Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Shoebox to Toolbox: Archival Affirmation in the AI-Assisted Processing of a Media-Art Archive

Selcuk Artut

sartut@sabanciuniv.edu

Sabancı Üniversitesi

Research Article

Keywords: archival description, artificial intelligence, archival affirmation, original order, controlled vocabularies, media-art archives

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11562692/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Large digitised media collections often remain unprocessed and unexplored, despite being technically preserved, and archival institutions increasingly turn to artificial intelligence (AI). Rather than pursuing autonomous archival systems, this article examines how AI can support archival description and classification while preserving the contextual, historical and provenance-related integrity of archival materials. Drawing on the archive of media artist Teoman Madra (1931–2025), the study develops and evaluates an AI-assisted archival tool for a large, initially unstructured collection. The proposed approach establishes an iterative, human-in-the-loop workflow in which AI-generated suggestions are critically assessed by the archivist, whose decisions subsequently inform the system's outputs. The study shows that the controlled vocabulary given to the model shapes what it can recognise, that the archivist's validated decisions guide classification reliably within an archival unit but not across units, and that the model's contribution declines as tasks demand contextual knowledge beyond what is visible. While this process can improve efficiency and reveal details a person might miss, it also introduces epistemic risks, particularly when plausible but inaccurate classifications remain undetected. Faster processing does not by itself mean more reliable description. The article conceptualises AI-assisted archival classification as archival affirmation, a collaborative process in which human judgement remains integral to the production and validation of archival knowledge. By situating the archivist's interpretive authority at the centre of this process, the article contributes to discussions of the responsible integration, critical oversight and governance of AI in archival practice.

1 Introduction

When an artist has maintained no systematic documentation of their work, as is often the case with media artists such as Teoman Madra (1931–2025), the archival process frequently begins with fragmented and unstructured materials, a shoebox of photographic negatives, unlabelled magnetic tapes, optical discs and envelopes bearing occasional handwritten annotations. The challenge extends beyond artists' archives: as recording and storage technologies have multiplied, from analogue photography and videotape to digital media, archival collections have grown substantially in volume and heterogeneity, complicating their organisation, interpretation and accessibility.

Contemporary archives increasingly comprise both born-digital materials and digitised analogue records, often amounting to tens of thousands of individual files. Although digitisation facilitates preservation and retrieval, it does not carry over the contextual information required to interpret the materials: collections that held whatever physical order they had may become extensive directories of files identified only by numerical or automatically generated filenames. Materials thus become accessible at a technical level while their intellectual accessibility remains limited, since effective searching presupposes descriptive metadata that is frequently incomplete or absent. Archival description therefore emerges as a critical bottleneck in the management of large-scale media collections, motivating growing interest in artificial intelligence (AI) to support archival processing, classification and contextual interpretation.

Vision-language models can generate descriptive accounts of still and moving images and assign predefined classification terms, while speech recognition models can transcribe audio recordings. Rather than pursuing the development of fully autonomous archival systems, whose limitations are demonstrated in this study, the article examines how AI can support archival description and classification without compromising the contextual, historical and provenance-related integrity of archival materials. Such considerations extend beyond the identification of visible content to questions that require archival and historical knowledge, including the series to which a photograph belongs, the occasion documented by a particular recording and the identities of individuals depicted.

The article argues that AI-assisted archival classification should be understood as a collaborative process in which human judgement remains essential to the interpretation, production and validation of archival knowledge. AI-assisted workflows accelerated archival processing and, in certain instances, identified visual details that might otherwise escape human observation, but they also introduced epistemic risks through plausible yet inaccurate classifications that may remain undetected; increased efficiency does not necessarily mean greater descriptive reliability. Model performance declined as tasks demanded knowledge beyond what is visible, from image orientation and content type to identifying individuals, reconstructing historical circumstances and establishing relationships between archival objects. Human expertise is therefore not merely a mechanism of verification but a constitutive element of responsible AI-assisted archival practice.

The study makes three further contributions. First, it demonstrates that classification can improve through the refinement of the surrounding archival workflow rather than through modifications to the underlying models. Without retraining or fine-tuning, improvements were achieved through a controlled vocabulary, classification criteria informed by the archivist's decisions, systematic review procedures and previously validated decisions used as reference examples. Where Rolan et al. (2019, pp 181–183) classify AI systems by whether archival expertise is embodied in them as rules or as training data, here it is embodied in the workflow around the models.

Second, the findings reaffirm the significance of provenance and original order in AI-assisted archival classification. Content-based indexing has been proposed as a way of organising and accessing archives beyond provenance and original order, although the same authors suggest that these principles may "find new life in a datafied world" (Colavizza et al. 2022, Discussion). The results support the latter view for archival description: contextual relationships remain essential to reliable classification. Decisions made by the archivist could be effectively applied to other materials within the same archival unit (a group of materials found stored together), whereas their applicability diminished substantially across different units. This suggests that the physical and contextual relationships embedded in archival groupings retain considerable interpretive value, even when the materials are processed through computational methods.

Third, the study identifies AI as a methodological instrument not only for processing archival materials but also for developing the archival workflow itself. Confronted with a large-scale, initially unstructured

collection, the archivist employed iterative experimentation with AI to explore possible approaches, evaluate their limitations and progressively establish a systematic methodology for organising and describing the materials.

The research draws on the archive of Teoman Madra, a pioneering figure in Turkish media art, and was conducted by the author, who has acted as both researcher and archivist since the archival work began in 2021. Section 2 reviews scholarship on AI and archival practice; Sections 3–5 present the archive, the method and the findings; Section 6 discusses original order, controlled vocabularies, methodological experimentation, automation bias and privacy; and Section 7 concludes.

2 Artificial intelligence and archival description

2.1 Artificial intelligence in archives

Early case studies of AI in archives, such as the four initiatives reviewed by Rolan et al. (2019), found it promising but "certainly not a production-ready 'silver bullet'" (p 195); later work has mapped a growing practice. Colavizza et al. (2022) identify automated appraisal, metadata creation and sensitivity review as the main areas of use. They note practical results alongside "a still limited reflexive attitude towards AI", in which biases and limitations go largely undiscussed. They also argue that content-based indexing of people, places and keywords lets users search across fonds, beyond the order given by provenance, while provenance itself "needs specific techniques to be captured and maintained". In a systematic review of work from 2019 to 2023, Shinde et al. (2024) find retrieval and classification most common, and few studies of arrangement and description.

Institutional reports frame these issues as questions of responsibility. Padilla (2019, pp 10–12) proposes a "track changes" record of description, whether human or algorithmic, and warns that machine descriptions, like human ones, carry varying certainty; Cordell (2020, p 12) asks libraries to lead on the ethics of machine learning. A call for collaboration between archivists and computer scientists recurs across this literature (Padilla 2019, pp 19–20; Cordell 2020, p 5; Jaillant and Caputo 2022, p 825; Shinde et al. 2024, s 7). This article answers that call within the Teoman Madra archive, where the archivist works directly with tools programmed for this archive on open models rather than commercial software, so that the effect of every decision can be traced and kept under the archivist's control.

2.2 Description, classification and the archivist’s affirmation

Archival description records not only what a record shows but how it came to be: who made it, in which activity, in relation to which other records. The Records in Contexts conceptual model (RiC-CM) of the International Council on Archives' Expert Group on Archival Description makes these relations explicit, describing record resources together with the agents and activities that produced them (EGAD 2023, s 1.2, s 1.5.3). By contrast, the vision-language model used in this study sees only an instantiation of each record, one image at a time; what it cannot see is largely what RiC-CM treats as context: the agents who

made or kept the record, the activity it documents, its date and place, and its relations to other records (EGAD 2023, s 2.1). In this archive, that is who made a photograph, a video or a digital work, on which occasion, and what else it was kept with.

Beyond what the model can see lies a second problem: the categories it is asked to apply are never neutral. Bowker and Star (1999, p 5) show that categories carry the values and blind spots of those who make them. Crawford and Paglen (2019) trace how the person categories of a widely used image dataset reproduce prejudice that no technical adjustment removes, and Crawford (2021, p 135) argues more broadly that every training dataset "contains a worldview". Archival practitioners place the same risk "at the stage of labelling and classification" (Jaillant et al. 2026, p 13). For archival description these studies point to the same design choice: the vocabulary should belong to the archive, its terms should be defined by the archivist, and a model should only choose from it, never write its own categories into the record. Jaillant et al. (2026, p 26) reach a similar recommendation: constrained captions that describe only what is visible and state their uncertainty.

Computer vision applied to collections has mostly served exploration: Arnold and Tilton's distant viewing (2023) sees many images at once while insisting that annotations never capture a whole image. This article asks a narrower archival question: when an annotation enters the catalogue, who is answerable for it, and how is that recorded?

