ISO/IEC JTC 1/SC29/WG1 N100099
94th Meeting, Online, January 17-21, 2022

ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG
Joint Bi-level Image Experts Group

JPEG
Joint Photographic Experts Group

TITLE: Use Cases and Requirements for JPEG Fake Media

SOURCE: WG1

PROJECT:

STATUS: Approved

REQUESTED ACTION: For distribution

DISTRIBUTION: Public

Contact:
ISO/IEC JTC 1/SC 29/WG 1 Convener – Prof. Touradj Ebrahimi
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
Tel: +41 21 693 2606, Fax: +41 21 693 7600, E-mail: Touradj.Ebrahimi@epfl.ch

Use Cases and Requirements for JPEG Fake Media

1 Summary

Recent advances in media creation and modification technologies allow the production of near realistic media assets that are often, to the human eye, indistinguishable from original assets. These developments open opportunities for creative production of new media in the entertainment and art industry. However, the intentional or unintentional spread of manipulated media, *e.g.*, modified media with the intention to induce misinterpretation, also imposes risks such as social unrest, spread of rumours for political gain or encouraging hate crimes.

Clear and transparent annotation of media creation and modifications is a crucial element in many usage scenarios bringing trust to the users. This has already triggered various organizations to develop a wide range of mechanisms that can detect and annotate modified media assets when they are shared. However, these annotations should be attached to the media in a secure way to deter them from being altered. In addition, to achieve a wide adoption of such an annotation ecosystem, interoperability is essential and this clearly calls for a standard.

This document introduces the Use Cases and Requirements for JPEG Fake Media. The scope of JPEG Fake Media is the creation of a standard that can facilitate a secure and reliable annotation of media asset creation and modifications. The standard shall address use cases that are in good faith as well as those with malicious intent.

2 Introduction

Current technologies permit the modification or synthetic creation of media assets. Some, like deep learning methods, can create media assets that are hard for people to distinguish from natural media assets. These developments open new, creative opportunities that are useful for the entertainment industry and other business usage, such as the creation of special effects, artificial but photorealistic scene production with actors in the studio, or restoration/re-colourisation. However, this also leads to issues relating to the use of manipulated media to spread disinformation. Misuse of manipulated media can cause social unrest, spread rumours for political gain, or encourage hate crimes.

In many application domains, creators may want, or need to declare the type of modifications that were performed on the media asset, in opposition to other situations where the intention is to obfuscate the existence of manipulations. Media modifications are not always negative as they are increasingly a normal and legal component of the production pipeline. Therefore, various governmental organizations plan new legislation. Some companies, including social media platforms and news outlets, are developing mechanisms that would clearly detect and annotate manipulated media when they are shared, attempting to avoid negative impacts.

While growing efforts are noticeable in developing technologies, there is a need to have a standardized way to annotate media assets (regardless of the intent) and securely link them together. Therefore, the JPEG standardization committee (under auspices of ISO, IEC, and ITU) has launched an initiative to identify the standardization needs related to the facilitation of the secure and reliable annotation of modified media through an in-depth analysis of various usage scenarios. While the initiative is called *JPEG Fake Media*, it is important to stress that it addresses both good faith and malicious usage scenarios.

JPEG initiates a standardization activity in order to ensure interoperability between a wide range of applications dealing with media asset creation and modifications. To reach this goal, JPEG has and continues to invite stakeholders to join the effort by helping to better understand applications and scenarios relevant to JPEG Fake Media use cases. This allows the JPEG committee to identify key requirements for a standard in this context. Initial findings suggest that a set of standard mechanisms to describe and embed information about the creation of media assets as well as modifications are needed. In addition, standard mechanisms for security and protection of integrity of media assets are desired. The latter is closely related to issues highlighted in media blockchain, which has been in progress for a few years in JPEG and therefore is considered as a natural continuation of that effort.

The JPEG Committee engaged with stakeholders to develop a clearly defined roadmap for standardization. This roadmap includes use cases from relevant industries, public bodies (responsible for legislation), technology providers and end-users.

3 Scope

The scope of JPEG Fake Media is the creation of a standard that can facilitate a secure and reliable annotation of media asset creation and modifications. The standard shall address use cases that are in good faith as well as those with malicious intent.

4 Definitions

To ensure a correct understanding of the descriptions in this document, this section defines terms and concepts as they are used in the context of this work.

