TAXONOMIES

The goal of the Trust in Archives Initiative (TAI) Taxonomies subgroup was to survey existing taxonomies and taxonomic guidance in order to describe AI generated or edited media (AI media). After surveying, the subgroup made a list of recommended taxonomic fields for cataloguing AI media. The IPTC Photo Metadata Working Group's recommendations, first released for public comment on Aug 1, 2025, were a major jumping-off point. The goal was not to create a technical manual or a format-specific guide but a general guide that reflects emerging standards for describing AI media.

SURVEYED TAXONOMIES AND TAXONOMIC GUIDANCE

Members of the subcommittee sought examples of metadata for AI media by reaching out to their professional networks, including open requests for input from the AMIA email list, and multiple metadata communities on Slack. The standards that came to light span multiple formats and vary in technical specificity.

- IPTC Photo Metadata Working Group
- Program for Cooperative Cataloging (PCC) Standing Committee on Standards
- AI4LAM Speech-to-Text Working Group (Transcript Provenance Metadata Elements)
- Coalition for Content Provenance and Authenticity (C2PA)
- Various open linked data ontologies

SYNTHESIS OF SURVEYED TAXONOMIES

Members of the subcommittee considered the following questions when synthesizing the surveyed taxonomies:

- What metadata is unique to AI media? What information should be captured for this media that is not necessary for other media?
    - This encompasses technical, provenance, and ethical concerns.
- What information is necessary to determine the veracity of AI media?
    - Authentication remains an evolving issue. IPTC and C2PA overlap in important ways, but they are not interchangeable, and adoption varies across vendors, platforms, and workflows. As a result, no single current standard guarantees complete provenance in every case. Cataloguers should therefore treat provenance metadata as partial evidence that must be interpreted in context.
- Who is doing this cataloguing and when is it happening?
    - The members of TAI come from diverse fields, including non-profit libraries and archives, media production, and stock footage libraries. The person cataloging the media might be its creator or a downstream repository. The media might have

known or unknown provenance. A cataloguer receiving the media may not be able to complete some of the fields recommended by TAI or may need to request specific metadata from the source of the media.

- Where does the metadata come from?
    - In some cases, the metadata recommended in this guide will not be available even when provenance is known. Much of the metadata described here must be generated at the point of creation and cannot be entirely reproduced retrospectively. Documentation practices vary widely: some model developers publish model cards, system cards, dataset documentation, or provenance manifests, while others provide only partial disclosures or none at all. Cataloguers should therefore expect uneven metadata availability and record both what is known and what could not be verified.
    - Institutions that need to catalog AI media will need to become advocates for metadata access and transparency. In some cases, there will be no leverage to request transparency from AI companies; in others, the brokering of archival media from institutions to AI companies that need large data sets may be a point of leverage. In this developing field, libraries, archives, media producers, and others can advocate for access to the metadata they need to document AI media.
    - The format of the metadata will vary depending on whether it was recorded at the point of creation or retroactively described. It can be useful in either case, but may lack uniformity.
    - Even when provenance is partly known, AI-related metadata may be incomplete, inconsistently recorded, or unavailable to downstream repositories. Some workflows preserve detailed technical records, while others leave only partial traces. Cataloguers may therefore need to distinguish between metadata captured at the point of creation and metadata reconstructed later from surrounding documentation.

SOURCES OF AI METADATA

- A model card is a document that accompanies an AI model and describes what the model is, how it was trained or evaluated, its intended uses, known limitations, and relevant ethical or bias considerations. When available, model cards can provide a useful starting point for archival description because they may identify the model, summarize training or

evaluation context, and document important constraints on interpretation and reuse. For a widely used introduction, see Mitchell et al., *Model Cards for Model Reporting*: https://arxiv.org/abs/1810.03993

Model cards are not consistently available, however, and their level of detail varies widely. Some open models are accompanied by substantial documentation, while many commercial systems provide only partial disclosures or no model card at all. For this reason, cataloguers should treat model cards as one possible source of evidence rather than as a complete or uniform record.

- Some relevant metadata may be embedded in the media itself or preserved in sidecar data, such as IPTC/XMP fields, C2PA manifests, timestamps, or exported workflow files. Other important information, including prompt text, prompt authorship, training-data disclosures, or internal processing notes, often must be requested directly from the creator, production team, or platform.
- When exact technical metadata is unavailable, cataloguers may need to combine creator-supplied free-text descriptions with controlled vocabulary and local notes. In practice, requests should focus on the records most useful for establishing provenance: model name and version, prompt text, prompt writer, reference media, workflow or processing steps, date generated, and any available model card or provenance manifest. These details may be requested through deposit forms, acquisition agreements, creator questionnaires, or direct follow-up with the depositor or vendor.

TAI AI TAXONOMY

Standards Comparison Grid

AI MODEL

Definition: The foundational model name and version used to generate this media.

Example: "Wan2.2-T2V-A14B-Diffusers Text-to-Video"; "CogVideoX-2B"

Source: Model cards, system documentation, creator-supplied workflow files, platform job history, API logs, provenance manifests, or direct confirmation from the creator or vendor.

- Hugging Face `Model Cards` docs: https://huggingface.co/docs/hub/en/model-cards
- Mitchell et al., *Model Cards for Model Reporting*: https://arxiv.org/abs/1810.03993

TRAINING DATA

Definition: The collection of data used to train an AI/ML model, shaping its capabilities and biases.

Example: A dataset of newsreels containing war footage.

Source: Model cards, technical documentation, research papers, vendor disclosures, or creator-supplied notes. In many cases, training-data information will be partial, generalized, or unavailable and should be recorded with appropriate caution.

AI TEXT PROMPT DESCRIPTION

Definition: The text instructions or other human-authored inputs provided to an AI system in order to generate, transform, or modify the media.

Example: “Enhance film grain, preserve original texture, upscale to 4K.”

Source: Source: Creator-provided prompt logs, workflow or project files, API request history, platform job exports, or written documentation supplied by the depositor. If the prompt is not embedded in the asset or a provenance manifest, it usually must be requested directly.

AI PROMPT WRITER NAME

Definition: Name of the person who wrote the prompt used for generating this media.

Example: "Kevina Tidwell"; "Unknown prompt writer"

Note: May not always be known or necessary to capture.

Source:

REFERENCE MEDIA

Definition: Media supplied to an AI system as an input, source, or base from which new or modified media was generated.

Example: "Getty Images 2218833057"

Source: The source media itself, asset management records, project files, edit decision lists, workflow inputs, or creator-supplied documentation identifying the base or input media.

AI WORKFLOW

Definition: The documented sequence of AI-driven steps, parameters, and tools applied to archival material. It describes how content was processed, from input to output.

Example: An Upscaling Workflow using ComfyUI, that can be easily replicated using the final output with the workflows embedded into a JSON file.

- See https://dripart.mintlify.app/tutorials/basic/upscale.
- Preserving Intent in Nonfiction Media: A Responsible Approach to AI Enhancement | Topaz Labs

Source: Exported workflow files (for example, ComfyUI JSON), project files, scripts, API logs, processing notes, standard operating procedures, or creator/vendor documentation describing the steps and settings used.

DATE GENERATED

Definition: The date on which the media was generated, materially modified, or output by an AI-assisted workflow.

Example: 2025-01-12

Source: Embedded file metadata, provenance manifests, platform job history, project files, or creator-supplied production records. If only an approximate date is known, that uncertainty should be recorded in a note.

NEXT STEPS

Further work should be done to guide cataloguers on what metadata should be recorded, where to get it, and how to ask for it from AI media creators and AI companies, whether for individual media objects or for collections acquired in bulk.