TECHNICAL REPORT

ISO/IEC TR 21000-11

First edition
2004-11-15

Information technology — Multimedia framework (MPEG-21) —

Part 11: Evaluation Tools for Persistent Association Technologies

Technologies de l'information — Cadre multimédia (MPEG-21) —

Partie 11: Outils d'évaluation relatifs aux technologies d'association persistante

Reference number
ISO/IEC TR 21000-11:2004(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

In exceptional circumstances, the joint technical committee may propose the publication of a Technical Report of one of the following types:

— type 1, when the required support cannot be obtained for the publication of an International Standard, despite repeated efforts;

— type 2, when the subject is still under technical development or where for any other reason there is the future but not immediate possibility of an agreement on an International Standard;

— type 3, when the joint technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example).

Technical Reports of types 1 and 2 are subject to review within three years of publication, to decide whether they can be transformed into International Standards. Technical Reports of type 3 do not necessarily have to be reviewed until the data they provide are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC TR 21000-11, which is a Technical Report of type 3, was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

ISO/IEC TR 21000 consists of the following parts, under the general title *Information technology — Multimedia framework (MPEG-21)*:

— *Part 1: Vision, Technologies and Strategy*

— *Part 2: Digital Item Declaration*

— *Part 3: Digital Item Identification*

— *Part 5: Rights Expression Language*

— *Part 6: Rights Data Dictionary*

— *Part 7: Digital Item Adaptation*

— *Part 8: Reference Software*

— *Part 9: File Format*

— *Part 11: Evaluation Tools for Persistent Association Technologies*

The following parts are under preparation:

— *Part 10: Digital Item Processing*

Introduction

Today, many elements exist to build an infrastructure for the delivery and consumption of multimedia content. There is, however, no "big picture" to describe how these elements, either in existence or under development, relate to each other. The aim for MPEG-21 is to describe how these various elements fit together. Where gaps exist, MPEG-21 will recommend which new standards are required. ISO/IEC JTC 1/SC 29/WG 11 (MPEG) will then develop new standards as appropriate while other relevant standards may be developed by other bodies. These specifications will be integrated into the multimedia framework through collaboration between MPEG and these bodies.

The result is an open framework for multimedia delivery and consumption, with both the content creator and content consumer as focal points. This open framework provides content creators and service providers with equal opportunities in the MPEG-21 enabled open market. This will also be to the benefit of the content consumer providing them access to a large variety of content in an interoperable manner.

The vision for MPEG-21 is to define a multimedia framework to enable transparent and augmented use of multimedia resources across a wide range of networks and devices used by different communities.

This eleventh part of MPEG-21 (ISO/IEC TR 21000-11) documents best practice in the evaluation of Persistent Association Technologies – that is, technologies that persistently link information to identify and describe content with the content itself. Its purpose is to allow such evaluations to be conducted using a common evaluation framework with more specific test methodologies for each of the discussed persistent association technology types or paradigms. This Technical Report is intended to give confidence to those relying on the results that they are:

— Appropriate tests of the technology that will predict its performance under real-world conditions and

— Comparable with results obtained from other tests conducted using the same methodology.

Information technology — Multimedia framework (MPEG-21) —

Part 11:
Evaluation Tools for Persistent Association Technologies

1 Scope

1.1 Introduction

MPEG-21 will provide an over-arching framework within which many elements of multimedia are brought together. In particular, coded representations of content will be juxtaposed with metadata descriptors and the Intellectual Property Management and Protection (IPMP) protection that apply to the content. This leads to a requirement for tools that can create and maintain (e.g. detect or extract) an *association* between content, metadata and IPMP elements within MPEG-21. Tools based on the techniques known as “watermarking” and “fingerprinting” offer a means to form associations between multimedia elements and the related information, where that association can be directly embedded within or inferred from the content itself. Furthermore, tools based on watermarking and fingerprinting allow such inferences to *persist* in the face of adaptation of the content. Such tools are termed Persistent Association Technologies (PAT) and within MPEG-21 there is a need to assess and evaluate these tools. This report sets out a process and plan for evaluating PAT. It does not provide information on how to normatively interpret results of tests conducted in accordance with this Technical Report as the selection of a set of specific evaluation procedures depends on the application scenario.

This Technical Report focuses on the evaluation of two classes of technology: watermarks and fingerprints (see Definitions in Subclause 2.1) when applied to Audio content.

It is expected that the scope of this Technical Report will be enhanced in future to cover other media types including video, still pictures and text.

This Technical Report describes evaluation methodologies for only some of the characteristics of these technologies. In particular, it does not attempt to define methodologies for evaluating the resistance of these technologies to deliberate attack on the association. Further detail is contained in Clause 7.

1.2 Background to ISO/IEC TR 21000-11

Recognising that the standardisation of Persistent Association Technologies (PAT) is not currently thought to be viable in the context of MPEG-21 and the wider international standardisation community, ISO/IEC JTC 1/SC 29/WG 11 (MPEG) examined options which would allow it to assist the adoption of PAT by industry.

A call for Requirements on PAT was issued and generated numerous responses. The analysis of these responses has allowed WG11 to understand the characteristics of PAT that may be required. This analysis also showed both a need and a possibility to establish a consensus approach to the *Evaluation* of such technologies which would be useful in selecting appropriate technologies for particular applications.

1.3 Organisation of the Document

ISO/IEC 21000-11 contains nine clauses. Clauses 1 and 2 set out the scope of this Technical Report, provide definition for terms and a list of abbreviations used and not used.

Clause 3 then familiarises the reader four different persistent association technology paradigms by providing a reference architecture for each of the discussed PAT paradigms[1)]. Clauses 4 and 5 contain short use case scenarios for how to use Persistent Association Technologies, and how to evaluate such technologies.

Clause 6 then lists the seven characteristic parameters of PAT that can be used to evaluate such technologies. Before the main discussion on the evaluation methodology is discussed in Clause 8, Clause 7 contains a discussion on issues such as security and malicious attacks.

2 Terms and Abbreviations

2.1 Terms and Definitions

For the purpose of this document, the following terms and definitions apply.

2.1.1 Computational Performance

Computer scientists generally refer to the computational complexity of an algorithm in terms of the number of processor cycles needed as well as memory requirements (including lookup tables and/or databases). On any given platform (e.g. RISC, CISC and DSP) this complexity manifests itself as the *Computational Performance* of an implementation.

Watermark embedding and detection entail digital signal-processing operations that are similar to compression. The signal processing during embedding is largely concerned with the analysis of the original audio signal to so that masking methods can be exploited to reduce the audibility of the embedded signal. Digital signal processing associated with watermark detection and recovery is principally determined by the need to establish synchronisation between the detector and the embedded signal. Its computational complexity is increased if transforms involving scaling and shifting of the signals is anticipated whilst watermark detection is still required.

Digital signal processing of fingerprinting involves both calculating the fingerprint and comparing extracted fingerprint fragments with a large database of candidate exemplars, while seeking a match. The computational task is eased if simplifying assumptions are made (such as assuming a particular offset within larger objects) but is compounded if transformations such as scaling/shifting are anticipated and if the offset of the fingerprint is unknown, e.g. within a streaming environment.

2.1.2 Fingerprinting

Fingerprinting is the term used for to a type of pattern-recognition techniques when applied to identifying content and associating information with content, albeit without modifying the content. It works by extracting characteristics of a piece of audio content and storing them in a database. When the technology is presented with an unidentified piece of audio content, characteristics of that piece are calculated and matched against those stored in the database.

One technique is standardised within MPEG-7 using the *AudioSpectalFlatness Low Level Descriptor*, but there are many other approaches.

2.1.3 Impairment

Any modification to audio signals can have an impact on the perceived quality of the material regardless whether the modification is associated with embedding a PAT or subsequent manipulation of the content. In the context of assessing PAT it is helpful to define the term *Impairment* strictly on the basis of deliberate signal manipulations introduced in a controlled way for the purpose of testing reliability or robustness of the PAT. This is distinct from perceptibility effects that may be associated with embedding of a PAT (see Subclause 2.1.4).

---

1) While Clause 3 introduces four PATs (headers, digital signatures, watermarks and fingerprints), the remainder of this Technical Report concentrates on watermarks and fingerprints only.

In practice, Impairments may arise as a consequence of common signal processing operations (such as MPEG-2 or MPEG-4 lossy compression at low bit rates) and may affect the ability of the PAT to maintain an association. Other signal transformations that would not normally be described as Impairments (e.g. tone control, down mixing, etc) may also have an effect on watermark or fingerprint detection and recovery and should be considered as Impairments for the purposes of PAT assessment.