- **Misinformation**: information that is false but not created with the intention of causing harm[1].
- **Disinformation**: information that is false and deliberately created to harm a person, social group, organisation or country[1].
- **Malinformation**: information that is based on reality, used to inflict harm on a person, social group, organisation or country[1].

- **Media asset**: digital assets including images, videos, audio or text. In the context of this document we mainly focus on images, however, other media types are not necessarily excluded from the scope.
- **Natural media asset**: sensor acquired media asset.
- **Synthetic media asset**: media asset generated at least partially by a computer programme.
- **Media asset content**: the portion of a media asset that represents the actual content, such as the pixel data of an image, along with any additional technical metadata required to understand or render the content (e.g. a colour profile or encoding parameters).
- **Media asset metadata**: the portion of a media asset that represents non-technical information about the media asset or its content, such as location, creator, annotations or IPR information.
- **Actor**: A human or non-human (hardware or software) that is participating in the media ecosystem. For example: a camera (capture device), generation or editing software, cloud service or the person using such tools.
- **Region of Interest**: subset within the media asset content identified for a particular purpose. For example: the face portion of a portrait image, an extracted foreground object(s) or scene cuts of a video.
- **Coordinate system**: a method of representing points in a space of given dimensions by coordinates.

[1] As defined by UNESCO: https://en.unesco.org/fightfakenews

- **Media asset origin**: the actor that created the media asset.
- **Media asset provenance**: a set of information about a media asset including the trail of modifications starting from an actor, for example, the media asset origin.

- **Media asset source**: media asset produced by a device or method without any modifications.
- **Digital master**: master media asset as intended by its creator.
- **Modified media asset**: media asset that has been changed.
- **Manipulated media asset**: media asset that has been changed with the intention to induce misinterpretation.
- **Composed media asset**: media asset composed of multiple media assets.

- **Media asset integrity**: lack of corruption of a media asset.
- **Authentic media asset**: media asset that is verifiable and/or trustworthy
    - Verifiable: able to be checked
    - Trustworthy: able to be relied on as being what it is asserted to be

- **Signing**: a process that establishes the relation between an actor and a media asset in a tamper-evident manner.
- **Signer**: an actor who digitally signs a media asset.
- **Registration**: the process of storing information (e.g. media asset, metadata or provenance) about a media asset, separate from the media asset itself.
- **Registrar**: an actor that performs a registration.

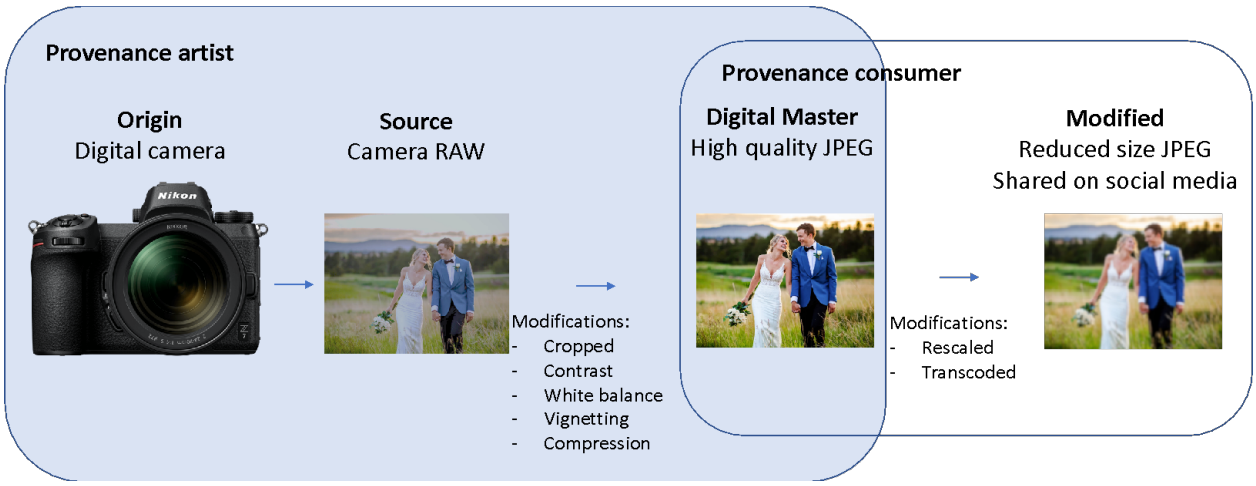

Figure 1: Illustrative example of selected definitions using wedding photograph scenario.

