Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 24, 2025

K. Williams
Independent Researcher

June 22, 2025

Classification and Tagging System for Digital Content to Preserve Clean Datasets for Machine Learning
draft-williams-ai-content-tagging-00

Abstract

This document specifies a classification and tagging system designed to identify and preserve the provenance of digital content (text, audio, video, and other media) to ensure the integrity of training datasets for machine learning systems. The framework described herein aims to support a standardized mechanism for tagging data with metadata that specifies whether the content was human-generated or AI-generated. This enables the exclusion of AI-generated data from training corpora where human-originated material is required, and it facilitates the maintenance of clean, verifiable sources for future AI development.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 24, 2025.

Table of Contents

1. Introduction

With the proliferation of generative AI models producing vast amounts of synthetic content, it is increasingly difficult to ensure the quality and originality of training datasets for future AI systems. This phenomenon, commonly referred to as "model collapse" or "data poisoning," occurs when models are trained on outputs of other models, compounding errors and losing alignment with human-authored knowledge and intent.

As Rear Admiral Grace Hopper stated in her 1982 NSA address, every data record must include an identifier. In keeping with this foundational principle, this RFC proposes a metadata-based classification system that can be attached to digital content at the time of its creation or publication to ensure traceability, discoverability, and reliability.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

HGC (Human-Generated Content): Data or media created entirely by human effort, without the use of AI assistance.

AIGC (AI-Generated Content): Content generated in part or in full by artificial intelligence systems.

Metadata Tag: An identifier attached to content that specifies generation origin, timestamp, authorship, and optional cryptographic signature.

Clean Dataset: A dataset composed exclusively of HGC, validated for provenance and integrity.

3. Metadata Structure

Each piece of digital content MUST be accompanied by a metadata record conforming to the following schema:

version: The version of this metadata specification.

origin: Identifies the method of content generation. Valid values are "human", "ai", or "hybrid". "Hybrid" indicates partial AI

involvement.

author: The creator or generating entity of the content.

creation_timestamp: ISO 8601 formatted timestamp of content creation.

license: The licensing terms under which the content is distributed.

checksum: SHA-256 hash of the content for integrity verification.

signature: Optional cryptographic signature for validation. RECOMMENDED for human content to validate integrity via digital signature.

toolchain: Optional field indicating the tools or AI models used in content generation.

model_identifier: Optional field specifying the specific AI model used for AIGC.

4. Implementation Mechanisms

The metadata MAY be embedded using one or more of the following methods:

* HTTP headers (e.g., X-Content-Metadata)

* Sidecar XML files for downloadable assets

* Embedded tags within HTML meta elements

* ID3v2 tags for MP3/MP4 audio files

* Exif/XMP tags for image and video files

5. Implementation Examples

5.1. HTTP Header for Human-Generated HTML Page

The following example shows the HTTP header format for a human-generated HTML page:

```
GET /article.html HTTP/1.1
Host: example.com
...
X-Content-Metadata: <metadata><version>1.0</version><origin>human</o
rigin><author>Jane Doe</author><creation_timestamp>2025-06-21T08:00
:00Z</creation_timestamp><license>CC-BY-4.0</license><checksum>3d2e
61...</checksum><signature>BASE64_PGP_SIGNATURE</signature></metada
ta>
```

5.2. Sidecar File for Image

For an image file named "sunrise.jpg", the accompanying sidecar metadata file would be "sunrise.jpg.meta.xml":

```
<metadata>
  <version>1.0</version>
```

```
  <origin>ai</origin>
  <author>AI Art Generator</author>
  <creation_timestamp>2025-06-20T19:23:00Z</creation_timestamp>
  <checksum>abc456...</checksum>
  <toolchain>StableDiffusion-v3</toolchain>
  <model_identifier>sd-v3.1</model_identifier>
</metadata>
```

5.3. HTML Meta Tags

The following example shows HTML meta tag implementation:

```
<meta name="X-Content-Origin" content="human">
<meta name="X-Content-Author" content="John Smith">
<meta name="X-Content-Timestamp" content="2025-06-22T12:34:00Z">
<meta name="X-Content-Signature" content="BASE64_SIGNATURE">
```

5.4. Audio File with ID3 Tags

For audio files, the following ID3v2 tags SHOULD be used:

* TXXX:Content-Origin = "human"
* TXXX:Content-Author = "Podcast Host"
* TXXX:Toolchain = (not present for human-generated content)
* TXXX:Checksum = "8f3a2b..."

6. Trust and Verification

Verification SHOULD be achieved through the following mechanisms:

* Public key infrastructure (PKI) for digital signature validation
* Trust registries for known human content sources
* Canonical checksums published to public hash registries

Content consumers SHOULD verify signatures and checksums when available to ensure the integrity and authenticity of metadata claims.

7. Application to Machine Learning Pipelines

Dataset construction pipelines MUST implement the following requirements:

* Filter and exclude content with origin values of "ai" or "hybrid" where HGC-only datasets are mandated
* Retain metadata lineage for dataset auditing and transparency
* Support backtracking of samples to their original signed content
* Maintain logs of all filtering decisions for compliance and auditing purposes

8. Security Considerations

Several security considerations apply to this specification:

* Malicious actors may attempt to falsify metadata. Digital signatures, cryptographic checksums, and origin auditing are necessary to prevent tampering.

* Repositories SHOULD maintain audit logs of metadata ingestion and content origin verification.

* The integrity of the metadata itself must be protected through cryptographic means where authenticity is critical.

* Trust anchors and certificate authorities used for signature verification must be carefully managed and regularly audited.

9. Privacy Considerations

Privacy protections must be considered in the implementation of this system:

* Personally identifiable information (PII) SHOULD NOT be embedded in metadata.

* Where authorship must be asserted, anonymized or pseudonymous cryptographic identities SHOULD be used.

* Content creators must have control over what identifying information is included in metadata tags.

10. IANA Considerations

This document requests the following IANA actions:

10.1. New MIME Type Registration

This document requests the creation of a new metadata MIME type:

Type name: application
Subtype name: x-content-metadata+xml
Required parameters: None
Optional parameters: charset
Encoding considerations: UTF-8 encoding is RECOMMENDED
Security considerations: See Section 8 of this document
Interoperability considerations: None known
Published specification: This document
Applications that use this media type: Content management systems, digital archives, machine learning platforms
Additional information: None
Person and email address to contact for further information: Keenan Williams <keenanwilliams@gmail.com>
Intended usage: COMMON
Restrictions on usage: None
Author: Keenan Williams
Change controller: IETF

10.2. HTTP Header Field Registration

This document requests the registration of the following HTTP header fields:

* X-Content-Origin

* X-Content-Signature

* X-Content-Toolchain

These header fields are intended for use in identifying content provenance as specified in this document.

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

11.2. Informative References

[HOPPER] Hopper, G., "Future Possibilities: Data, Hardware, Software, and People", NSA Lecture Series, 1982, <https://www.youtube.com/watch?v=si9iqF5uTFk>.

[ISO8601] International Organization for Standardization, "Date and time -- Representations for information interchange -- Part 1: Basic rules", ISO 8601-1:2019, February 2019.

[NIST-FIPS-180-4] National Institute of Standards and Technology, "Secure Hash Standard (SHS)", FIPS PUB 180-4, DOI 10.6028/NIST.FIPS.180-4, August 2015, <https://doi.org/10.6028/NIST.FIPS.180-4>.

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, DOI 10.17487/RFC3986, January 2005, <https://www.rfc-editor.org/info/rfc3986>.

Author’s Address

Keenan Williams
Independent Researcher

Email: keenanwilliams@gmail.com