In testing, representative Impairments will be introduced in a controlled way, to explore the PAT performance according to the anticipated deployment scenario.

Impairments might also arise through deliberate attacks on the PAT that involve manipulation of the content signal in order to prevent the PAT from maintaining an association.

The perceptibility of audio Impairments would typically be measured using formal techniques such as PEAQ, MUSHRA or double blind ABX testing.

2.1.4 Perceptibility

Perceptible artefacts may arise from the embedding of watermark into a content signal. Artefacts associated with embedding may be designed to be perceptually transparent or perceptually significant depending on the design goals of the watermarking scheme. Assessment of the *perceptible effects of embedding* is considered a distinct topic, separate from the generation (and assessment of the perceptibility of) Impairments.

2.1.5 Persistent Association

In the context of MPEG-21 as well as outside, it is often necessary to create and recover associations between content items and related information (e.g. MPEG-7 metadata, unique identifiers or copy control information). A multitude of solutions encompasses the use of mark-up, tags, databases, file headers, etc. Such associations can be fragile, in the sense that if tags are stripped away (e.g. as happens when content is sent over legacy interfaces) then the association is lost.

The field of “persistent association” is also concerned with techniques by which non-fragile associations can be established or by which “broken” associations can be re-established. Presently, the main techniques of the field are watermarking and fingerprinting. The main application areas are where analogue signals (such as audiovisual content) are contained within a digital environment.

2.1.6 Persistent Association Tool

Tool for linking information to identify and describe content contained in MPEG-21 Resources with the content itself.

2.1.7 PAT Evaluation Configuration

The PAT Evaluation Configuration is defined as the combination of PAT algorithm, its parameter settings, the test stimuli and PAT payload and the set of signal Impairments used in an evaluation.

2.1.8 Robustness

The robustness property of persistent association describes the extent to which the association is maintained in the presence of processes that impact the persistent association (e.g. signal Impairment).

2.1.9 Survivability

Synonym for Robustness. The term Robustness is preferred in this document.

2.1.10 Watermarking

Modification to the content values to introduce patterns that can later be detected and interpreted as a data payload. Watermarking can be fragile or robust, or it can be perceptible or imperceptible. Different emphasis may be appropriate in differing applications. Normally the patterns embedded into signals will be assigned in such a way that they represent a symbol alphabet or carry a data payload that can be used to carry message data. Furthermore, there are classes of cryptographic watermarking algorithms where either the embedded patterns or the symbols assigned to them are manipulated according to keying material, and access to this keying material is required in order to recover the patterns or symbols. Watermarks can include any data, such as content identification and control information.

2.1.11 Feature Extraction

Feature extraction is a common sub-process in pattern recognition by which signal values are subject to transformations that yield features of the signal. This process is used within fingerprinting. Feature extraction is not synonymous with fingerprinting, but is a constituent part of it.

2.2 Terms not used in this Technical Report

Within this technical report, the following terms have *not* been used in order to avoid any confusion:

— Fuzzy hash;

— Thumb print;

— Foot print;

— Robust signature; and

— Tattoo.

2.3 Abbreviations

API
Application Programming Interface

EBU
European Broadcasting Union

IPMP
Intellectual Property Management & Protection

PAT
Persistent Association Tool

3 (Persistent) Association Technologies

3.1 Introduction

This section contains a list of classes of technologies – not specific products – that can persistently associate information with content. It will include a brief summary of the ways in which the technologies work. As part of this summary, a reference model for each PAT is provided following the below generic model:

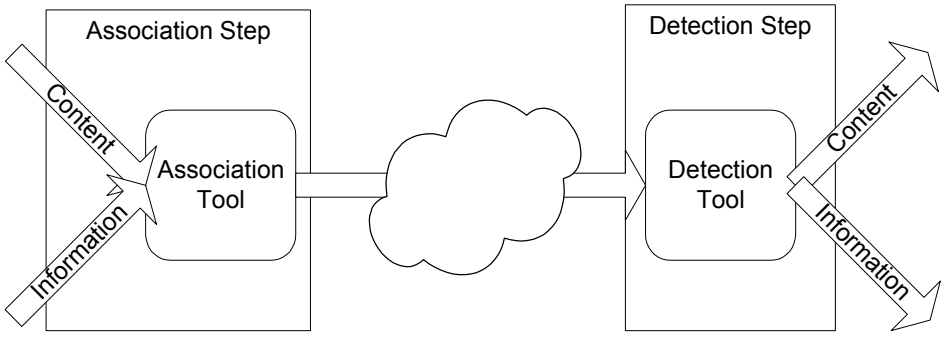

Figure 1 — Generic PAT Reference Model

It is important to note that both Association Tool and Detection Tool may use external mechanisms in performing their function. One such mechanism could be the deposit and retrieval of information from an external database.

3.2 Headers

Adding information into a header of a file is one of the most direct and simplest mechanisms for associating information with the content. Within the context of the Multimedia Framework as specified by various parts of ISO/IEC 21000, there are various slightly different flavours of this approach:

a) Adding the information to the Digital Item Declaration (DID) into a `Statement`;

b) Adding the information to the beginning of a Resource referenced from the DID (i.e. in a "header" in the narrow sense);

c) Adding the information at the end of a Resource referenced from the DID (i.e. in a "footer"); and

d) Adding the information into one or several data blocks that are interspersed throughout of a Resource referenced from the DID (i.e. in data blocks as utilised by many file formats, including MP4 defined in ISO/IEC 14496-12).

The mechanism of this technique – using cases (a) and (b) as examples – is depicted below. The Detection Tool, upon receipt of the Digital Item is then able to read the data out of the DID/Resource.

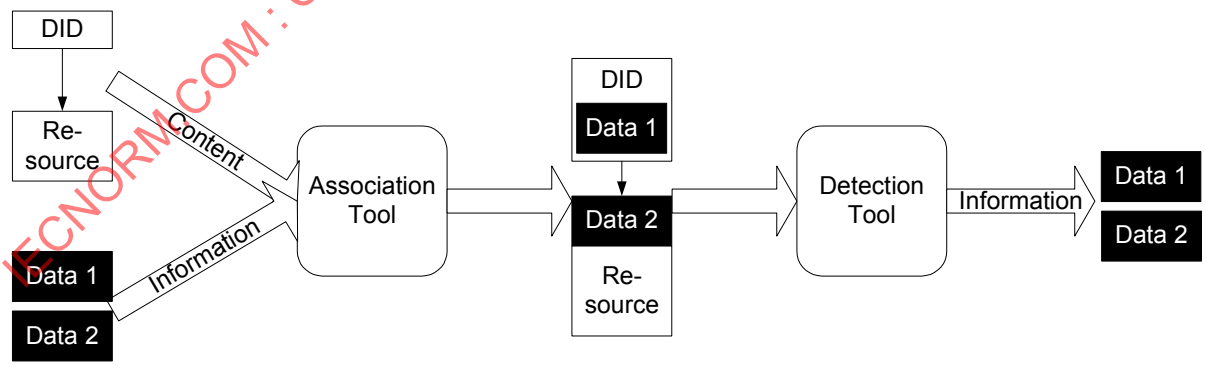

Figure 2 — Header Reference Model

In its simplest form the data will be included in clear text. This offers the benefit that finding and extracting the data is comparatively simple. However, it also has a drawback insofar as the data can (a) be read by all parties that get hold of the Digital Item and (b) that the data can easily be removed. While the first issue can be addressed by using cryptographic algorithms to cipher the content (in that case a key distribution mechanism needs to be addressed in addition), the latter can be addressed by using digital signatures (see Subclause 3.3).

In addition to storing all information in the "header", many applications use an external database to store the data in a database, add an identifier (or database key to the data) into the "Header" and then allow the Detection tool to query the database to get to the data.