Figure 1 provides an illustrative example of some of the definitions using the scenario of a wedding photograph. The *media asset origin* is the digital camera used to capture the image, the image produced by this camera is the *media asset source*, in this case stored in a camera RAW format. The photographer (artist) makes some changes to the *source image* such as cropping, enhancing the contrast, correcting the white balance, and applying vignetting. Finally, they store a high-quality JPEG image version which is called the *digital master*, which is shared with their client

(consumer). Note that the *digital master* is a version that has been modified from the *source image*. At this point the *provenance* of the media asset spans from the *origin* to the *digital master*. When the client shares the image via social media, a newly *modified* version is created that entails rescaling to a lower resolution and transcoding to a lower quality JPEG. Since the client may not have access to any provenance information prior to the *digital master*, the *authenticity* for the client therefore relies on her/his trust in the photographer/artist who has created the *digital master*.

5 Use Cases

One of the key objectives of the JPEG Fake Media initiative is to better understand topics and use cases that fall under its scope and to analyse their implications, especially from a standardization point of view. Currently, the JPEG committee has identified the following topics and use cases:

- **Misinformation and disinformation**
  - Deepfakes
  - Manipulated media
  - Media intentionally used out of context
- **Forgery / Media forensics**
  - Document forgery (e.g. IDs and passports)
  - Insurance fraud (e.g. pictures of accidents)
  - KYC (Know Your Customer) (e.g. fake identity)
  - Impostering (e.g. impersonating a celebrity)
- **Media creation**
  - Use of deepfakes for special effects
  - Green screens, media processing and composition
  - GAN (Generative Adversarial Network) images
  - Short content bursts
  - UGC (User Generated Content) e.g. TikTok, Triller, Adobe Spark
  - Picture and movies production
- **Media modification**
  - Image editing software
  - Movie preservation
  - Film enhancement
  - Restoration of old movies or photographs

Based on the above, the following sections provide an overview of illustrative use cases. Both the topics and use cases may be extended in the future.

5.1 Misinformation and disinformation

5.1.1 Media usage in breaking news

In his coverage, a journalist wants to use images from a social media post depicting police violence during protests. The journalist has to make a fast decision but of course he wants to be sure the image in the post is genuine and taken at the mentioned place and time.

5.1.2 Deepfake detection

A news host wants to double check if a video she received of the president making questionable claims is genuine and not a deepfake.

5.1.3 Content authenticity checking

An investigative journalist wants to verify if an image depicting past atrocities is actually from that era and place.

5.1.4 Content usage tracing

A photographer wants to find out where and how some of the images from his portfolio have been used and check whether they are used in a genuine context.

5.1.5 Academic research

An academic journal reviewer might want to know that an image used as evidence for a successful experiment hasn't been altered and is accurately used.

5.1.6 Photographic framing

A journalist received images of the Grand Place in Brussels in the aftermath of the terroristic attacks. Due to the specific framing, the images give a frightening impression of the situation. Therefore, the journalist wants to compare with other images taken at the same place and time but from different perspectives to better evaluate the actual situation.

5.2 Forgery/media forensics

5.2.1 Insurance fraud

In the context of insurance fraud, an insurer might want to check whether an image used as evidence has been manipulated.

5.2.2 Mileage reporting photo

A car insurance company provides a discount program for the customer of limited annual mileage and demands the annual-reporting photo showing the mileage and the time displayed on the front

panel of the customer's car. This insurance company might want to check whether the photo reported has been manipulated.

5.2.3 Photo for cost charge

A series of before & after photos is frequently used for charging repair-costs in modern digital society. In this case, the integrity of a series of photos with the timing information from the origin to the final needs to be authenticated.

5.2.4 Evidence of Trial

A prosecutor wants to verify whether a movie recorded by Closed Circuit TV Security System was really taken at the location and the time.

5.2.5 Media sharing on social media

A media consumer (end user) wants to verify the credibility of a news article shared on social media and he/she would like to trace who created, modified and published the image on it.

5.2.6 Credibility of AI training image data sets

An online auction service buys a set of training image data from a stock photo service and wants to check if each image was really taken by a camera instead of being created synthetically.