Keeping the archivist in the loop is the usual answer to that question, but it carries its own risk. Automation bias is the tendency of people using decision aids to accept their errors and to miss what they fail to flag. It has been documented in experts as well as novices, and training does not remove it (Parasuraman and Manzey 2010, pp 381, 391). A method that asks an archivist to confirm machine proposals must therefore measure the proposals independently of the archivist's acceptance of them. Section 5 reports such checks wherever acceptance was nearly complete.

This article calls the alternative archival affirmation. Affirmation is more than the confirmation of machine outputs, the role that human-in-the-loop designs usually assign to people. It is the iterative process by which the archivist sets and defines the terms a model may use, weighs the contextual evidence the model cannot see, validates or corrects each proposal and records the decision with its provenance. A machine observation becomes archival description only through this accountable act. Affirmation does not make the archivist infallible: the archivist's categories carry assumptions of their own (Bowker and Star 1999, p 5). It is understood here as accountable and revisable judgement, documented so that it can be questioned and changed.

3 The Teoman Madra archive

Teoman Madra (1931–2025) was a pioneering figure in the development of media art in Turkey, whose artistic practice encompassed experimental photography, light, sound, video and computer-based technologies. Spanning more than five decades, from the 1960s to the 2010s, his body of work reflects a

sustained engagement with evolving technological media, tracing a trajectory from analogue experimentation to early forms of digital artistic production.

These technological transitions are reflected in the composition and material diversity of his archive. Between the 1960s and 1980s, Madra worked predominantly with analogue photographic technologies. In the late 1980s, he began experimenting with the Amiga computer, producing digital artworks using the graphics software Deluxe Paint. These computer-generated works were subsequently transferred to analogue video formats, including VHS, Betamax and Hi8, and later to digital videotape formats such as MiniDV. Consequently, the archive encompasses both born-digital works and digitised analogue materials, alongside digital content preserved on analogue carriers. The collection comprises approximately 20,000 photographic slides, negatives and prints; 600 hours of video recordings; four hard drives; and 1,700 CD-ROMs. This heterogeneous assemblage of media formats and storage technologies illustrates the archival complexities that arise when artistic practices evolve across successive technological generations.

At the outset of the archival process, the collection lacked a systematic classification structure. Photographic rolls, prints and various media carriers were stored in envelopes or plastic bags, often accompanied by little or no descriptive information. During digitisation, the materials were organised according to their physical containers. Each archival unit comprises the materials found stored together, such as the negatives of several rolls in one envelope or bag, a single videotape or an optical disc, and was assigned a separate digital folder. These units largely preserved the original physical groupings of the materials, regardless of whether such arrangements reflected deliberate decisions by the artist or emerged through everyday storage practices. They are therefore treated as inherited physical groupings, which carry local context whether or not they record the artist's own arrangement.

To maintain the relationship between digitised materials and their physical originals, the archivist introduced a consistent identification system consisting of a carrier-related prefix (such as NEGA for photographic materials, VHS for videotapes and BETA for Betamax recordings) followed by a numerical identifier. This system enables individual digital files to be traced to their corresponding physical sources, thereby supporting subsequent examination, verification and preservation activities.

In addition to the digitised media content, each archival unit includes photographic documentation of any accompanying handwritten or printed annotations, such as carrier labels, protective sleeves and information cards. These supplementary records are treated as a distinct source of contextual evidence within the proposed methodology, providing information that may not be recoverable through automated analysis of the media content alone. Once confirmed by the archivist, these records are attached to every item of their unit as context but are not shown to the model, so that the machine reading stays independent of what the archive already knows.

Three characteristics make the archive a demanding case for automated description. First, it mixes artworks with their documentation: a print of a light drawing (a long exposure of a light moved in time with music) is itself a work, while a photograph of an installation only records one, yet the two can look

alike. Second, its subjects range from exhibitions and theatre to family life, so that descriptive categories such as performance, event and portrait overlap. Third, much of its context lies outside the material itself: in the artist's published chronology, in the notes found with the material and in the archivist's own knowledge.

4 Method

The study draws on four sets of material, summarised in Table 1. The evaluation focuses on the still images; the videotapes, including the born-digital works, were processed with the same pipeline but are discussed only briefly here.

Table 1
Material used in the study

Set	Content	Use
Development sample	179 still images from five units; 2,958 tape segments	Building the pipeline; full review by the archivist; six controlled re-runs
Five unseen units	253 images from five units, chosen to differ from the sample (theatre, social occasions, portraits, objects)	Transfer test and review-loop test
Full archive	26,157 images in 658 units	One machine reading; two review rounds; orientation, unit proposals and faces
Artist's chronology	91 dated entries (exhibitions, performances, biography)	Linking images to occasions

4.1 Three layers and their provenance

The method treats a machine description as a first reading that the archivist confirms, corrects and completes, not as a catalogue entry in its own right. Every record keeps three layers apart:

1. **Machine reading**: what the model proposes, such as a description, visible text, terms and a count of people, recorded together with the model that produced it.
2. **Archivist's correction**: changes to the machine reading, such as corrected terms, description or orientation, with the model's value kept beside each correction.
3. **Archivist's context**: what the material itself cannot show, such as the series a work belongs to, the occasion it records (linked to the artist's chronology), the people in it and inscriptions the model missed.

Each descriptive value within an archival record, including individual classification terms, retains its provenance, enabling users to distinguish between information proposed by an AI model, validated or modified by the archivist and derived from existing archival knowledge. This approach ensures

transparency and traceability throughout the descriptive process. Contextual information is recorded at the level of granularity at which it can be reliably established. For example, information concerning an artistic series or a documented occasion often applies to an entire archival unit or visually related group of items. Such information is therefore recorded once at the appropriate level and subsequently associated with all relevant items, ensuring descriptive consistency while avoiding unnecessary duplication.

4.2 Local models and their selection

The processing pipeline operates entirely on a single laptop equipped with an NVIDIA GeForce RTX 5090 Laptop GPU (24 GB VRAM), ensuring that archival materials remain within a local computing environment throughout processing. This configuration is particularly significant for a privately held archive containing personal materials and works with unresolved intellectual property rights. Previous research has identified the use of cloud-based AI services as a potential risk to personal privacy and the legal ownership of archival records, motivating the adoption of locally deployed systems (Rolan et al. 2019, pp 191, 196). All models used are therefore openly available and run on this machine, so that no archival material left the archive's computer.

Within these constraints, each model was chosen for its task. Qwen2.5-VL-7B-Instruct (Bai et al. 2025) describes still images and video frames, assigns terms from the controlled vocabulary and transcribes carrier annotations. It was selected because it reads text in images well, follows an instruction to choose only from a given list of terms and handles Turkish as well as English, while fitting within the 24 GB of GPU memory. Whisper large-v3 (Radford et al. 2023) transcribes speech; its multilingual training suits recordings in Turkish and English, although its known tendency to invent text over music or silence required such passages to be flagged. SigLIP (Zhai et al. 2023), in its base version, produces image embeddings, referred to throughout the study as fingerprints, which are used for visual groups, reference examples and orientation; it is compact enough to fingerprint the whole collection in a short time. An Audio Spectrogram Transformer trained on AudioSet (Gong et al. 2021) classifies the sound of the recordings as speech, music, electronic sound, silence or noise. YuNet (Wu et al. 2023) and SFace (Zhong et al. 2021), distributed with the OpenCV library, detect and compare faces; they are small, run on an ordinary processor and carry permissive licences. Separately from these models, the pipeline and the search tool were developed in an iterative dialogue with an AI assistant (Claude, Anthropic): the author set out the intended system, tested each version on the archive and directed its revision, while the assistant wrote and debugged the code. The assistant was also used to run the analyses and to draft and edit the text; all review decisions, vocabulary terms, definitions and interpretations are the author's own.

All models operate with fixed parameters, without additional training or fine-tuning; the orientation classifier described in Section 4.6 is the only component trained, and only on the archivist's decisions. Although fixed model weights contribute to methodological consistency, they do not independently guarantee deterministic outputs, which may also depend on inference settings, software versions and hardware configurations. In this study, the vision-language model was set always to choose its most

likely answer rather than sampling among alternatives (greedy decoding). Under this configuration, with bfloat16 precision, PyTorch 2.11, CUDA 12.8 and Transformers 5.19 on the same machine, repeated analysis of 100 images yielded identical content-type classifications for all images and word-for-word identical descriptions for all 95 comparable responses (Section 5.2). Claims of reproducibility are therefore restricted to the tested configuration. Maintaining fixed model parameters also makes it possible to attribute changes in descriptive outcomes to adjustments in the surrounding archival workflow, including archivist decisions, controlled vocabularies and review procedures, rather than to modifications of the models themselves. The methodological transferability of the approach derives from the combination of general-purpose models deployed without further adaptation and archival procedures tailored to the characteristics and requirements of individual collections.