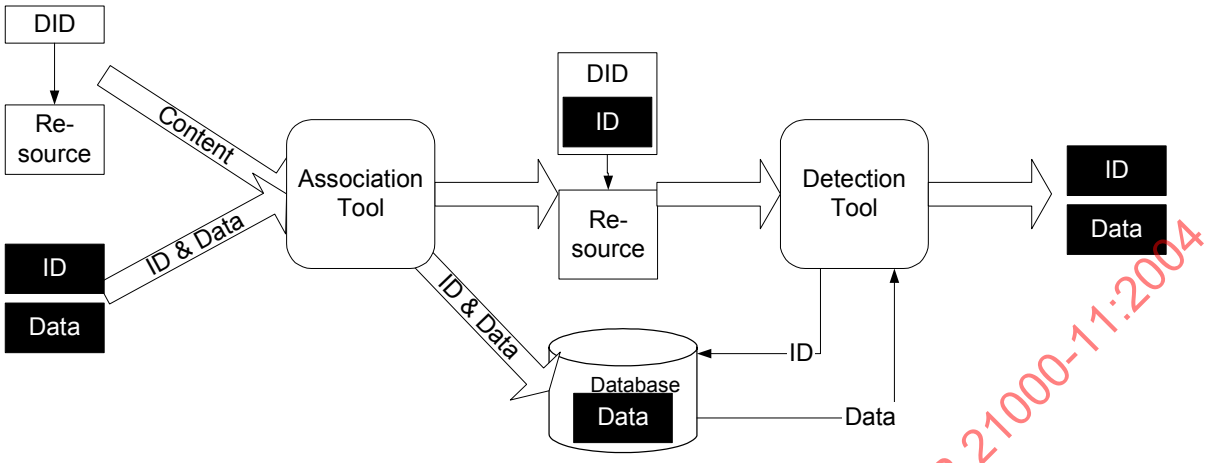

Figure 3 — Extended Header Reference Model

3.3 Digital Signatures

As indicated above, digital signatures can be used to authenticate information that has been associated using other technologies, including but not limited to information provided via headers. Also, information provided via watermarking and fingerprints can be authenticated using digital signatures. This Clause, however, uses the case of digitally signing a DID to illustrate how digital signatures can be utilised to persistently associate information with content.

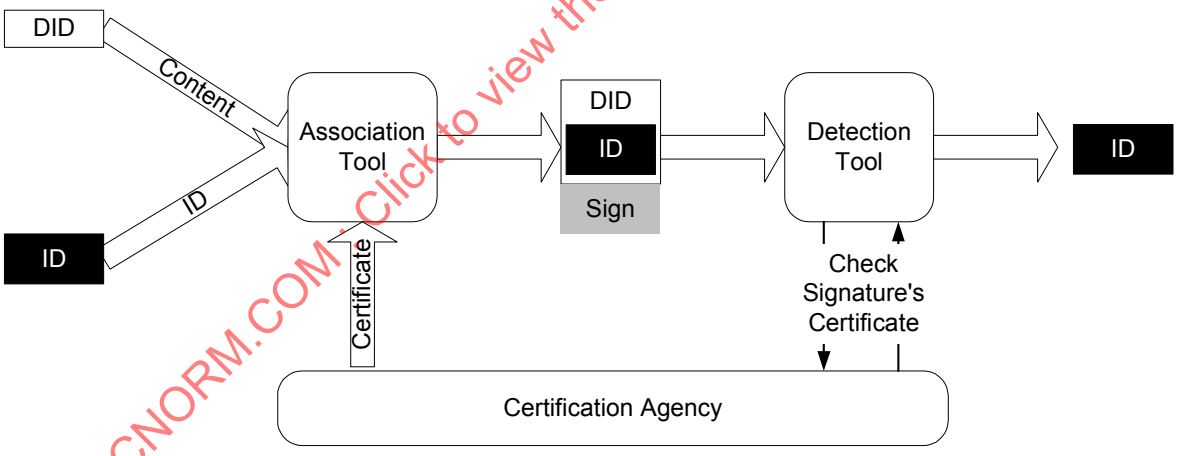

Figure 4 — Digital Signature Reference Model

The Peer or User wanting to use digital signatures acquires a certificate from a certification agency and then uses this certificate and a signature algorithm[2] to sign the DID with the ID included in the following steps:

— Calculate a hash sum over the DID with the ID;

— Use the signature algorithm and the certificate to create a signature using the hash sum;

2) An asymmetric cryptographic algorithm.

— Add the signature into the DID.

Upon receipt of a digitally signed DID the Detection Tool will:

— Calculate a hash sum over the DID – without the signature

— Use the signature algorithm together with the hash sum and the Association Tool's certificate to verify that the signature is indeed the signature of the DID created by the Association Tool.

Only when these two match can the Detection Tool be certain that the DID has not been altered and that the ID contained therein is genuine. Thus, this mechanism does not prevent the information contained in the DID to be removed. It does, however, allow for detecting changes.

3.4 Fingerprinting

Fingerprinting, or content-based identification, technologies work by extracting characteristics of a piece of content and storing them in a database. When the technology is presented with an unidentified piece of content, characteristics of that piece are calculated and matched against those stored in the database.

The following reference architecture illustrates the potential functionality of the core technology of fingerprinting. Two distinct aspects are shown in Figure 5:

a) Populating the database (top of the diagram): A series of sound recordings are presented to a fingerprint generator. This generator processes audio signals in order to generate fingerprints derived uniquely from the characteristics of each sound recording. The fingerprint that is derived from each sound recording is then stored in a database and may be associated with an identifier or other metadata for that particular sound recording.

b) Content Identification (bottom of the diagram): Audio, in either streaming or file format, is presented to the input of a fingerprint generator. The generator function processes the audio signal to produce a fingerprint. This fingerprint is then used to query the database. If a match is found, the resulting Track ID is retrieved from the database. A confidence level or proximity associated with each match may also be given.

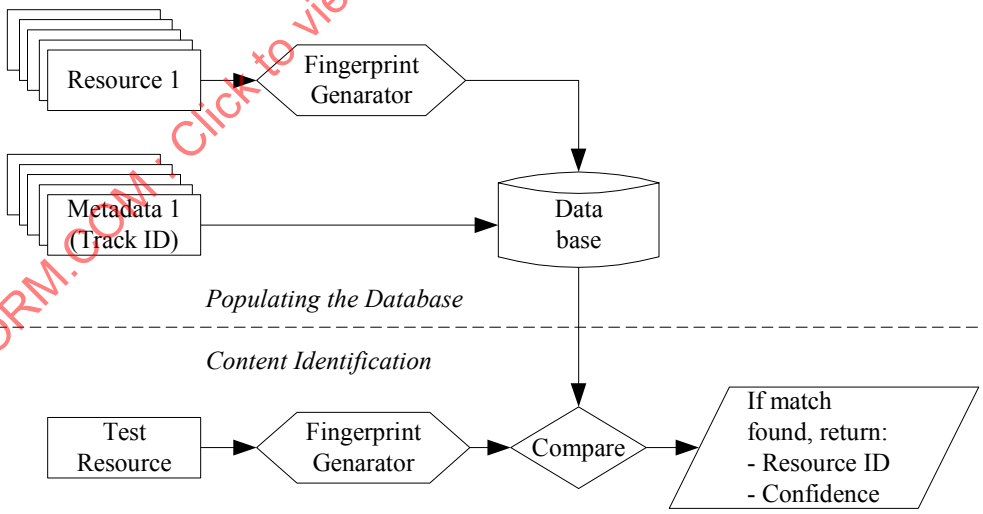

Figure 5 — Fingerprinting Reference Model

As indicated in the above diagram, fingerprinting is a technology that can be used with non-text-based Resources (e.g. audio clips, video streams and pictures) only.

3.5 Watermarking

While fingerprints are not affecting the signal quality of the content that has some information persistently associated with it, watermarks do have an affect insofar as the information that is to be associated with the content (usually called the payload) is embedded into the content (the carrier signal). As with fingerprinting, watermarking is usually used with non-text-based Resources (e.g. audio clips, video streams and pictures) only.[3] Certain watermarking applications can be implemented without the need for an external metadata database (e.g. where the watermark carries a small payload such as copy control information).

The figure below provides a reference model for watermarking.

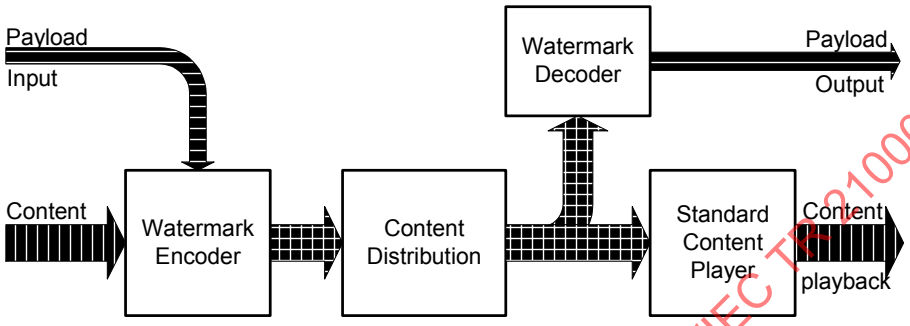

Figure 6 — Watermarking Reference Model

Note that not all of the elements in the reference models are necessarily present in every watermarking system:

— Not all of watermarking systems do need a key for embedding and detection;

— Some watermarks do not carry a message. Only their presence/absence can be detected;

— Some watermarks allow message decoding/reading, even though it is impossible to prove that the watermark was actually embedded.

