB creator should take account of how whitespace should be handled in postal or zip codes. Failure to match on postal codes should be taken to indicate that the items are different addresses.
- Within a given postal code, the DEB creator should where possible try to match on a numerical parameter; they should bear in mind the possibility of flat numbers or other sub-components.

When handling email addresses, the DEB creator should be aware that they are not case sensitive. The DEB creator should consider whether it is sufficient to match only on the domain.

E.5 Locations

When comparing locations, the DEB creator should identify and follow an appropriate standard if there is a need to transform Latitude/Longitude to/from grid references.

Where appropriate, the DEB creator should be aware that there are functions which can be used to find a distance between points e.g. Pythagoras [i.2] for grid references or Haversines [i.3] for Latitude/Longitude.

E.6 Dates and times

The formatting of a date should be considered by the DEB creator. In particular the DEB creator should consider whether days are written before months (DD/MM) or months before days (MM/DD). Particular care is needed if some of the data has come from a country in which a different default applies (e.g. USA compared to UK). A simple way to ensure dates are clear is to use standard ISO 8601 [i.4].

The formatting of times should be considered by the DEB creator There is less scope for confusion in the ordering of elements (i.e. it would normally be reasonable to assume HH then MM then SS). Timezones should be specified and taken into consideration. If timezones are not specified, consideration should be given as to whether it can be assumed that local time applies, with care being taken around the point where clocks are changed (start or end of Daylight Saving).

Where relevant, consideration should be given to the accuracy of the clocks involved. Consideration should be given to extending a search window by a given amount of time in order to take account of inaccuracies in clocks (of course this is not always appropriate e.g. if this would include a disproportionate amount of extra results).

E.7 Identifiers

The DEB creator should consider whether identifiers are unique within a given scope or context. The context should be understood (e.g. unique within a country). If identifiers are being used outside of their context then the appropriate additional information should be considered (e.g. add a country code).

E.8 Text in general

The DEB creator should consider whether text is case sensitive.

EXAMPLE: Names, email addresses, addresses are not be considered case sensitive but passwords are in general case sensitive.

The DEB creator should: consider whether spaces are relevant (e.g. postcodes); consider other whitespace characters (carriage returns); consider handling of foreign character sets.

Annex F (informative): Testing a DEB

F.1 Introduction

This annex provides a recommendation for how DEB software should be tested. This annex applies to DEBs using PDTs only.

Software which claims to be a DEB should be tested in two ways:

- A conformance statement should be produced when the software is complete (see clause F.2).
- If material is challenged in legal proceedings, the procedure in clause F.3 should be followed.

F.2 Conformance statement

The producers of software claiming to be a DEB should produce a conformance statement once the given version of the software is complete.

It should state:

- Version number of software and date of production.
- Input file types and formats, how dates and identifiers are logged.
- Any PDTs used inside the DEB including their version numbers.
- Schemas for any internal data formats (i.e. after some but not all PDTs have been applied).
- Schema for the output file.

The conformance statement should list a set of test files that were used to check the software, including edge cases, "misuse" cases (where incorrect procedures were followed), empty cases and overflow conditions.

F.3 Checking when challenged in legal proceedings

If material is challenged during legal proceedings, some or all of the following items may be followed:

- Examine the conformance statement for the version of the software used: it should align with the details in clause F.2.
- Check the output and input files match the schema required.
- If necessary, re-run the processing within the DEB.
- Also check the hashing / signatures as appropriate.

Annex G (normative): Hash assurance function

G.1 Requirements

G.1.1 Functional requirements

The hash assurance function shall be able to receive a hash associated with an identifier (called an exhibit number). It shall store the hash and associate the exhibit number with the hash. The hash assurance function shall store the approximate (within 30 seconds) time of submission of the hash.

NOTE: The hash assurance function would help provide assurance about other meta-data e.g. start/end time of data gathering, location, identity of submitter (see clause 5.2) because those details are included within the information that was hashed. However, the hash assurance function would not explicitly be aware of any meta-data beyond the exhibit number and time.