The pipeline comprises a sequence of processing stages (Fig. 1) designed to generate structured descriptive metadata while preserving the integrity of the original materials. First, all files are inventoried and their checksums calculated to enable integrity verification. Accompanying carrier annotations are subsequently transcribed, while still images are encoded as visual fingerprints and grouped according to their similarities before being described and classified using a controlled vocabulary. Audiovisual recordings are segmented into shots with PySceneDetect, an open-source program that detects cuts by comparing successive frames without machine learning, to facilitate the successive analysis of audio signals, speech and visual frames. Each processing stage produces independently accessible output files and a dedicated review interface through which the archivist can examine individual items alongside their corresponding machine-generated descriptions and classifications. Throughout the workflow, archival master files remain unmodified, maintaining a clear separation between preserved source materials and the descriptive metadata generated through AI-assisted processing.

4.3 Controlled vocabulary and versioning

The classification framework employs a controlled vocabulary organised into five facets: content type, depicted technology, visual characteristics, language and sound. Each term has an English and a Turkish label and a concise definition of its meaning and scope. Restricting the model to these terms keeps machine-generated classifications directly searchable and comparable with the archivist's assessments, avoiding the later reconciliation that unrestricted tagging requires.

The vocabulary was refined through review. Where existing terms proved insufficient, the archivist introduced new categories, including architectural, experimental and casual photography; the model was required to answer with term names rather than codes; and where recurring errors revealed ambiguous boundaries, definitions were formulated from the archivist's decisions and approved by the archivist. Artwork documentation, for example, was defined as "a photograph of an existing work, not a photograph that is itself the work".

This process produced two related but distinct vocabularies: the **archivist's vocabulary**, which expands whenever review identifies a need for a new term, and the **model's vocabulary**, the subset included in the

model's classification prompt. The distinction is necessary because the model reads the vocabulary as an interconnected whole: changing one term or definition can alter classifications in categories that were not themselves changed (Section 5.2). New terms are therefore designated *archivist only* by default. The model's vocabulary changes only through explicit versioning: each proposed version is tested by reprocessing a previously reviewed set of items against the archivist's validated decisions before it is adopted.

4.4 Review: the tool, the change log and review rounds

After the initial automated processing, review moved into a locally hosted search and review interface in Turkish and English. The archivist holds full administrative rights; editor and viewer roles exist for future collaborative work. Every modification is recorded in an audit log with three values: the model-generated value, the value before the change and the value after it. The log, not the search database, is the authoritative record: the database is rebuilt from the pipeline's outputs and the logged modifications are reapplied, so that later automated runs never overwrite validated human decisions and every change remains traceable and reversible.

To reduce selection bias, further review is organised into sampling rounds rather than left to browsing, which tends to follow the archivist's interests and favour frequent categories. Each round is a reproducible, seeded random sample of unreviewed images, assessed one by one. Balanced rounds draw roughly equal numbers from each content type the model assigned, so that rare types are seen; simple random rounds preserve the collection's proportions. Both take at most one image per visual group and spread their selections across units. Because balanced rounds overrepresent rare types, their overall estimate is a weighted average of the type-specific results, weighted by each type's size in the collection; 95% confidence intervals are obtained by bootstrap resampling of the judged images within each content type. Resampling whole units instead, to allow for clustering within units, gives a comparable interval (61–78% against 61–77%).

4.5 Proposals from the archivist's decisions

Certain distinctions proved difficult to express through textual definitions alone, notably that between a photograph that is itself an artwork and a photograph documenting an existing work. The method therefore uses the archivist's validated decisions as reference examples. For each unreviewed image, the five most visually similar reviewed images vote on its content type and artistic series, each vote weighted by its similarity. A proposal is presented only when at least 80% of the weighted vote agrees and the nearest reviewed image reaches a similarity of at least 0.80. The second threshold was introduced after a unanimous vote among visually distant images produced an incorrect proposal on the first unseen unit; together, the two criteria account for both agreement and visual proximity. Series are only ever proposed, never assigned automatically, since visually similar abstract compositions may belong to different series.

Following the transfer test (Section 5.3), which showed that decisions from one unit could not be reliably generalised to others, proposals were restricted to the unit. In a **unit round**, about 20% of a unit's images,

and at least ten, are sampled at regular intervals across it, with no more than one per visual group. Once the archivist has classified these, the remaining images of the unit receive proposals voted only by reviewed images from the same unit, under the same thresholds. The archivist deselects incorrect proposals and applies the rest collectively; images without a sufficiently confident proposal are reviewed individually. Accepted and rejected proposals are both recorded, preserving a traceable account of the decisions and allowing the mechanism itself to be evaluated (Fig. 2).

4.6 Orientation, people and faces

Orientation is handled separately from the vision-language model, whose own orientation judgements proved unreliable (Section 5.5). Each image preview is embedded with SigLIP in four rotations (0°, 90°, 180° and 270°), and a lightweight linear classifier, trained on the archivist's validated rotation decisions, compares the four jointly to identify the upright one. Experimental photographs without an unambiguous orientation are excluded from training. This classifier is the framework's only trained component, and it learns solely from the archivist's decisions, leaving the pretrained models unchanged. Proposals whose classifier-assigned probability for the chosen rotation is at least 0.6 are presented unit by unit, already rotated, for the archivist to accept or reject.

Naming individuals remains exclusively the archivist's task. The vision-language model is instructed not to identify anyone and describes people generically, as "a man" or "a woman". Faces are detected in the images, oriented as the archivist decided, and embedded for comparison. Faces of sufficient size and quality are grouped greedily: each joins the most similar existing group if its similarity to the group's centre is at least 0.45, and otherwise starts a new one, a threshold chosen against identities established by the archivist.

Groups are provisional, not identifications. In the review interface the archivist deselects faces of other people, assigns a name and confirms the rest; the name is then added to each selected image as a logged modification, and excluded faces are recorded as not that person. Unnamed faces are proposed for a named person when at least 0.50 similar to one of that person's faces (0.55 for small or ambiguous faces), always subject to the archivist's verification.

Names can also be assigned while reviewing a single photograph. When the archivist enters a name, the system asks which face it belongs to and retrieves similar faces, first from the same unit and then from the rest of the archive; if no face was detected, it offers the unit's unnamed faces, nearest frames first. Facial similarity, archival proximity and human verification thus work together, while computational proposals remain distinct from identities the archivist has validated.

4.7 Evaluation design

The primary evaluation measure was agreement: the share of evaluated images whose model-generated content type exactly matched the archivist's final validated classification. This is an exact-match rate, not a measure of inter-rater reliability such as Cohen's kappa. It treats the archivist's decisions as the

reference standard and shows how closely the model reproduces them, not that the classifications are historically or archivally correct. This standard rests on the archivist's long familiarity with the collection and with the artist's work, circumstances and circle, knowledge that neither the model nor an outside reviewer could bring; that it is a single person's judgement is a limitation discussed in Section 7. Images the archivist could not classify at first were set aside and decided before scoring, so every evaluated image has a decision. Within the development sample, modifications to the classification framework were assessed through controlled re-runs, each changing a single variable against the baseline reviewed by the archivist (Run 1): terms added during review (Run 2), definitions for three terms (Run 3), term names in place of codes (Run 4), a definition separating photographs that are artworks from photographs documenting works (Run 5) and refined definitions for social-activity categories (Run 6).

Proposals derived from previously reviewed examples were evaluated by leave-one-out prediction and, more stringently, with the target image's entire visual group withheld, to test generalisation beyond near-duplicate images. Transfer to new material was examined in a review-loop experiment on five unseen units, with reference sets growing from the development sample alone, to examples from the other unseen units, to evenly distributed selections from the target unit itself (5–20 images, approximately 10–50% of its contents) and finally the rest of that unit. Human effort was measured as the number of images still requiring an individual decision by the archivist.

Vocabulary versions were evaluated by reprocessing the first review round with each candidate alongside an unchanged control configuration and comparing both with the archivist's validated decisions. The orientation classifier was tested on a review round excluded from its training data and by withholding entire units from training in turn (leave-one-unit-out). Finally, unit-based proposals were assessed retrospectively on four fully reviewed units: an evenly distributed 20% of each unit was treated as reviewed, the proposals for the remaining images were scored against the archivist's existing classifications, and the replay was repeated for 20 evenly spaced starting positions and averaged.