5.3 Media creation

5.3.1 Movie special effects

A creative movie production company has created several shots for a movie that are computer generated but almost indistinguishable from real footage. The generated footage is labelled to allow consumers to identify that the content is computer generated. Since the final movie is a composition of generated and real footage, the entire movie can be labelled frame by frame.

5.3.2 Media transcoding

A photographer develops multiple versions of an image for different purposes. This includes the camera RAW image, rendered JPEG, moderately enhanced image and varying quality versions for web preview or print. During each transcoding step, authenticity and IPR information is retained from the parent version to the child version. In addition, authenticity information might be updated to describe modifications inherent to the transcoding process such as loss of quality when transcoding to a lossy format.

5.3.3 Chroma keying or silhouette extraction

Using chroma keying or silhouette extraction, a reporter can be virtually placed in a different location. Labelling the content allows media consumers to identify whether the shots were actually taken at the location or not.

5.4 Media modification

5.4.1 Image colorization and restoration

A developer has created an algorithm that uses deep learning to colorize grayscale images and enhance image quality. The output images are labelled to allow consumers to identify that these images have been processed and may not accurately reflect original colours.

5.4.2 Photo editing

A photographer uses photo editing software (e.g. Photoshop) to edit model pictures for a magazine. The final images are labelled to indicate that they are post-processed. The labels allow signalling of how “severe” the changes are to distinguish simple contrast and tone enhancements from changes where content has been added, removed, modified or manipulated.

6 Threat Vectors

Establishing the authenticity of a media asset is fundamentally an issue of trust. Threat vectors refer to the different possible approaches to compromising that trust. Some threat vectors include:

- Disassociated metadata
- Broken provenance chain
- Replaced trust
- Fractured initial trust
- Inaccessible resources
- Impostor signing

6.1 Disassociated metadata

An application or online service may remove embedded metadata, including provenance information. This removal would lead to the provenance being disassociated from the media asset.

> *An example solution to this problem could be the use of a perceptual hash algorithm that would enable digital content to be matched even if the underlying bits differ. Each signatory could supply their own algorithm; the details not being specified by this standard and may change over time as new algorithms are developed.*

Another option is to store either the original or a copy of the asset's metadata in a separate location from the asset itself. This would require a data store which may be publicly or privately accessible. This method could be used in conjunction with a perceptual hash approach as well.

6.2 Broken provenance

It is very common for online services and image gallery applications to re-encode images. Lossy image formats, such as JPEG, can change the bit-wise data stream value each time it is re-encoded. Provenance systems that only identify the content based on a cryptographic checksum, such as MD5, SHA1, or SHA256, will fail to identify the content if it is re-encoded. Therefore, the provenance must be updated to reflect any modification to the asset.

In addition, it is important that the entire chain can be validated. It is therefore important that each signer authenticates the prior provenance records. For example:

- Signer A validates the picture.
- Signer B denotes a handling change, validates provenance record A and incorporates that into provenance record B.
- Signer C denotes another handling change, validates provenance record B and incorporates that into provenance record C.
- Signer D denotes another handling change, validates provenance record C and incorporates that into provenance record D. As long as D can be validated, then the state of A, B, and C is valid.

A future actor who cannot validate one or more previous signers may only need to validate the final signer in order to identify the provenance. In this example, the recipient may only need to validate D in order to determine that A and B were valid at one time.

6.3 Replaced provenance

A common metadata attack replaces authoritative metadata with metadata from another source. Systems that only evaluate the metadata may not notice that the values associated with a media asset were replaced. Provenance signatures that only cover other types of metadata can be copied without detection. Accordingly, any approach needs to cover not only the metadata but also the asset's content as well.

6.4 Incomplete provenance

When adding provenance to a file, the signer receives a media asset and signs it. However, the signer does not authenticate the media asset before signing. For example, a photographer captures a picture and may perform a few touch-ups before sending the picture to a signer for signing. The signer signs what was received, but this does not authenticate the original.

6.5 Inaccessible resources

The entities that digitally sign and/or register the media asset and associated provenance are not expected to be around indefinitely. At some point the media asset, media asset metadata, media asset content or its associated registration information may no longer be able to be linked to the provenance and/or validated.

6.6 Impostor signing

It is important that an actor can determine if the signer of a media asset is trustworthy.