The hash assurance function shall be set up:

- either to allow the submission of material only by authorized users; or
- so that anyone can submit a hash, though it shall be checked that this did not provide a route for denial-of-service attacks, and a technique shall be implemented to prevent clashes in exhibit numbering.

EXAMPLE: A member of the public wishes to demonstrate that a recording they have made was created and submitted contemporaneously with the event itself, which would provide some assurance that it had not been subject to extensive processing or manipulation post-event. The member of the public hashes their recording and submits the hash to the hash assurance function immediately.

If the hash assurance function is given an exhibit number, it shall be able to provide the hash and approximate time associated with the exhibit number (or respond that it does not exist within the store). If the hash assurance function is given a hash, it shall be able to give the approximate time associated with the hash or respond that it does not exist in the store.

G.1.2 Non-functional requirements

The hash assurance function shall be able to demonstrate hashes are unchanged in a way which meets the following criteria:

- It shall be resistant to a denial of service attack (flooding the system).
- It shall survive all of the following: any single company going bankrupt, any particular piece of source code being found to have a bug in it, a power outage at a specific facility.
- It shall survive a long-term transition to newer/longer hash functions.
- It shall survive a single individual being corrupt in the hash storage/management process i.e. any single person at the hash assurance function shall not be able to spoil or remove assurance from a piece of submitted data. An individual shall not be able to prevent a hash being submitted (or if they do, it is visible to the submitter within 2 weeks).
- It shall be possible for any member of the public (who can use hashes and who can follow the present document) to be able to check the working of the hash store and prove that the hashes have been followed correctly without needing to purchase any specific piece of software.
- It shall survive over a period of at least 10 years.
- It shall prevent situations in which exhibit numbers are duplicated in a way which could damage the integrity of the system. As a minimum it shall prevent exhibit numbers being repeated within any two-week period.

G.2 Example of a hash assurance function (informative)

G.2.1 Introduction

The following details are intended to provide an illustration of techniques which could be considered for meeting the requirements in clause G.1. It is not guaranteed that the function described in clause G.2 meets all the requirements in clause G.1 as many implementation details would need to be checked and added.

It consists of three components:

- Primary - for receiving data and responding to queries.
- Secondary - for checking (submissions and calculations) and as a back-up query function.
- Tertiary - for physical publication of data.

G.2.2 Specification of primary functionality

PRIMARY Hash store functionality - receiving data:

- Receive a hash over an interface. The interface requires a username and password which are managed in accordance with basic principles (each username has contact details known by the hash assurance function). *Purpose:* This is to prevent a Denial-of-Service attack and to give a way to inform the submitter if there has been a problem, but the identity of the submitter is not recorded or assured.
- The interface supports giving immediate feedback that the record has been accepted by the application. The hash store checks that the submitted time (on the material) matches (e.g. within 1 minute) the time of the hash store's clock. *Purpose:* this is to prevent clocks getting out of synch, or to prevent claims that clocks are out of synch.
- The hash store checks that the exhibit number has not been used within the previous 7 days. *Purpose:* this is to prevent people creating lots of different pieces of "evidence" with an identical identifier, and to prevent people claiming that lots of different versions of the "same" piece of evidence were submitted.
- The hash store creates a list of all hashes submitted in the most recent "chunk" (a chunk covers 1 minute). It adds them to a single list in the format "Start-time-of-chunk; hash-of-previous-chunk; identifier1-hash1; identifier2-hash 2; …; identifierN-hash N; End-time-of-chunk". This is published on its website, along with (separately) a hash of this chunk. The website publishes the primary store's system clock.
- At the end of each week, a weekly summary is created in the format "Start-time-of-week; chunk1hash; chunk2hash; …; chunk10800hash; End-time-of-week" and this is hashed to create the "end of week hash". This is put on the website, marked as pending.
- The primary hash store has a subscription to the publication (see below) and each week will change "pending" status to "published" once the hash appears in the publication.