5 Results

5.1 Vocabulary refinement and visually identifiable categories

The initial evaluation of the development sample revealed that most classification errors stemmed from limitations in the controlled vocabulary rather than failures of visual interpretation. The archivist revised the content-type classifications of 128 out of 179 images, of which 106 required terms that were absent from the original vocabulary. In the absence of appropriate categories, the model tended to assign the closest available general terms, most frequently artwork documentation (51 images), exhibition view (45 images) and unidentified (43 images). Across six controlled re-runs, the progressive introduction of archivist-defined terms and refined category definitions increased agreement between model-generated classifications and the archivist's validated decisions from 51 of 179 images (28%) to 105 of 179 images (59%), more than doubling it (Table 2). Because the same sample was used to formulate these changes,

the re-runs measure their effect on familiar material rather than generalisation; transfer to unseen units is assessed separately in Section 5.3.

Table 2
Images on which the model's content type agreed with the archivist, development sample. Run 1 was reviewed by the archivist and is inflated by anchoring; later runs were scored without being seen

Archivist's term (images)	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
Experimental photography (62)	0	14	48	13	20	21
Architectural photography (37)	0	27	29	27	27	27
Artwork documentation (24)	19	18	17	21	21	20
Portrait (15)	7	0	0	12	9	8
Landscape / cityscape (15)	12	9	15	14	14	14
All images (179)	51	81	121	98	105	105

The controlled re-runs show that a refinement works only when a category can be recognised in the image. Architectural photography, identifiable from visible features, was classified correctly in 27 of 37 images once introduced. Experimental photography, defined by how an image was made, was not: Madra's light drawings were frequently read as artwork documentation rather than as works in their own right. A definition separating a photograph of an artwork from a photograph that is the artwork (Run 5) added only seven correct classifications; the distinction depends on knowledge of artistic production that the image does not show.

5.2 The controlled vocabulary as part of the model configuration

Re-reading the first review round (100 images from the full archive) with two revised versions of the model's vocabulary lowered agreement with the archivist in both cases, from 66 to 46 and to 53 of 100 (Fig. 3).

The control repeated every content type and every comparable description (95 of 95) word for word, so each difference comes from the vocabulary. The rewording fixed 3 of the 34 errors and broke 23 correct answers, and the three targeted terms did not improve.

5.3 Reference examples and process-defined categories within units

Certain descriptive categories depend on the processes through which images were produced rather than on their observable visual characteristics. Experimental photography provides a particularly revealing example: a light drawing constituting an artwork may appear visually similar to a photograph documenting an existing work, making the distinction difficult to establish through textual definitions alone. However, reference examples validated by the archivist proved substantially more effective. Within

the development sample, similarity-weighted voting among the five most visually similar reviewed images correctly classified all 62 experimental photographs, compared with 20 correctly classified by the vision-language model. Combining example-based proposals with model-generated classifications—using the weighted vote when both the agreement and similarity thresholds were satisfied, and retaining the model's classification otherwise—increased overall agreement from 59% to 85% under leave-one-out evaluation. Agreement reached 82% when the target image's entire visual group was excluded from the reference set. Among proposals meeting the established thresholds, agreement with the archivist's decisions reached 94% under leave-one-out evaluation and 90% with the visual group excluded.

These improvements, however, did not generalise reliably to previously unseen archival units. Across five units excluded from development (253 images), the reference examples exhibited substantially lower visual similarity to the new material. The median similarity between an unseen image and its nearest reviewed counterpart was 0.74, compared with 0.89 within the development sample. Consequently, only 19 proposals satisfied the minimum similarity threshold, of which nine agreed with the archivist's classifications (Table 3). Incorporating validated examples from other unseen units did not consistently improve performance and, in some cases, introduced misleading proposals. Visually similar materials, such as photographs of theatrical sets and costume parties, could belong to distinct archival contexts and therefore require different classifications. These findings demonstrate that the effectiveness of example-based classification depends not only on visual resemblance but also on the contextual relationships within which reference examples are situated.

Table 3
Review-loop test on the five unseen units (253 images). The model alone agreed on 67%

Examples available	Nearest similarity (median)	Vote agreed	Clear proposals: share of images, agreed	Combined with the model
Development sample only	0.74	12%	8%, 47%	67%
+ the other unseen units	0.77	28%	11%, 31%	64%
+ first 5 images of this unit	0.78	49%	14%, 74%	68%
+ the rest of this unit	0.86	78%	53%, 95%	79%

Decisions from within the same unit, by contrast, helped strongly. With about ten images or a fifth of a unit reviewed, clear proposals agreed with the archivist 92–96% of the time; gains levelled off after 20–30%, and reviewing half the unit did not reduce the remaining work further. In this archive, the relevant context is found within the unit.

5.4 The full archive

Following the development phase, the model was applied to the full archival collection. Its classification performance was evaluated through two balanced review rounds comprising 100 images each. Based on these assessments, the estimated agreement between model-generated content-type classifications and the archivist's validated decisions was 69% across the collection (95% confidence interval: 61–77%). Of the 200 images reviewed, 143 classifications (72%) agreed with the archivist's decisions. The collection-level estimate is slightly lower than this unweighted sample agreement because the balanced rounds over-represent rare content types, while some frequent types, above all exhibition view, showed low agreement and therefore weigh more heavily in the estimate (Fig. 4).

Exhibition view is the clear failure: the model applies it far more broadly than the archivist, and its 12 errors scatter over seven other types, from artwork documentation and portraits to landscapes and casual snapshots, so that no single refinement of the definition could correct it.

5.5 Orientation learnt from the archivist's decisions

A third of the scans in the first review round (33 of 100) needed turning, and the model's own flag was no guide: scans it called rotated right were turned left as often as right. The orientation classifier was first trained on only 444 images whose orientation the archivist had decided, under 2% of the 26,157 images in the collection; Table 4 gives its results at a probability threshold of 0.6, together with a second pass trained on 6,329 decisions, and Fig. 5 shows two examples.

Table 4
Tests of the orientation classifier

Test	Turn proposals correct	Turned scans found	Upright scans wrongly turned
Review round 1, kept out of learning (84 images)	16 of 16	16 of 25	0 of 59
Whole units left out in turn (360 images)	64 of 67 (96%)	64 of 105	3 of 255
Second pass, learnt from 6,329 decisions: round 1	21 of 21	21 of 25	0 of 59

Across the full collection of 26,157 images, the orientation classifier proposed 7,079 rotations in two successive passes (6,122 and 957), representing approximately 27% of the collection. This proportion is broadly consistent with the initial review round, in which approximately one-third of the images required reorientation. During unit-by-unit review, the archivist accepted 6,772 of the 6,773 evaluated proposals. At an average review time of 1.7 seconds per image, the first pass required approximately three hours, substantially reducing the manual effort that would otherwise have been necessary to inspect and rotate each image individually.

The exceptionally high acceptance rate, however, raises concerns about potential automation bias, particularly because proposed rotations were presented as selected by default. To assess their reliability

beyond the archivist's immediate acceptance, two independent validation procedures were conducted. First, a visual examination of 38 proposed rotations, including 24 drawn from the lowest-confidence range, confirmed that all visually interpretable cases were correctly oriented; one dark, abstract image could not be assessed conclusively. Second, the subsequent review round included 17 images previously rotated according to the classifier's proposals, all of which were independently confirmed to have the correct orientation. The five additional images requiring manual rotation in that round belonged exclusively to cases for which the classifier had not generated sufficiently confident proposals. Together, these findings provide supporting evidence for the reliability of the orientation procedure while highlighting the importance of evaluating machine-generated proposals independently of their acceptance during routine archival review.

5.6 Unit proposals

Unit proposals put the finding of Section 5.3 into practice: once the archivist has classified about a fifth of a unit, the remaining images receive proposals voted only by reviewed images from the same unit (Section 4.5, Fig. 2). To test them, the four fully reviewed units were replayed as if new; Table 5 shows the results, averaged over 20 evenly spaced selections of the images treated as reviewed.

Table 5
Unit proposals replayed on four fully reviewed units, means over 20 selections

Unit	Images	Treated as reviewed	Clear proposals for the rest	Agreed
NEGA00356	156	32	39.0 of 124	38.6
NEGA00226	80	16	18.0 of 64	17.2
NEGA00119	35	10	12.5 of 25	10.4
NEGA00051	25	10	0 of 15	0

The unit-based proposal mechanism achieved 95% agreement with the archivist's validated classifications among proposals meeting the established thresholds, with a coverage of approximately 30% of the remaining images (an average of 69.5 out of 228). Coverage refers to the proportion of items for which the system generates proposals meeting its predefined confidence criteria (Rolan et al. 2019, p 188). The number of qualifying proposals varied between 60 and 88 across individual selections, indicating that the choice of initial reference images substantially influences the effectiveness of subsequent classification. Extrapolating these findings to a unit of 100 images, the archivist would initially review 20 images, accept approximately 23 proposals, correct one and individually classify the remaining 56. This would avoid approximately 23 independent classification decisions while requiring the review of 24 machine-generated proposals; the actual time saved would depend on the effort required to assess and correct these proposals.