7 Requirements

Based on the identified use cases, a number of JPEG Fake Media requirements have been identified and organized in three main categories:

- Media creation and modification descriptions
- Metadata embedding and referencing
- Authenticity, integrity, and trust model

The sections below list the already identified requirements for each identified category.

7.1 Media creation and modification descriptions

- R1.1 The standard shall provide means to describe **how, by whom, where and/or when** the media asset was created and/or modified. For example:
    - How:
        - Natural: Sensor (e.g. camera model or specific user device)
        - Synthetic: Software used for creation or modification (name, creator, version, …)
    - Who:
        - Creator (person, just name or identifier, website, …?)
        - Social media platform (e.g. when transcoding)
        - Software that created a synthetic media asset
    - Where & when:
        - Timestamp of creation
        - Timestamp of modification
        - GPS coordinates creation
        - Other time and location information
- R1.2 The standard shall provide means to reference the asset(s) on which the modifications were applied and/or that were used for its creation.
- R1.3 The standard shall provide means to describe the **type** (for example: transcoding, contrast, brightness, colour temperature, adding annotations, …) and **category** (for example: global, local, restoration, enhancement, composition, ...) of modifications.
- R1.4 The standard shall provide means to describe the **region of interest (ROI)** where the media asset was modified.
- R1.5 The standard shall provide means to describe the **purpose of a modification**.
- R1.6 The standard shall provide means to **signal the extent of modifications** compared to a reference version, for example by providing an objective similarity metric. The standard shall also provide means to signal which method was used.

- R1.7 The standard shall provide means to describe (algorithmically or by humans) the **probability of the existence of a modification** and which method was used to determine that probability. The probability can be specific to a particular region of interest or modification.
- R1.8 The standard shall provide means to **keep track of the provenance of media assets and/or of specific modifications**.
- R1.9 The standard shall provide means to signal IPR information related to media assets and/or to specific modifications.

7.2 Metadata embedding and referencing

- R2.1 The standard shall provide means to **embed provenance, authenticity and IPR information** into media assets.
- R2.2 The standard shall **comply with the JPEG Systems framework** and should **retain backwards compatibility**.
- R2.3 The standard shall consider **privacy** of individuals and locations.
- R2.4 The standard shall support **anonymization and/or obfuscation of private information** if demanded by the user.
    - The standard shall provide means to explicitly denote anonymous, obscured, or redacted information. If the information is not provided, then it is considered anonymized.
- R2.5 The standard shall **accommodate non-JPEG formats**.
- R2.6 The standard shall be **viable as a self-contained structure.**
- R2.7 The standard shall provide means to **verify the integrity** of the media asset by supporting:
    - various hashing methods;
    - various signing methods;
    - various digital fingerprinting methods;
    - the ability to embed multiple signatures, hashes or fingerprints with different scope:
        - Ability to cover JUMBF boxes, entire metadata, subset of metadata, asset content, ...
- R2.8 The standard shall provide means to **protect** media asset metadata, including provenance information.
- R2.9 The standard shall provide means to **provide conditional access** to media asset metadata.
- R2.10 The standard shall provide means to **compress embedded descriptions**.
- R2.11 The standard shall provide means to **embed references** to externally hosted descriptions, methods and services.
- R2.12 The standard shall provide means to **keep track of modifications made to the media asset content** and provide means to compare with or rollback to a previous version.
- R2.13 The standard shall provide means to **keep track of modifications made to the media asset metadata** and provide means to compare with or rollback to a previous version.
- R2.14 The standard shall provide means to signal what should happen with **embedded JUMBF boxes in case modifications are applied**: carry over, remove, update, warn the user about potential inconsistencies. The action may depend on the type of modification and can differ depending on the type of the specific JUMBF box. For example:

- Modifications that do not impact semantics or coordinate system:
  - Transcoding
  - Modest contrast, brightness, exposure, shadows, highlights, vignetting, ...
  - Colour temperature, tint, saturation, vibrance, color curve, ...
  - Sharpening, blurring, noise reduction, ...
- Modifications that impact the coordinate system:
  - Cropping, rotating, warping, resizing, …
- Modifications that change the semantics:
  - Object removal
  - Object shape modifications
  - Composition
  - Deep fake (face swapping, …)