PRIMARY Hash store functionality - querying tool:

- The hash store has a query functionality which - if given a hash - reports if it is present. It returns a yes or a no, together with the time (i.e. which minute-long chunk) and the associated submitted exhibit number.
- The hash store has a query function which - if given an exhibit number - reports the hash and approximate timestamp.

G.2.3 Specification of secondary functionality

SECONDARY Hash store functionality - part 1: submission checking:

- Receive a hash over a trusted interface. The interface asks for a username and password which are managed in accordance with basic principles. When the username is created, the hash store records contact details for someone who can be contacted in the case of a problem. The interface supports giving immediate feedback that the record has been accepted by the application.
- The hash stores a list of all hashes submitted. It checks that they are present on the primary system and that the exhibit number matches and that the times are consistent. If they are not, every day it reminds the submitter to re-submit.

SECONDARY Hash store functionality - part 2: calculation checking:

- This hash store takes a copy of the primary store's list of submitted hashes every minute and also publishes this on its website.
- This hash store checks its own clock matches the primary's clock (within 30 seconds).
- The hash store independently re-creates each chunk, and independently calculates the hashes, and publishes this on its website if it agrees.
- The hash store independently creates an "end of week" hash and publishes this on its web site.
- The secondary function would also receive a copy of the publication (see Tertiary) and would check the published weekly hash.

SECONDARY Hash store functionality - part 3: backup query function:

- This function is identical to the "querying tool" component of the primary hash store.

The primary function is built and run independently from the secondary functions (built separately using totally independent source code, different OS and compiler/language, separate implementations of off-the-shelf algorithms, hosted in a different physical building and by a different domain provider). Both primary and secondary functions are run with appropriate back-ups. They have a locked-down interface to the internet, with only the above very limited functions being able to be accessed.

G.2.4 Specification of tertiary functionality

TERTIARY - Physical publication company:

- Each week: The tertiary hash store visits the primary hash store's web site to identify the "end of week hash". The tertiary hash store visits the secondary hash store's web site to identify its "end of week" hash (see "calculation checking". If the two end-of-week hashes are identical, then this end-of-week hash is put into the tertiary hash store's next publication.
- The publication has a print-run of at least 50 physical copies which is available and stored historically in at least 5 public places in the country (e.g. major libraries).

The publication is sent to the PRIMARY and SECONDARY sites.

G.2.5 Use cases

The police or person submitting a hash:

- Makes a file which contains the evidence itself and has the information from clause 5 clearly present inside the file.
- Takes a hash of the file (or more if more than one is required - see below).
- Submits the hash to the primary store using an exhibit number and adds the time from his/her own system.

- Submits the same information to the Secondary function ("submission checking function"). They watch for communications from the secondary function (i.e. on the address they gave as their contact details) to be able to follow-up on any problems e.g. to re-submit if a problem is reported.
- Can zip the files together if there is a large number of files being created in one place at the same time.

If material needs to be checked:

- It is possible to check via the primary system web site that the hash is present in the store and that it was submitted at the time stated.
- It is possible to check this on the secondary web site also.
- Members of the public can check (using their own tools) that hashes are present on the primary and secondary sites.
- Members of the public or legal experts can check (using their own tools) that the hash chain is complete through each week and up to a value which is published in a publication e.g. at a public library. There is no need for any proprietary algorithms or software, beyond a basic hash function.

EXAMPLE Regarding sizing concerns, if an organization had 100 000 members who each wished to submit a hash every 5 minutes, this is approximately 1 billion records per year. It is reassuring to note that the chance of a collision (two identical 256-bit hashes) in 1 billion records is approximately 1 in 10^60.

Annex H (informative): Change History

Date	Version	Information about changes
November 2021	1.1.2	Implementation of CR1 (CYBER(21)027009r1) to introduce the new concept of Assured Digital Transformation
January 2022	1.2.1	Publication

History

Document history		
V1.1.1	January 2020	Publication
V1.2.1	January 2022	Publication