5.7 Face grouping and the identification of individuals

In a pilot on 128 units (4,140 images), 2,623 faces were found and grouped. The archivist judged 13 groups in about eight minutes, naming 207 faces and marking 129 as someone else. Excluding one catch-all group of small, blurred faces, 206 of 214 faces in named groups were the person the archivist named (96%).

Name proposals were first tested by leaving out one named face at a time: 202 of 202 matched the archivist's names. The archivist's decisions on real proposals told a different story: 41 of 75 accepted (55%). The test had been flattered by near-duplicate frames of the same person within a group. Below a similarity of 0.50 only 7 of 30 proposals were accepted; at 0.50 and above, 34 of 45 (76%). Family resemblance and matches across units were the hard cases.

On the full archive, 12,643 faces were found in 6,220 images in 458 units. Of these, 8,225 were clear enough to group, into 2,564 groups. The archivist then made 2,173 face decisions in one morning: 1,831 faces named and 342 marked as someone else. In total 1,883 faces of 19 people are named, in 1,562 pictures across 227 units. They include the artist (218 faces) and members of his family, and two private persons with 142 and 101 faces, whose names are withheld in this article. Two groups still mix family members, and face fingerprints separate infants poorly.

5.8 From visual recognition to archival knowledge: a gradient of tasks

Taken together, the results form a gradient (Table 6). How much can be proposed depends on how far the knowledge lies from what the picture shows, and what the proposal must start from moves steadily from the image to the archivist.

Table 6
What the method proposes, from what, and how well

What is proposed	Proposed from	Agreement	Who decides
Orientation	a small classifier on image fingerprints	96% of turn proposals correct on units left out; 6,772 of 6,773 accepted	the archivist confirms, unit by unit
Content type	the model with the archivist's vocabulary	69% of the collection (95% confidence interval 61–77%)	the archivist, image by image
Content type, rest of a unit	the archivist's decisions in that unit	95% of clear proposals, for about 30% of the rest	the archivist accepts or corrects
Identity	faces grouped by likeness; the archivist's names	76% of name proposals	only the archivist names
Series	the archivist's decisions on similar images, as proposals only	44 of 61 series proposals agreed (72%); 44 of 46 series images found (development sample)	only the archivist assigns
Occasion	not proposed	—	the archivist

6 Discussion

6.1 Original order returns

Content-based indexing has been proposed as a way of extending access beyond the structures established by provenance and original order, complementing rather than replacing archival arrangement (Colavizza et al. 2022, Organising and accessing archives). Description presents a different challenge, and here the findings give empirical support to the same authors' suggestion that provenance may acquire renewed significance in data-driven archival environments (Colavizza et al. 2022, Discussion). Classifications derived from the archivist's validated decisions transferred to other images within the same unit but did not generalise across units. A unit frequently corresponds to an occasion, a period or a working session, and its items share contextual relationships that visually similar material elsewhere need not share. Even where original order was informal or only partly preserved, these inherited groupings provided useful boundaries for machine proposals. The evaluation does not, however, separate the contribution of contextual relationships from the visual similarity of items within a unit.

The identification system introduced during digitisation preserves the same groupings while linking each digital file to its physical source, so original order serves both the machine's interpretation and the archivist's return to the negative, print or tape. Rolan et al. (2019, p 186) caution that a model developed for one context may fail in another, "even one that is seemingly close"; in this archive, the unit marks where validated decisions remain useful and where renewed human interpretation becomes necessary.

Original order thus emerges not only as a principle to be preserved during AI-assisted processing but as an active source of context that makes machine proposals more reliable. It also follows that review cannot be a single calibration stage before automated classification: the archivist establishes a reference set for each new unit before proposals are extended to the rest, the "iterative human evaluation" Cordell (2020, p 22) calls for, with human judgement an ongoing part of description rather than a final check.

6.2 Vocabulary versioning as a form of archival control

The improvements demonstrated in this study resulted from refinements to the archival workflow surrounding the AI models rather than from retraining or modifying the models themselves. This distinction is particularly relevant to smaller archival institutions with limited technical and computational resources. The principal challenge of AI-assisted archival description lies not solely in computational capacity but also in the design of procedures through which machine-generated outputs are structured, evaluated and integrated into archival practice. Such procedures, including controlled vocabularies, systematic review mechanisms and the incorporation of archivist-validated decisions, can potentially be adapted to other collections and implemented with different AI models.

The observed sensitivity of model outputs to changes in the controlled vocabulary (Section 5.2) transforms an established archival concern into a technical requirement. The vocabulary supplied to an AI model is not merely a reference list but an integral component of the classification process, capable of influencing descriptive outcomes even when the underlying model remains unchanged. Consequently, revisions to model vocabularies require systematic documentation, version control and evaluation, comparable to the governance practices applied to archival descriptive standards. Associating each machine-generated classification with the specific vocabulary version used to produce it extends the principle of provenance beyond the archival object to the process of its description. This approach enables subsequent researchers and archivists to reconstruct the conditions under which descriptive metadata was generated, assess the effects of vocabulary revisions and distinguish changes in archival interpretation from changes in computational procedures.

6.3 Experimenting with the archive

Beyond archival description and classification, AI played a less immediately quantifiable role in the development of the archival methodology itself. Confronted with tens of thousands of initially unclassified items, the archivist faced fundamental uncertainties: where to begin, which descriptive categories to establish, how much material could feasibly be reviewed and in what sequence. AI-assisted experimentation allowed alternatives to be tested before committing to a workflow: a representative sample could be processed within an hour, and a proposed vocabulary evaluated against the archivist's validated decisions, or a review round generated and assessed, within a single afternoon. The methodology presented here therefore emerged through experimentation rather than being fully established in advance; the separation of archivist and model vocabularies, structured review rounds,

unit-based proposals and the logging of modifications were each introduced in response to limitations identified in successive tests.

This finding resonates with Rolan et al. (2019, p 196), who suggest that archival processes may require substantial adaptation before AI can contribute effectively to description, and identify the insufficient documentation of AI-assisted decisions as a persistent concern. In the present study, the methodology and its computational components evolved together, while the audit log and version-controlled vocabulary documented the conditions and decisions underlying each description, answering Padilla's (2019, p 10) proposal for a record of every change to description, whether human or algorithmic. For large-scale, unstructured collections, the value of AI may therefore extend beyond processing efficiency: by reducing the effort required to test alternative procedures, it enables an experimental archival practice in which systematic methods are developed, evaluated and revised in response to the characteristics of a collection.

6.4 Affirmation and automation bias

The archivist's role varies with the increasing demands for contextual knowledge illustrated in Table 6. For image orientation, human involvement consists mainly of confirming machine-generated proposals, at about 1.7 seconds per image; content-type classification requires more sustained judgement, often of each image individually; and for identifying individuals, assigning artistic series and establishing the occasions documented, the archivist becomes the primary source of contextual and historical knowledge rather than merely a validator of computational outputs. Throughout, the authority to determine what enters the catalogue remains with the archivist, irrespective of the model's reported confidence. The relationship is nevertheless not unidirectional: computational image analysis occasionally revealed details that had escaped the archivist's attention, prompting revisions to earlier assessments. Human authority should not be equated with human infallibility; it ensures that every observation, whether it originates from the archivist or the system, is evaluated before it enters the record.

The increasing efficiency of human confirmation also introduces epistemic risks. Plausible but inaccurate classifications may be difficult to detect during rapid review, and automation bias increases the likelihood that they are accepted without sufficient scrutiny (Parasuraman and Manzey 2010, p 391). Two methodological implications follow. First, the acceptance of proposals should be reported as agreement with the archivist's decisions rather than as an independent measure of accuracy, and supplemented where possible by separate validation, as in the evaluation of image orientation. Second, retrospective assessments on previously reviewed materials may overestimate reliability: identity proposals appeared entirely correct against faces already identified by the archivist, yet only 55% were accepted for previously unnamed faces. Systems should therefore be tested on genuinely unreviewed archival objects.

The design of the review interface also influences how responsibility is exercised. Presenting proposals as selected by default (Fig. 2) reduced review time for the experienced archivist conducting this study,

but the same design may encourage uncritical acceptance in collaborative settings. There, presenting proposals without preselection or introducing previously validated cases into review rounds may help sustain critical attention and individual accountability, since responsibility can become diffused when decisions are shared between human operators, automated systems or multiple collaborators (Parasuraman and Manzey 2010, p 393). Effective oversight therefore depends not only on formal decision-making authority but also on review procedures that actively support independent judgement.

6.5 Privacy and naming

Facial recognition introduces ethical and legal responsibilities beyond those of other archival classification tasks. Facial embeddings generated to identify individuals constitute biometric data, a special category of personal data under Turkey's Personal Data Protection Law (KVKK, Law No. 6698, Article 6) and subject to specific protections under the European Union's General Data Protection Regulation (GDPR, Article 9). In a privately held archive, those depicted may never have consented to biometric processing or to inclusion in a searchable catalogue. The Madra archive illustrates the range of circumstances: public figures attending exhibitions and performances appear alongside private individuals, including family members photographed in domestic settings over several decades, some during childhood. The distinction between public and private contexts therefore shapes how personal information is processed, documented and made accessible.