All watermark applications share the property of being designed to communicate information through a multimedia signal. Depending on the application, the properties below may or may not be desired:

— Transparency – The perceptual difference between the marked signal and the host signal is limited;

— Robustness – The detection can be carried out robustly, i.e., even when the received signal is a degraded version of the marked signal;

— Security – Unauthorised people are neither able to detect, remove or modify the watermark, nor to embed a watermark;

— Large Payload – A large amount of information can be transmitted trough the watermark channel (hence, therefore often only identifiers are actually embedded, with a database containing the full set of information that is associated with the content (see also the extended header reference model within Subclause 3.2).

Moreover, error rates (false positive; false negative; bit errors in decoded message) should be small.

3) It should be noted that there are algorithms to embed a watermark into text resources, e.g. by varying line and character spacing.

4 Use Cases for Persistent Association

4.1 Introduction

This Subclause contains several of use cases for Persistent Association Technologies. In general, Persistent Association Technologies will find application in areas where metadata and content are simultaneously required. Application areas include multimedia databases, content monitoring and tracking on networks or broadcast transmissions, and in the management of digital rights (for example as a component within IPMP Systems).

4.2 Rights and Content Management

The technology used in content production has permitted the implementation of schemes that formalise the rights and content management processes. Initially, rights information is, however, not robustly linked with the content and it is very easy to separate the content from its metadata (which may include rights information).

In this situation a unique and robust identification of the work, from the moment of its creation, will allow that appropriate metadata is robustly associated with the content. This can help, for example, rights holders in demonstrating their ownership.

The techniques that exist and allow the robust linking of such metadata to content are PATs.

4.3 Audio Content Tracking and Reporting

Sound recordings are communicated over broadcast channels, cable and other networks such as the Internet. Rights holders, service providers and other stakeholders often seek to monitor the use of such sound recordings. Some of the many examples in this field include:

— Reporting. Fingerprinting and watermarking applications could track audio content played by broadcasters and webcasters to report usage. Royalties could then, where appropriate, be distributed to the correct rights holders. In such applications it may be advantageous if the technology could distinguish between different versions of a song;

— Compiling Charts. Airplay/netplay monitoring for automated charts compilation in unicast and multicast scenarios.

The metadata necessary for such applications would typically associated with the content using unique identifiers (both for manifestations[4)] and abstractions[5)]). Such identification schemes are typically administered and governed by an international agency.

4.4 Internet Audio Content Services

The Internet has brought new opportunities for distributing audio content to consumers. Distributors may need assistance in ensuring that only audio content authorised for distribution is transmitted. In such an environment, fingerprinting and watermarking technologies may be useable to identify which tracks are authorised.

4.5 Anti-Piracy Investigation and Enforcement

Anti-piracy investigators often need to analyse unlabelled or mis-labeled CDs to determine the identity of recordings on the CD. Fingerprinting technologies may help to automatically identify such works. Typical applications would include:

---

4) For music content one could use the International Standard Recording Code (ISRC).

5) For musical works one could use the International Standard Work Code (ISWC).

— Verifying that a suspected infringing recording is in fact the same as a recording whose ownership is known;

— Repertoire analysis, e.g., identifying unidentified audio content recovered from CD plants and distributors in anti-piracy investigations; and

— Screening of master recordings at CD manufacturing plants.

Forensic applications of watermarking can help to track content in its distribution path, including unauthorised disseminations.

4.6 Authentication and Integrity

Some watermarks can verify that the content is genuine and from an authorised source. Watermarks can also be used to assure the integrity content (i.e. that it has not been altered) for example by using “fragile watermarks” or by embedding digest information in the payload.

4.7 Value Added Services

Content-related services that identify a sound recording can use such identification to offer additional services such as:

— Providing informational metadata about the track, e.g., artist name, track title, lyrics, etc., by, for example, sending a fingerprint or a content identifier (e.g. an ISRC or GRid in the case of music content) to an on-line audio content metadata resolution service;

— Integration with on-line purchase services (“click-to-buy”); and

— Offering special promotions and other incentives to consumers playing certain recordings.

5 Considerations for the Evaluation of Persistent Association Tools

Persistent Association Tools involve distinct technologies such as watermarking and fingerprinting. There are many developers or vendors offering different implementations geared to different media in different applications. Questions such as the following may arise in the context of an application scenario:

— Which PAT paradigm is better suited to the particular application?

— Which PAT implementation is better suited to the application requirements?

— Which parameter settings in the PAT are needed for the application or can be set to give the best performance trade-off in the application?

An evaluation framework for PAT should assist the application developer in refining questions such as these and in providing answers to those questions.

This document has been prepared from the standpoint of evaluating PAT in a standardisation context. Here, the level of consideration must be somewhat generic in order to provide flexibility in the breadth of applications to which a standard may be applicable. But the evaluation methodologies and criteria herein should be equally applicable within the more focussed context of a specific application. Indeed, it is envisaged that PAT evaluation may assist the implementation of PAT, both within and without of a standardisation framework. This leads to several initial use-case scenarios:

— Evaluation of PAT within a standardisation process;

— Evaluation of standardised PAT during an implementation process; and

— Evaluation and/or certification of PAT implementations for use within commercial applications.

In some circumstances it may be advantageous to have an evaluation methodology constructed according to an international consensus and to be able to demonstrate that the methodology has been applied to select the best available technology and implementation and to achieve the best system performance.

6 Characteristic Parameters of Persistent Association Technologies

6.1 Introduction

This clause provides an overview of the parameters that will aid users of Persistent Association Technologies when choosing (i) the class of PAT to use and (ii) a specific solution within the chosen class of technologies. Seven characteristics have been identified as being most important. While this list of criteria is not exhaustive, a properly conducted test using them will enable the user to make an informed decision.

The seven criteria are:

— Fingerprint Size;

— Watermark Payload;

— Granularity;

— Perceptibility;

— Robustness;

— Reliability; and

— Computational Performance.

By introducing them in turn, their importance for the PAT evaluation process is explained. The *way* in which they can be used to make decisions will, however, be discussed in Clause 8 below.

6.2 Fingerprint Size

There are two distinct aspects to the Fingerprint Size that both need to be taken into account when evaluating fingerprinting systems:

a) The storage size in bytes of the fingerprint. A smaller fingerprint has an advantage as such fingerprints need to be stored in a database, processed, and/or sent over a network.

b) The total amount of source material that is encompassed by a single fingerprint. If the fingerprint covers more content, the greater is the number of features that can be included in that fingerprint. This can then aid a reliable recognition of the source material. This aspect of fingerprint size is closely related to the granularity of the algorithm).