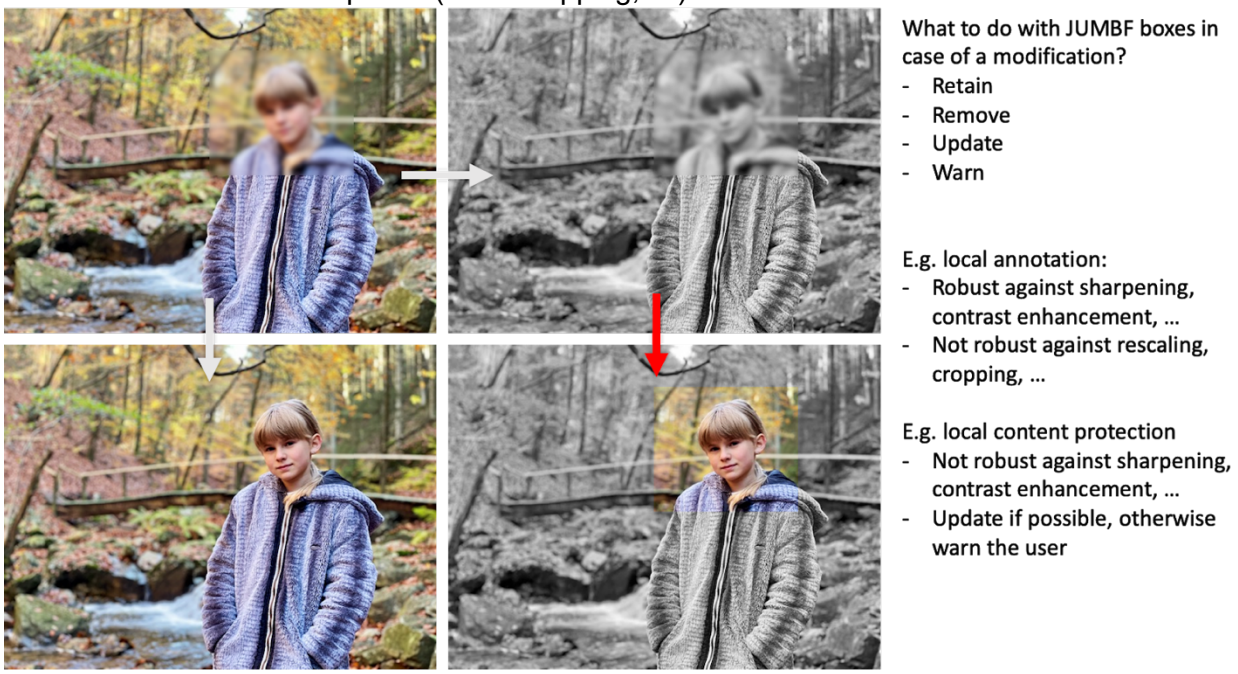

Figure 2: Illustrative example of an embedded Privacy & Security JUMBF box that needs to be updated in case a modification to the media asset content is made.

7.3 Authenticity, integrity, and trust model

- R3.1 The standard shall support means to establish and revoke trust of actors.
- R3.2 The standard shall support **digital signing** of media assets, metadata and content by actors as a means to establish authenticity and integrity.
  - The signing actor can be a capturing device, application, individual or organization.
- R3.3 The standard shall support means to **verify the authenticity** of the media asset.
  - The provenance must be updated to reflect any modification to the asset.
  - The standard shall provide means to identify if a media asset contains modifications that are missing from the asset's provenance.
- R3.4 The standard shall support verifiable **integrity** of media assets.

    - Modifications that are missing from the asset's provenance are treated the same as an invalid or unverifiable provenance chain.
- R3.5 The standard shall support **registration of media assets, media asset metadata and media asset content along with additional registration information**.
- R3.6 The standard shall support **registration of the actors** involved in the media asset creation, modifications and distribution.
- R3.7 The standard shall support both **decentralized and centralized registration solutions**.
    - The standard shall provide means to identify a **decentralized or centralized registration repository** where a media asset, media asset metadata and/or media asset content is registered even if the metadata of the asset is disassociated. Once the repository is identified, it should be possible to request the asset, metadata, content and/or additional registration information.
- R3.8 The standard shall provide means to **identify the origin, source or digital master** of the media asset while also supporting **anonymization or obfuscation** of that information if demanded by the use case.