The workflow therefore incorporates several safeguards. All biometric processing is performed locally, without transmitting facial images or embeddings to external services. The system does not assign identities autonomously; it generates similarity-based proposals that require explicit validation by the archivist. Each confirmed identification is recorded in the audit log with the person who made the decision and its time. Validated names become descriptive metadata subject to the archive's access policies and to data protection requirements; public disclosure of the identity of a private individual requires an appropriate legal basis, such as valid consent, and does not follow automatically from inclusion in the catalogue.

As access to the archive expands and more contributors become involved, identification may become a collaborative process drawing on the knowledge of many participants. This can improve the completeness and accuracy of the record, but it requires clearly defined responsibilities and differentiated permissions: the authority to contribute an identification should be distinct from permission to access or disclose personal information or biometric data, so that collaborative identification remains subject to human oversight and appropriate privacy safeguards.

7 Conclusion and future work

AI can take an archive out of the shoebox by giving the archivist a toolbox, but not on its own. In the Teoman Madra archive, locally deployed models, combined with an archive-specific controlled vocabulary and structured review, reached approximately 70% agreement with the archivist's content-type classifications, and facial similarity-based grouping enabled the archivist to identify nearly 2,000

faces within a single morning. AI-assisted experimentation also contributed to the development of the archival methodology itself. Yet increased efficiency did not by itself produce greater descriptive reliability, and every improvement depended on archival affirmation, the archivist's accountable and revisable judgement: the terms and definitions, the reference examples validated within units and the historical and biographical knowledge that could not be inferred from the images alone.

Three methodological principles emerge with potential relevance beyond the collection examined. First, machine-generated descriptions, archivist-validated modifications and contextual information derived from archival sources should remain distinguishable, each with its provenance documented. Second, the controlled vocabulary supplied to an AI model should be treated as a versioned component of the descriptive process, subject to evaluation and governance. Third, original order should set the boundaries within which machine-generated proposals are considered reliable, since contextual relationships within archival units may support classification better than visual similarity across otherwise unrelated materials. Together, these principles position AI-assisted description as a collaborative and accountable process in which computational capabilities extend, rather than replace, archival expertise.

The study is subject to several limitations. Its findings derive from the collection of a single artist, and agreement was assessed against the decisions of one archivist, who also conducted the research; the study cannot establish whether independent archivists would reach comparable decisions. Some findings rest on relatively small samples, with results pooled across review rounds. Several classifications were validated while the model's suggestions remained visible, introducing the possibility of anchoring and automation bias; the controlled re-runs and unchanged control configuration (Section 5.2) reduce but cannot eliminate these effects. Finally, audiovisual recordings, although processed by the pipeline, receive limited attention here, a significant gap given that the videotapes also preserve Madra's computer-generated works from the late 1980s onward.

Future research should address three directions. First, unit-based proposals should be evaluated prospectively on previously unreviewed units rather than through retrospective simulations, to measure the time saved and the human intervention required under actual working conditions. Second, independent review of the same samples by an additional archivist would enable the measurement of inter-annotator agreement, particularly for ambiguous or context-dependent categories, and provide a more appropriate basis for interpreting agreement between human and machine classifications. Third, the open-weight models permit local adaptation: low-rank adaptation (LoRA) could be trained on archivist-validated classifications from balanced review rounds and evaluated on an unseen, randomly sampled test set, distinguishing the benefits of adaptation from those of iterative human–AI interaction. Adaptation may prove most effective for visually grounded ambiguities, such as exhibition views and artwork documentation, and least for classifications requiring historical or contextual knowledge unavailable in the images themselves.

Ultimately, the findings suggest that the sustainable integration of AI into archival practice depends less on the pursuit of autonomous classification than on the development of transparent, adaptable and critically evaluated workflows. The significance of AI lies not only in accelerating archival processing but also in enabling archivists to experiment with, refine and systematically document their descriptive methods. Within this framework, human judgement remains essential not because it is infallible, but because archival meaning depends on contextual interpretation, accountable decisions and the preservation of relationships that computational recognition alone cannot establish.

Declarations

Author Contribution

Use of generative AI. The author developed the pipeline and the search tool in an iterative dialogue with an AI assistant (Claude, Anthropic): the author set out the intended system, tested each version on the archive and directed its revision, while the assistant wrote and debugged the code. The assistant was also used to run the analyses and to draft and edit the text. All review decisions, vocabulary terms, definitions and interpretations are the author's own, and the author takes responsibility for the article.

References

1. Arnold T, Tilton L (2023) Distant viewing: computational exploration of digital images. MIT Press, Cambridge, MA
2. Bai S, Chen K, Liu X et al (2025) Qwen2.5-VL technical report. arXiv:2502.13923
3. Bowker GC, Star SL (1999) Sorting things out: classification and its consequences. MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/6352.001.0001
4. Colavizza G, Blanke T, Jeurgens C, Noordegraaf J (2022) Archives and AI: an overview of current debates and future perspectives. ACM J Comput Cult Herit 15(1):4. https://doi.org/10.1145/3479010
5. Cordell R (2020) Machine learning + libraries: a report on the state of the field. Library of Congress, LC Labs, Washington, DC
6. Crawford K (2021) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven
7. Crawford K, Paglen T (2019) Excavating AI: the politics of images in machine learning training sets. AI Now Institute, New York University
8. EGAD, Expert Group on Archival Description (2023) Records in Contexts – conceptual model (RiC-CM), version 1.0. International Council on Archives, Paris. https://www.ica.org/app/uploads/2024/01/ric-cm-1.0_0.pdf
9. Gong Y, Chung Y-A, Glass J (2021) AST: Audio Spectrogram Transformer. In: Proceedings of Interspeech 2021, pp 571–575. https://doi.org/10.21437/Interspeech.2021-698

10. Jaillant L, Caputo A (2022) Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data. AI Soc 37(3):823–835. https://doi.org/10.1007/s00146-021-01367-x
11. Jaillant L, Kidd M, Zhao L (2026) Should sensitive archival images be shown? The case for applying AI to colonial archives. Arch Sci 26:36. https://doi.org/10.1007/s10502-026-09552-x
12. Padilla T (2019) Responsible operations: data science, machine learning, and AI in libraries. OCLC Research, Dublin, OH. https://doi.org/10.25333/xk7z-9g97
13. Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52(3):381–410
14. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I (2023) Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning, PMLR 202, pp 28492–28518
15. Rolan G, Humphries G, Jeffrey L, Samaras E, Antsoupova T, Stuart K (2019) More human than human? Artificial intelligence in the archive. Arch Manuscr 47(2):179–203. https://doi.org/10.1080/01576895.2018.1502088
16. Shinde G, Kirstein T, Ghosh S, Franks PC (2024) AI in archival science: a systematic review. arXiv:2410.09086
17. Wu W, Peng H, Yu S (2023) YuNet: a tiny millisecond-level face detector. Mach Intell Res 20(5):656–665. https://doi.org/10.1007/s11633-023-1423-y
18. Zhai X, Mustafa B, Kolesnikov A, Beyer L (2023) Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 11975–11986
19. Zhong Y, Deng W, Hu J, Zhao D, Li X, Wen D (2021) SFace: sigmoid-constrained hypersphere loss for robust face recognition. IEEE Trans Image Process 30:2587–2598. https://doi.org/10.1109/TIP.2020.3048632