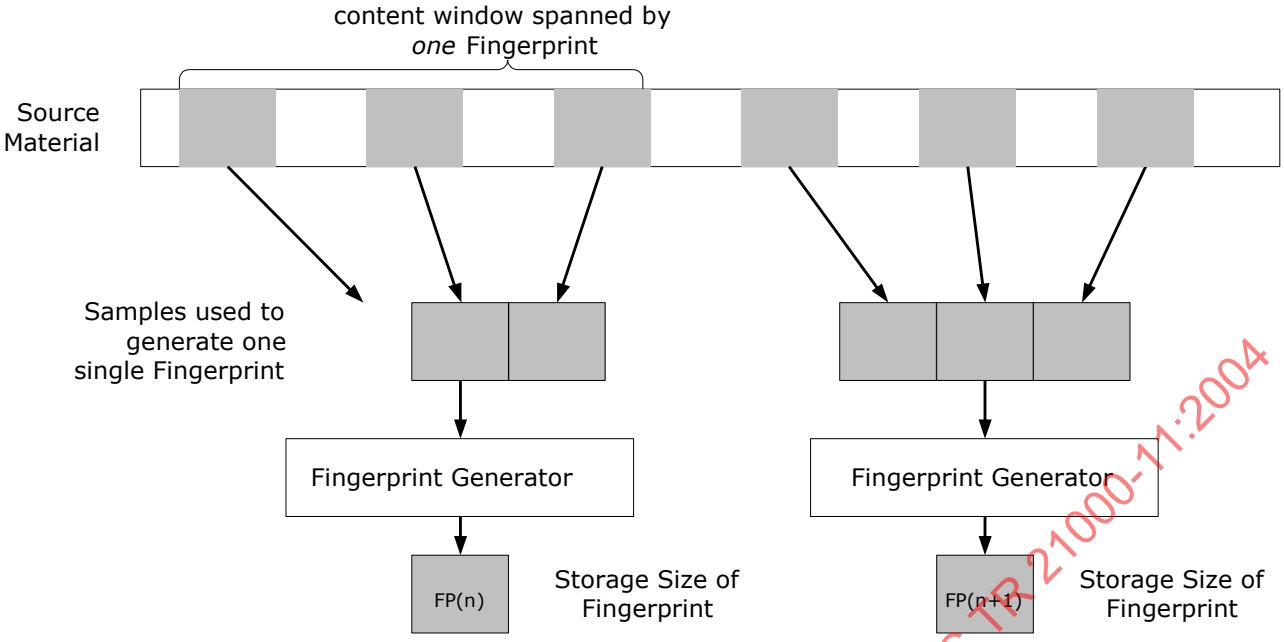

Figure 7 — Two Aspects of Fingerprint Size

The effectiveness of the Feature Extraction relates these two aspects to one another. Typical application would distinguish between (known) reference and (unknown) test fingerprints for matching. In some systems, especially in streaming applications, reference and test fingerprints have different sizes.

6.3 Watermark Payload

The payload of a watermark[6] is the number of bits that can be embedded in the input signal. This is usually described in bits per second of source material. Within the PAT evaluation, the focus is on “raw” payload size made available to the user to carry information in his application. Underlying this and within the algorithm additional error correction and/or synchronisation bits may be used. These will not be measured as they are not available to the user for his application. Similarly if the user chooses to partition data within his payload (e.g. for additional error correction such as CRC bits) or to repeat the payload redundantly across multiple watermark cycles, such application-dependent details are not within scope of the evaluation methods described in this document.

6.4 Granularity

The parameter called "granularity" is concerned with the smallest unit of source material necessary to achieve an association. This parameter is usually determined by the application domain (e.g. requirements on repeating insertion cycles of watermarks or minimum recognition latency).

6.5 Perceptibility

The parameter Perceptibility (distinct from the measurement of perceptibility), within this document, is considered to refer to the perceptible artefacts associated with embedding a watermark.

As described in Subclause 3.5, watermarks work by adding a signal to the content signal, thus altering the original content. While some watermark systems add a clearly perceptible watermark (such as television

6) This criteria is only applicable to watermarks – the equivalent criteria for fingerprints is the “size” criteria described in Subclause 6.2.

station logo inserted into the pictures of a TV programme), most system work towards watermarks that are "imperceptible" to the human auditory or visual system.

There is a six-way trade-off between all parameters described in Clause 7. However, the most users, the pertinent trade-off is between Perceptibility, Payload Size and Robustness. As a general rule, reducing the perceptibility or increasing the payload size will negatively affect the robustness. However, using more signal processing or a better algorithm may allow a better trade-off overall. For example, in order to improve imperceptibility while maintaining the other parameters, watermarking systems often employ psycho-acoustic models.

6.6 Robustness

Robustness is defined as the ability to reliably maintain an association with increasing Impairments to the signal under consideration. This criterion is distinct from Reliability (see Subclause 6.7). A reliable system may lack robustness if, with the onset of an Impairment, the ability to maintain an association is significantly degraded.

The term Impairments (as defined in can Subclause 2.1) covers a wide range of modification to audio signals. The table below provides a non-exhaustive list of Impairment types to audio content. It should be noted, however, that PATs may need to be robust to not only survive isolated Impairments but also their combinations.

Table 1 — A selection of various Impairment types for audio content

Impairment types to Audio Content
Conversion between the analogue and digital domain
Perceptual audio compression
Various filters and equalisers
Conversion between mono, stereo and multi-channel audio
Addition of noises and echoes
Changes to the time scale and pitch
Cropping and excerpting

Different signal Impairment types have different effects on different PAT systems. For example one audio PAT system may be able to survive even very strong perceptual compression while being very overcome by only slight changes in the time scale.

While many of these signal Impairments occur in the normal value chain of Digital Items, some may be deliberately employed to break the PAT. However all of the Impairment types listed in the table above can fall into either category. While this Technical Report concentrates on providing evaluation methods for the "normal" type of signal Impairments, a discussion on "malicious attacks" is provided in Clause 7).

A further aspect of Robustness is that even when an Impairment is such that the association is broken, some systems are able to infer the former presence of an association even if the exact association remains unrecoverable. This has value in some applications.

6.7 Reliability

The reliability of Persistent Association Technologies is described in statistical terms as the ability of the technique to maintain the correct association between a content item and associated payload or metadata within a given environment. In general an attempt to determine such an association will yield a result falling into one of four categories:

— *Correct Associations*: An association was created and the association method properly retrieves the correct payload/metadata.

— *No associations*: No association was created and the association method correctly declares the absence of an association.

— *False Positives*: An incorrect association to payload/metadata was found. False positives fall into two subtypes:

  — No association was created, but the association method incorrectly retrieves spurious payload/metadata; and

  — An association was created, but the association method incorrectly retrieves the wrong payload/metadata without a proper indication of error.

— *False Negatives*: An association was created but the association method does not detect any association. Since association technologies may include approximate matching or other confidence measures, false negatives may be further categorized into two subtypes:

  — No association is detected; and

  — One or more associations appear to be present, but with an indication that the payload integrity is damaged. With this, further processing may provide the accurate payload/metadata.

Evaluation of reliability should be carried out in respect of the payload/metadata that is made available to the user by the PAT implementation. Within the PAT there will typically be some levels of statistical confidence measures or other error-correction that is not visible at the user level. PAT assessment should not be concerned with the internal operation of the algorithm. Similarly, at the application level, the available payload/metadata channel can be utilised to provide addition levels of error correction e.g. through repeat embedding, through the use of parity or CRC bits, etc. The PAT assessment should concern the payload/metadata bits made available and not how they might be utilised. In contrast, by directly assessing, say, the bit error rate of the payload channel, an optimum utilisation of the available bits could be devised.

Within a controlled environment (predefined set of content elements, metadata, Impairments, etc) a series of tests can be conducted to determine statistics that will be an approximation to the reliability of the PAT in that environment.

This document will help the reader to establish a meaningful, realistic set of environments that will be useful in probing the performance envelopes of different persistent association methods for his application. These environments should be specified in such a way that repeatable experiments can be carried out on different methods, and the experimental results compared on an objective basis. Thus a procedure for conducting such tests can be established based on statistical confidence intervals or other applicable methods. This will permit the scientific evaluation of different persistent association techniques.

6.8 Computational Performance

The Computational Performance of the PAT has an impact on the performance of the overall system. In some cases this may require greater Computational Performance that the rendering of the content. When considering different PAT approaches it is necessary to understand these computational requirements in terms of the amount of processing performance required and latency added to the processing of audio content.

Computational Performance needs to be assessed both in terms of creating an association and of recovering the association:

— Creating an Association

  — The *processing and memory requirements for insertion of a* watermark. The computational requirements to perform the insertion of the watermark into the content need to be assessed. This

should be measured as the processor cycles/time required whist embedding the content and peak and average of the memory requirements over time. In certain embedding applications (e.g. re-watermarking on a user terminal) the available computational resources may be limited.

— *The processing and memory requirements for creating a reference fingerprint*. The creation of a fingerprint from source material does not create a particularly high computational overhead. However, most fingerprinting applications require the population of a reference database containing a large number of such reference fingerprints. Thus the computational requirement can be significant for (i) storing the source audio material and metadata, (ii) the extraction of the fingerprints, (iii) adding the generated fingerprints into the database. This should be measured as the processor cycles/time required, the amount of data that has to be accessed and time needed for database access. The latter may be variant depending on the number of items already in the database.

— Recovering an Association

Detection performance is important since the detector may be implemented in consumer devices that are resource-constrained, and also since the recovery step may involve bulk processing, especially in Internet-based applications. Recovery of an association may incur an overhead or latency prior to processing the content item and/or during the ongoing processing of that item:

— It may be necessary to understand how much computation is required to recover an association *prior to* processing content e.g. prior to playback (in a rendering application) or forwarding (in a network application). This should be measured as the processor cycles/time and process latency required and the amount of data that has to be accessed, and the peak and average of the memory requirements over time.

— It may also be necessary to understand how much ongoing computation is required to recover an association *while* processing content. If the PAT requires further processing during the rendering of the content this should also be quantified. For example, in the case of a watermark based PAT repetitive recovery may be required to ensure that attacks involving the splicing of non-watermarked and watermarked content are not possible. This should be measured as the processor cycles/time required whist rendering the content and peak and average of the memory requirements over time.

As well as the number of cycles or cycles per second required, some other information is required to allow comparisons to be performed:

— The test case used (length, bit rate, type, etc);

— The processor used, including the environment in which it was being operated (Memory speeds, operating environment, etc); and

— The data access requirements (amount of data required or data-rate).

Some architectures utilise distributed extraction and comparison steps involving a network interface between the two. The component parts of such architectures can readily be assessed in modular fashion. In addition, it may be helpful to compare the PAT processing to the compression processing algorithms, as many times the PAT technology can share processing and memory requirements and efficiencies (e.g. parallel processing) with the compression algorithm.

7 Issues in Persistent Association

This clause raises issues that are important in the context of identifying appropriate PAT for any application. However, these issues will not be taken into account in the Evaluation procedures presented in Clause 8 because they are difficult to resolve in general and are highly application dependent.

7.1 Robustness to Malicious Attacks

Malicious attack tests measure the ability of an attacker to modify or remove the association, whilst leaving the content as unimpaired as possible. Two forms of measuring the attack are possible – expert analytic attack or an open challenge. Such attacks will involve manipulating the content to prevent the detection of a persistent association. Thus determining a successful attack would involve measuring both the reliability of the PAT after the attack, and the level of Impairment.

7.1.1 Impairment Attacks

While fingerprinting and watermarking achieve Robustness in different ways, they are susceptible to similar Impairment attacks. In both cases a failure may occur when the signal is sufficiently impaired to break the association. For an Impairment attack to be considered "successful", the Impairment must not be such that subsequent use of the material remains perceptibly acceptable.

Deciding what level of Impairment is acceptable is difficult as many "gross" Impairments (e.g. time stretching, notch filtering, etc) may be acceptable – or even expected in certain circumstances – by the human user. The threshold at which an attack is deemed "successful" is highly application dependent.

Using formal tools to assess the affects of such Impairment attacks can be unproductive as such tools

a) Usually evaluate the perceptibility or the Impairment without giving a value judgment on the extent to which the Impairment is acceptable in a given application;

b) Are designed to measure minute or intermediate Impairments and may not deal effectively with "gross" Impairments introduced by malicious attacks; and

c) May use psycho-acoustic models for the measurements which may not be suitable to measure the specific artefacts introduced by malicious attacks, and which may nor fit such psycho-acoustic models.

Thus, such tools may not be appropriate for assessing the (gross) Impairments resulting from a malicious attack.

7.1.2 Synchronisation Attacks

A class of attacks can be described as preventing recognition by disrupting the analysis window over which any comparison is performed. Both watermarks and fingerprints are embedded or extracted over a time window and if recovery can be disrupted to attempt recognition in a different time window then the attack will succeed.

Synchronisation attacks may depend in turn on Impairments (phase distortion, cropping the start of a file, etc) in which case a method of measuring Impairments may be possible. Assessment of susceptibility to synchronisation attacks could also utilise jittering the signal and statistical analysis of the results. However the parameters and thresholds for such analyses can only be set in the context of the application.

7.1.3 Cryptographic Factors

PAT evaluation within the context of this document is concerned with testing the "persistence" properties and the "association" properties of PAT and is not concerned with how the payload or associated metadata is actually used within an application. Therefore, cryptographic factors concerning the enciphering of payload bits, for example, are not of interest from the standpoint of this document. Instead, the recovery of bits embedded as a payload, and any error rates associated with the payload bits are the focus.

Certain watermarking schemes combine cryptographic and steganographic principles in the embedding and/or detection step – e.g. to "obscure" the embedding technique by varying the embedding parameters according to a secret sequence. Such a sequence may be derived through the use of keying material, such that recovery depends on the extraction tool having access to the keying material that was used in embedding.

Fixed watermarking schemes can be interpreted as an extreme case where the fixed embedding step is considered a global secret, and security depends on secrecy of the algorithm. In this case, a leak of the encoder technology or a successful steganalysis would be considered to compromise security.

In any watermarking scheme, the payload could be encrypted and/or digitally signed, thus further increasing security to unauthorised modification and reading. However this is out of scope of this document.

It should be possible to apply standard assessment tools of information security to these aspects of the algorithm. Namely, issues include whether the algorithm:

— Relies on a global secret;

— Has a single point of failure; and

— Offers renewability.

It must be re-emphasised that this is in respect of embedding techniques, not the security of the payload data.

Fingerprinting systems can be seen to offer renewability in some sense, because new fingerprints can be derived at any time from reference material, without the need to access specific content items in the field. The practical utility of this depends on how easy it is to update fingerprint databases and extraction code in fielded devices – this cannot be taken as given. Therefore, some useful analysis can be done in respect of application architectures. But this will not typically have a bearing on the core algorithm.

7.2 Scalability

Scalability is a term that can describe:

— The number of repertoire items that can be covered (e.g. in a fingerprint system database); and

— The extent to which recognition time can be reduced by using additional processors.

Scalability has different aspects in fingerprinting and watermarking.

7.2.1 Scalability of Fingerprinting

A general principle in pattern recognition applies to fingerprinting – that is, the error rates are very nonlinear with respect to the number of items in the recognition database. This is because the ability to cleanly partition the feature space (with a decision boundary) is adversely impacted by overcrowding the feature space with many similar exemplars. It is thus axiomatic that test results obtained on databases of, say 100 and 1000 items, can *not* be used to extrapolate a predicted error rate for a larger database, say of 10,000 items. The opposite, however, is true – i.e. the error rate on a large dataset is a reliable upper bound for a smaller subset of the overall dataset.

A similar issue concerns recognition of items drawn from different sample distributions in the feature space. Results from testing on one distribution can *not* predict performance on samples from a different distribution. Development on a limited test set can lead to “specialisation” e.g. repeated testing on a “genre” of music may optimise for that genre, but may decrease performance on others.

This “scalability problem” also applies to search time and Computational Performance.

Fingerprinting systems generally scale well to handle high-volumes of query traffic, by adding additional computational resources in parallel.

A computational burden may be evident in fingerprinting within a streaming environment, as there is no general way of finding the analysis window in a stream. Solutions using multiple fingerprints at different analysis windows, or using a sliding window, are generally computationally intensive and will also impact the error-rate for a given repertoire database (as compared to recognition in a file-based environment).

7.2.2 Scalability of Watermarking

In watermarking systems the scalability is a design feature and depends on the number of payload bits the watermark is designed to carry.

In terms of query traffic volumes, watermarking, as with fingerprinting, will generally scale well by adding additional computational resources in parallel.

7.3 Interactions

There are many scenarios where multiple watermarks may be included in content, e.g.:

— Watermark “remarking”;

— Different schemes at different points in the value chain; and

— Fragile and robust watermarks.

Questions concern the interaction of multiple watermarks, both on recognition performance and on perceptibility. Tests conducted for one technique in isolation may not be directly applicable when multiple marks interact. In general, tests should be conducted (and results considered applicable) within the bounds established for testing. Watermarks may utilise different secret keys controlling the embedding processing order to minimise interaction between different watermarks. This can help to “layer” different watermarks within the same content.

It is not likely that watermarking and fingerprinting schemes would adversely interact as fingerprints usually operate far above psycho-acoustic thresholds and most watermarks are below those thresholds[7]. Furthermore, fingerprints cannot, by definition, modify the content – marked or otherwise.

By design, watermark and fingerprint systems can be usefully combined to gain benefits of both techniques in a single system. However, issues related to the architecture for such systems are not a concern of this document. This document should enhance the ability to quantitively assess such combined systems.