Figures

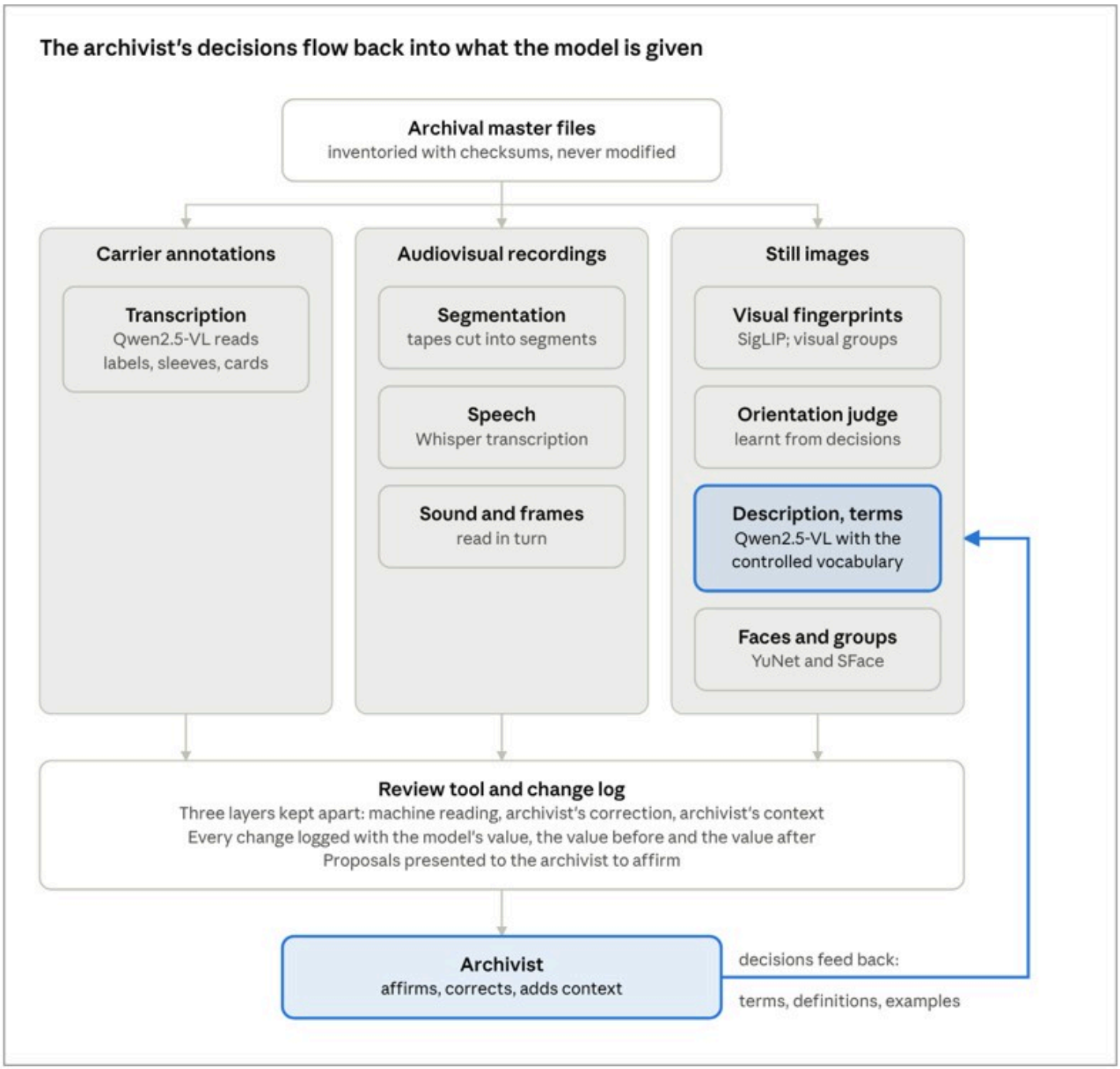


Figure 1

The processing pipeline and the archivist's review loop

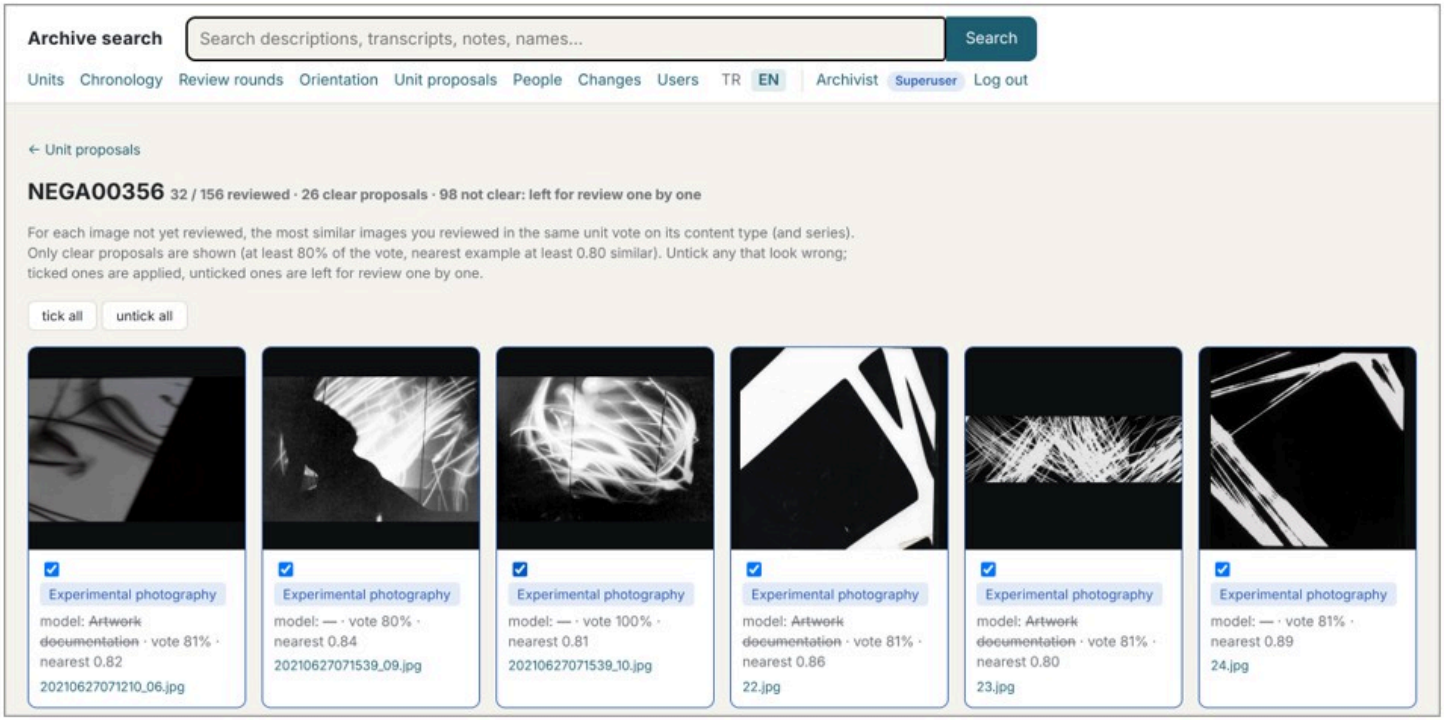


Figure 2

The unit-proposal page of the review tool, shown for unit NEGA00356 with 32 of its 156 images treated as reviewed (a replay): proposals arrive ticked, and the archivist unticks any that look wrong

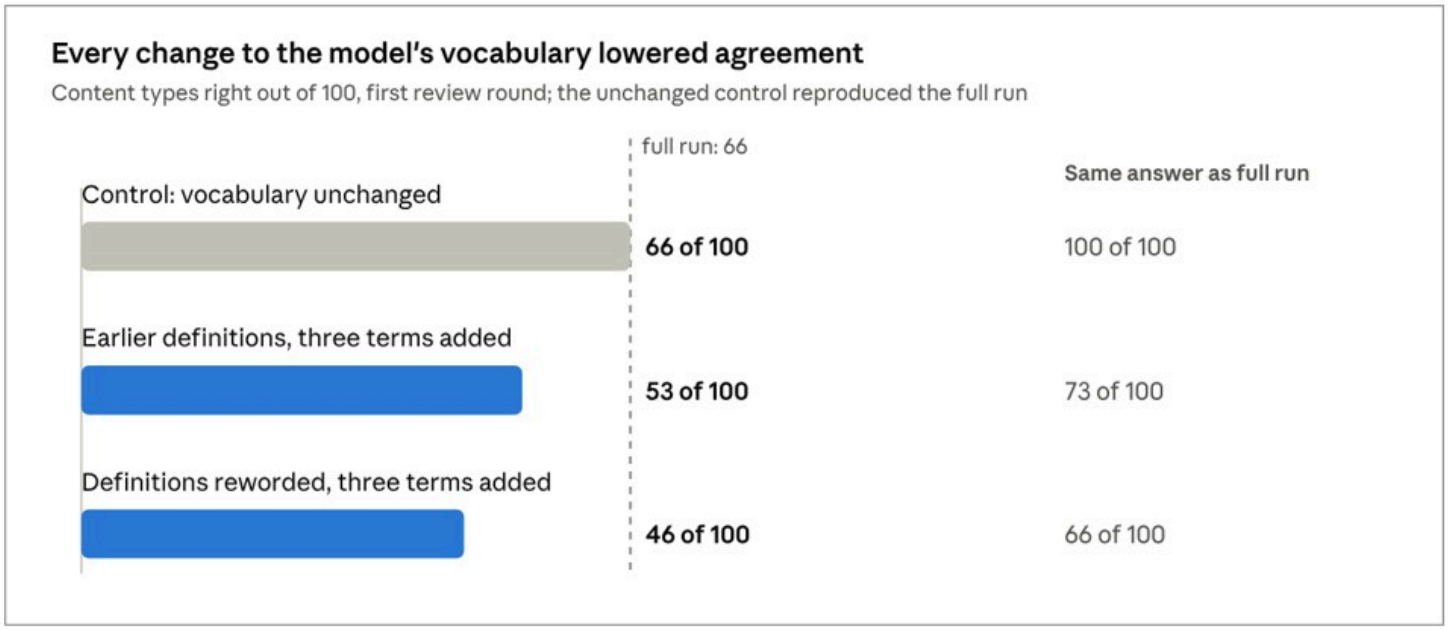


Figure 3

Re-reading review round 1 (100 images) with different versions of the model's vocabulary