8 Evaluation Methods for Persistent Association Technologies

8.1 Introduction

PAT evaluation occurs at two distinct levels. At a *high level*, general characteristics of PAT can be compared against the application requirements. At a second, lower, level, PAT parameters and performance can be evaluated. Within this section, the two distinct aspects of the evaluation of PAT are described. Firstly, a general framework is set out describing guidelines for PAT evaluation within a generic framework. Secondly, and at a more detailed level, the seven main parameters of PAT (see Clause 6) are described in the context of how they interact and how to test them. Recommendations are set out in this document for evaluating the seven parameters. It is also important to understand what the evaluation results *mean*. In general, there is no meaningful notion of *ideal* results except in the context of a tightly defined set of application criteria, utilising a deep understanding of the market, problem and possible solutions.

8.2 Generic Framework and Methodology for Evaluation of PAT

The first step in any PAT evaluation concerns setting out the framework within which the application will be utilised, and then qualitatively examining different approaches to PAT within that framework. Initial questions will concern the *applicability* of the PAT paradigms. Questions may be asked such as:

---

7) Naturally, when the watermark *sufficiently* manipulates the source material there may be affects to the reliability of fingerprinting systems that are very sensitive.

— Is the information in the PAT self-contained, or is a database required?

— Does the PAT need to scale across a large repertoire of distinct content files?

— Are false positives and false negatives of equal importance in the application?

— Will the PAT be subject to malicious attacks?

— Does the PAT need to carry cryptographic elements, or be embedded using cryptographic techniques?

— What computational resources are required for embedding or recognising the PAT?

Questions such as these are not a part of the PAT Evaluation Configuration and they are not questions for which quantitative results can be gained as they depend upon the business model and threat analysis. For this reason, guidelines rather than recommendations regarding these aspects are provided in this document. These guidelines include information on:

— What data will be used for testing including:

  — Type of data (analogue, digital, etc.)

  — Genre of content (classical, jazz, etc.)

  — Scale of database (1000's, millions, etc.)

  — Metadata type (identifiers, information payload, etc.)

— What characteristics will be measured to obtain results?

  — False positives/negatives, "similarity", etc.

— How will results be measured?

  — Is 'not recognised' a permitted result in testing?

  — Statistical evaluation based on multiple "passes" or evaluation steps?

— How will results be analysed and used to "evaluate" the PAT configuration?

  — What weighting is applicable to false negatives / false positives (application issue)?

  — What weighting is applied to different parameters that may be measured (perceptibility, error-rate, etc) (application issue)?

  — What does the evaluation result "mean" in the context of the application?

In general, it is not possible to set out a single test environment against which PAT can be evaluated. Instead, in actually implementing a PAT evaluation process, questions such as the above must be posed and answered as part of both understanding the applicability of the PAT and of building the test environment to assess the PAT.

In general an evaluation will be used to measure raw results, such as error rates, perceptibility, etc. Then, in an application, given the raw channel bits and error rate, standard information theory can be used to determine how best to allocate the bits in a given application (e.g. partitioning some bits for CRC, redundant additional watermarks or fingerprints, etc). In the context of the application, there will need to be a framework for interpreting the results. This framework is a step in the general methodology of Subclause 8.4 and is application specific.

8.3 PAT Evaluation Configuration

Once the Evaluation Configuration is defined to give a general testing framework, a detailed consideration of the PAT parameters can be undertaken. In general, PAT implementations involve many interacting parameters whose settings influence the performance of the tool in profound ways. Since multiple interacting parameters are available, certain parameters must be selected as controlled parameters whose values will be fixed or defined and other parameters will become dependent variables. For example, the payload size and granularity of a watermark may be fixed along with the reliability and perceptibility and then the robustness and Computational Performance can be measured.

The combination of a defined test environment and a defined set of parameters for testing is termed a PAT configuration. The type of test results obtained will depend strongly on the configuration.

Various PAT configurations can be evaluated against given design requirements. However, results for any given configuration of PAT cannot meaningfully be compared against results generated for a PAT that was configured for a different application. This would be an 'apples and oranges' comparison. Therefore we can state three evaluation rules at the outset:

— An evaluation is carried out in respect of a PAT Evaluation Configuration, not in respect of a PAT;

— Evaluation results are valid for only the PAT Evaluation Configuration that was tested; and

— Results for a PAT Evaluation Configuration cannot in general be compared against results for a different Configuration. Results can only meaningfully be compared if particular controlled parameters are varied in a systematic way to explore performance trade-offs.

8.4 Generic PAT Evaluation Process

Regarding the general evaluation framework, whilst the Evaluation Configuration together with test criteria and weighting of importance to different criteria are all strongly application dependent, the general methodology around those is not. Therefore it is possible to set out a recommended evaluation framework as below. Some parts of this recommended framework are accompanied by guidelines and some parts by further recommendations, depending whether the parts are application-specific or PAT specific. The framework comprises a sequence of steps:

a) Evaluate the applicability of the technology to the application (guidelines, as throughout this document):

   — Define functional requirements and map to existing technologies;

b) Select the criteria for the evaluation process (guidelines):

   — Controlled parameters vs dependent variables;

   — Weighting of importance;

c) Structure the test stimuli and outputs (guidelines):

   — Specify the source material and payload data;

   — Define inputs/outputs;

   — Define a methodology for results/data collection;

d) Prepare a methodology for analysing results. (Recommendations, expanded in Clause 8 as below):

   — Reliability (see Subclause 8.5);

   — Perceptibility (see Subclause 8.6);

- Payload/Size (see Subclause 8.7);
- Robustness (see Subclause 8.8);
- Granularity (see Subclause 8.9);
- Complexity (see Subclause 8.10);

e) Conduct the Evaluation and collate test results;

f) Interpretation of results in light of the requirements (guidelines):

- Determine the suitability or performance level of the PAT within the application at hand.

While this process requires careful planning in order to obtain valuable results, certain elements can be automated. Some recommendation on the automation of PAT evaluations can be found in Subclause 8.11.

8.5 Evaluation of Reliability

This clause provides various metrics by which reliability (as introduced in Clause 6) can be established and how these metrics can be used.

The reliability of Persistent Association Technologies is described in statistical terms as the ability of the technique to maintain the correct association between a content item and associated metadata within a fixed environment.

Test results will depend on the test set (stimuli) used in testing and it is essential to specify the data set used in testing – not only the number of items in the data set, but including information on what each item is. Subclause 8.2.1 gives guidelines on the importance of the scale of the test set, and the effect on decision boundaries within the feature space.

Depending on the application a different weighing is put on false positives and false negatives. False positives can be difficult to measure. In watermarking applications they are often measured theoretically since false positives rates can be around 10^{-9} to 10^{-12} which is often expensive to test.

As the system reliability increases, false negatives can also have very low occurrence rates. For content with only very slight Impairments (since the creation of the association), the false negative rates can be so low that a theoretical analysis is the only practical method for assessing the error rate.

For evaluating reliability, the following recommendations are given.

Recommendation: Wherever possible reliability should be assessed using practical tests. Theoretical assessments should only be used to augment such practical tests, when error rates are expected to occur with a very low probability. When results combine theoretical and practical assessments, the basis for all results should be clearly indicated.

Recommendation: the dataset used in testing should at minimum be sufficiently large to provide results that are statistically significant, and which can include confidence intervals. The data set should also be large enough to represent the number of items that will be included in the application domain.

Recommendation: The test dataset should have the same statistical distribution as the items that will be encountered in the application domain.

It should be noted that bias can be introduced if the same dataset is used repeatedly in assessing a PAT during a development or implementation phase, and that evaluation in general should not be based on the same dataset as was used in development. Methodologies for dealing with this issue are used in statistical analysis and machine learning – such as partitioning the data set, the ‘stop-out’ method, etc. These methods may be of use, but may not be fully appropriate in PAT evaluation.

Recommendation: The same dataset that was used in development should not be used in evaluation. The smaller the overlap between the development dataset and the evaluation dataset, the more independent will be the results.

In general an attempt to determine such an association will yield a result falling into one of four categories, (See Subclause 6.7):

— Correct association;

— False Positive;

— False Negative; and

— No Association.

Recommendation: In the evaluation process, results falling into each of the categories shall be captured to form an evaluation result. The weighing of importance of each class of result is application dependent, and in some applications the cost or value placed on a particular class of result may fall to zero.

In watermarking applications that carry an information payload, results can be gathered at a more detailed level than merely assessing whether an association can be established. The payload bits can be recovered and examined to determine a channel error rate.

Recommendation: In Watermarking applications, the payload bits should be recovered and an error rate determined for the watermark data channel.

Once the raw channel error rate is known, information theory may be applied to analyse the reliability of the channel. This would, for example, allow the implementer to know whether some payload bits need to be allocated for CRC or error-correction purposes.

As described above, evaluation of reliability is conducted in a fixed environment, i.e. at a fixed level of Impairment.

Recommendation: The reliability evaluation should be repeated for different levels of Impairment as described in Subclause 8.8 (Robustness).

8.6 Evaluation of Perceptibility

Perceptibility applies to only those forms of PAT that manipulate the actual content data. Evaluations should be conducted to determine any perceptible artefacts. The measurement of perceptibility deals with subjective and objective testing methods that estimate the degradation of persistently associated content in comparison to the respective original content. Subjective testing means that human beings are involved in the test, whereas objective measurements try to simulate the outcome of subjective tests by means of a (computer) machine.

There are a number of standardised methodologies that describe methods for evaluating perceptibility and that are appropriate for use in PAT evaluation. These methodologies are listed in Table 2 below.

Methods of testing that are not standardised are also widely used, particularly with trained listeners in a studio environment. Such methods include various forms of AB testing, ABX testing and ABX double-blind testing.

Recommendation: Formal methods are recommended for perceptibility evaluation. Methods such as double-blind ABX may be applicable in perceptibility evaluation if used with care. Such methods should be used within a framework that ensures rigorous statistical interpretation of the results and any associated confidence intervals.

Recommendation: Subjective tests are the preferred method. When implementing a subjective test, the use of both skilled and trained listeners as well as unskilled listeners is recommended.

Table 2 — Impairment Test Methods: Audio

Test Method	Test Type	Description	Target quality
BS.1116	Subj.	The ITU Recommendation BS.1116 “Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems” is targeted on assessing small Impairments for high quality audio material.	High
Pair Test	Subj	This test is targeted on assessing very small Impairments for very high quality audio material. It can be used for proving audibility of Impairments. The opposite, i.e. inaudibility, however cannot be proven with this test.	Very high
MUSHRA	Subj	The ITU Recommendation ITU-R BS.1534 "Method for subjective assessment of intermediate audio quality"	Medium
PEAQ	Obj.	Draft new Recommendation ITU-R BS.[10/20] – “Method for Objective Measurements of Perceived Audio Quality” is targeted on assessing medium quality audio material. The PEAQ system has been developed in order to get a perceptual measurement scheme that estimates the results of real world listening tests as faithfully as possible. In listening tests for very high quality signals, the test subjects sometimes confuse coded and original signal and grade the original signal below a SG of 5.0. Therefore the difference between the grades for the original signal and the signal under test is used as a normalized output value for the result of the listening test.	Medium-high
PAQM	Obj.	PAQM derives an estimate of the signals on the cochlea and compares the representation of the reference signal with that of the signal under test. The weighted difference of these representations is mapped to the five-grade Impairment scale as used in the testing of speech and audio coders. As such the PAQM system tries to estimate test results on a scale used for BS.1116 tests.	
NMR	Obj.	The third system used in the evaluations is the NMR. The interesting output value is the overall NMRtotal value expressed in dB to indicate the averaged energy ratio of the difference signal with respect to a just masked signal (masking threshold). Usually, at NMRtotal values below –10 dB there is no audible difference between the processed and the original signal. This system historically led to the PEAQ system and is now a subset thereof.	Medium-high

Recommendation: If objective tests are to be used, it is recommended to use PEAQ as PEAQ is a successor to NMR and PAQM.[8)]

For all tests mentioned in the above table it is important to select an appropriate body of test items.

Recommendation: Typically items should not exceed 15-20 seconds in length and should be selected by a pre-selection test before the actual listening test. It is highly recommended to indicate audio signals that are known to be critical for audio coding since many watermarking schemes apply methods that are also used in the field of audio coding. If non-critical material is chosen for these tests the results are of no value.

8.7 Evaluation of Payload/Size

The payload size of a watermark is the amount of information that is carried by the watermark that is made available for use in applications. Depending on the underlying media-type, this amount is usually measured in bits/second. There is a direct trade-off between payload size and other parameters, like robustness, perceptibility and reliability. In many cases Payload/Size is not measurable and has been predetermined from

8) As observed elsewhere, these objective tests have been designed for large changes in the content and environment, whereas watermarking usually results in minimal changes in the content.

the business and application requirements. On the other hand, if Payload/Size is measurable, recommendations are given below.

There are three classes of watermarking systems:

a) Watermarking systems that have a fixed payload size. For such systems the payload size is fixed within the algorithm and only robustness and reliability will vary within the evaluation.

b) Watermarking systems with a variable payload size, which can be set by the user. This case can be dealt with in the same way as the fixed payload case: the evaluator sets a relevant payload size and subsequently compares the other indicators of performance of the watermarking systems.

c) Watermarking systems where the payload size is dynamically determined by the embedder (usually from the source signal statistics). Since the instantaneous payload rate will vary in such systems, measuring the payload size will require determining the resulting payload sizes from a sufficiently rich set of test signals, and statistically analysing the results (e.g. mean; median, variance).

Recommendation: Measurements of the payload size should be limited to the "raw" payload size and not consider application-dependent usage of the date (e.g. to carry error correction information).

The size of fingerprints is determined in terms of the number of bytes that is (a) stored in the reference fingerprint database or (b) generated from unknown source material and that is forwarded (possibly via a network) to the "matching engine".

Recommendation: The fingerprint size should be assessed for suitability with respect to the intended application.

Fingerprints have no payload size. However, both fingerprints and watermarks span a window in the source material as described in Subclause 8.9 on evaluating Granularity.

8.8 Evaluation of Robustness

Robustness is the ability of a persistent association method to maintain an association in an environment where signal modifications occur. These modifications can be divided into signal Impairments which arise in the normal operation of the system and malicious "attacks". Although the nature of both Impairment types may exhibit similarities, this document is only concerned with those that arise in the normal operation of the system.

Recommendation: In any evaluation of the robustness of a PAT, an appropriate set of Impairments need to be defined. Moreover, levels for each type of Impairments need to be established according to the application requirements. For audio content, Table 3 includes a non-exhaustive list of relevant Impairments.

Table 3 — Recommended Impairments

Impairment Class	Impairment Type	Parameter Types
Perceptual Coding	ISO/MPEG Layer II as defined in ISO/IEC ISO/IEC 11172-3	At various bit rates
	ISO/MPEG Layer III as defined in ISO/IEC ISO/IEC 11172-3	
	ISO/MPEG-2/4 Advanced Audio Coding as defined in ISO/IEC 13818-7 and ISO/IEC 14496-3.	
	ISO/MPEG-4 High-Efficiency Advanced Audio Coding as defined in ISO/IEC 14496-3.	
	Commercial codecs such as Dolby AC-3, Dolby E, Microsoft WMA9	

Impairment Class	Impairment Type	Parameter Types
Tandem Coding		Using various perceptual coders and bit rates
DA/AD Conversion		Several times
Filters	High-Pass	Various frequencies and roll-off
	Low-Pass	
	Band-Pass	
	All-Pass	
Down-mixing, up-mixing	Multi-channel to stereo	
	Stereo to mono	
	Multi-channel to Dolby Surround	
	Dolby Prologic	
Signal Addition	Pink Noise	dB levels relative to peak/average source level
	White Noise	
	Voice-over	
Time scale modifications	Changing the sampling rate	Various speeds
	Pitch corrected time scaling	
	Speed change (render items at non-nominal sampling frequency)	
Studio techniques	Pitch shifting	Various levels
	Multi-band equalization	
	Echo addition	
Cropping or excerpting of content		Various lengths of excerpts
Combinations of the above		

Recommendation: Robustness testing needs to be applied iteratively at successive Impairment levels for each appropriate type of Impairment while measuring the Reliability of the association at each step. The methodology should allow the Reliability of the PAT to be established in terms of error rates at each step of Impairment. In many cases a threshold can then be determined at which the PAT fails to maintain an association reliably.

8.9 Evaluation of Granularity

The performance of PAT systems is dependent on the content window over which the PAT works. Certain applications may require a short window, to allow recovery of the association from a small content fragment. Alternatively, there may be a need to carry a large payload in the watermark or to include many features in a fingerprint, and this would lead to a need for an increased window length.

The evaluation of granularity is intended to provide the evaluator with a means to investigate the system performance in respect of window size. In particular, the evaluation of granularity will allow the evaluator to determine the reliability of association as small fragments of content are provided for analysis.

Test material needs to be provided in order to retrieve the associated information at a certain confidence level. For watermarking systems this is related to the length and repetition cycle of the embedded message. For