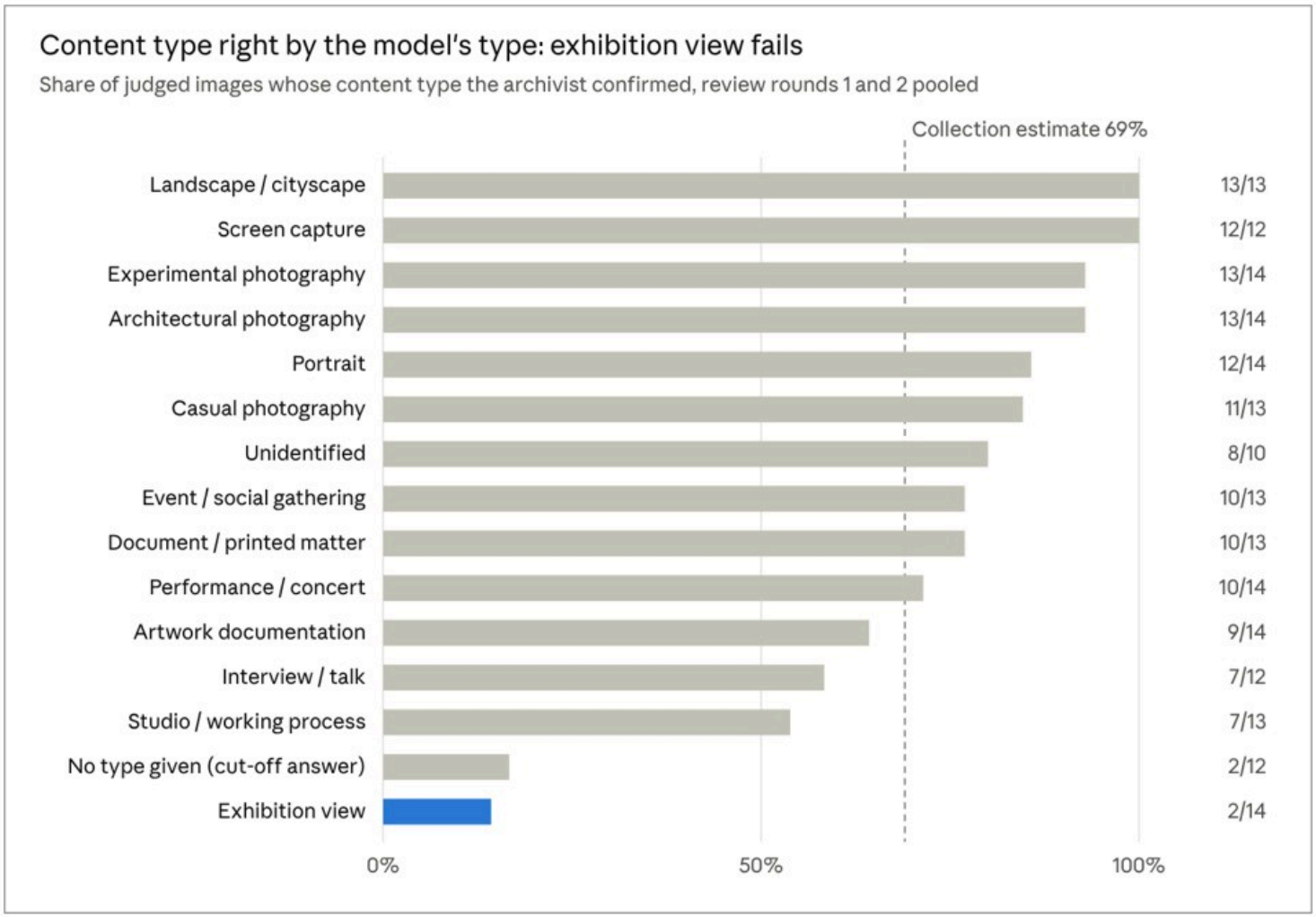


Figure 4

Review rounds 1 and 2 (24 October 2026), the 15 content types with at least ten judged images (195 of 200)

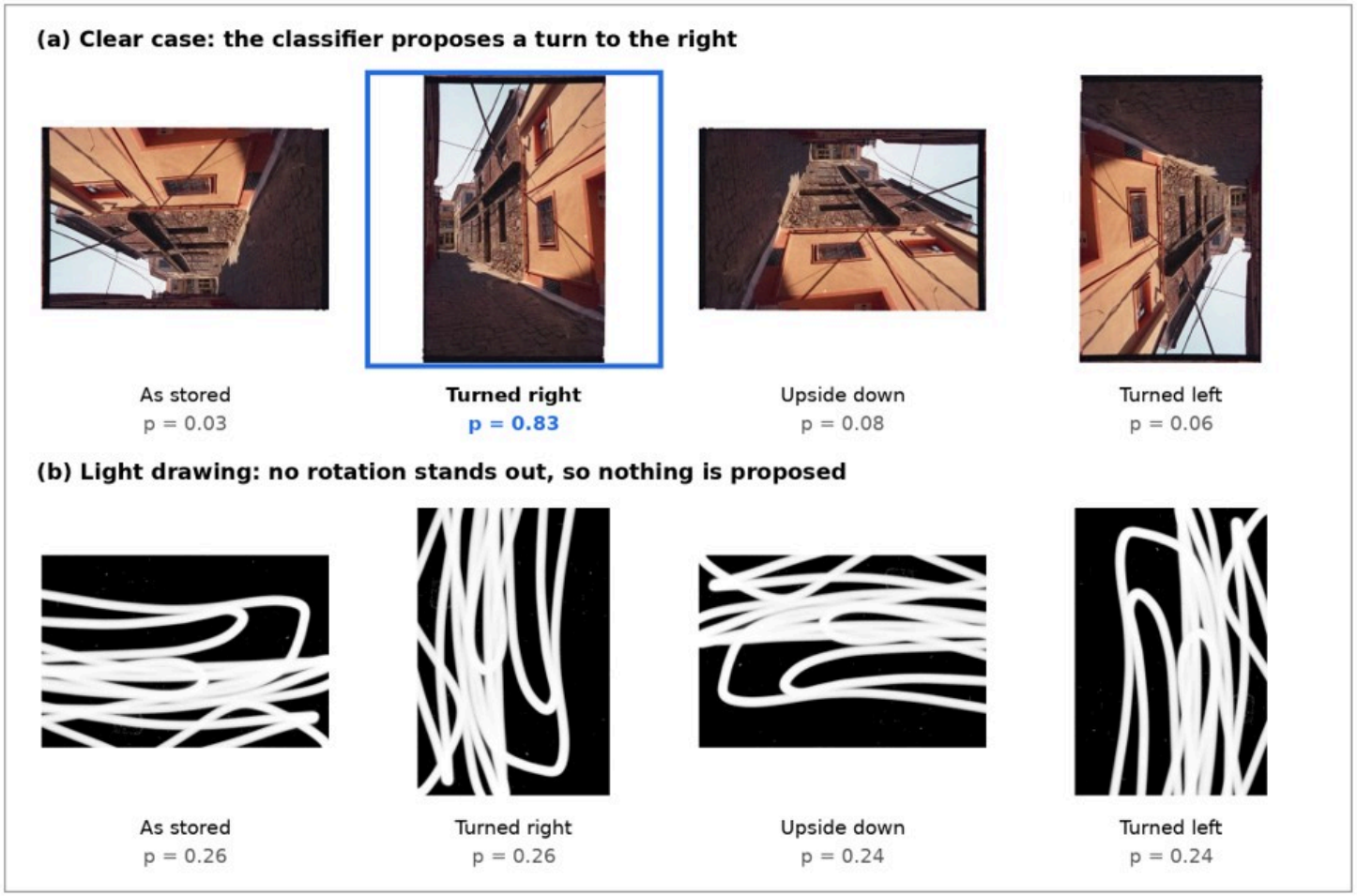


Figure 5

The orientation classifier on two scans from the archive: (a) NEGA00642-0026, proposed for turning right (p = 0.83); (b) NEGA00093-0001, a light drawing with no natural up, for which the four rotations score almost equally (0.24–0.26), so none reaches the 0.6 threshold and nothing is proposed