Cognitive Nutrition for Foundation Models:

Empirical Validation of the PEST Framework for Provenance-Verified, Semantically Dense Training Data

Tad MacPherson[1]

[1] Metavolve Labs, Inc., Research Division
nexus@metavolvelabsinc.com

Abstract

In *The Entropy of Recursion* [1], we introduced the Provenance-Embedded Semantic Transport (PEST) framework as a theoretical response to model collapse and context blindness—two converging failure modes threatening the integrity of foundation model training. That paper argued that training data requires not merely volume but *Cognitive Nutrition*: verifiable provenance, dense semantic scaffolding, and preservation of human-grounded variance.

This paper presents the empirical realization of that vision. We report on the design, implementation, and deployment of a production metadata factory that has processed 10,097 cultural heritage artifacts into provenance-verified, semantically enriched "Aeternum Assets," each carrying 2,000–6,000 tokens of structured context across 111+ fields organized into four ontology layers. Central to our approach is the *Translation Paradigm*: enrichment is not AI synthesis from nothing but machine-assisted articulation of existing human knowledge—museum curatorial records, domain expertise encoded in model weights, and visual analysis that renders explicit what trained specialists recognize implicitly. The enrichment pipeline employs a *multi-model Oracle Synthesis*, orchestrating complementary frontier models to achieve depth, nuance, and structural precision that no single model produces alone.

We present system architecture, operational metrics, and dataset-level statistics from the *Alexandria Aeternum* collection. We propose a mathematical framework for evaluating Cognitive Nutrition using topological data analysis (persistent homology for distributional drift detection) and causal disentanglement (back-door adjustment for separating causal from spurious features). We introduce the *Amendability Framework*—a living curation model where originating institutions retain governance over their enriched collections, enabling versioned corrections, human verification tagging, and trust-level filtering that allows downstream consumers to select purely human-verified data subsets. Finally, we propose a rigorous evaluation methodology—controlled ablations, recursive contamination stress tests, and enrichment scaling studies—designed to falsify or validate the hypothesis that semantically dense, provenance-verified data measurably improves model robustness.

Keywords: Model collapse, training data integrity, provenance, semantic topology, perceptual hashing, data stewardship, GLAM, topological data analysis, causal inference, multi-model synthesis

---

1. Introduction

1.1 From Theory to Factory

In [1], we established that the era of indiscriminate Data Accumulation must yield to Data Stewardship. We identified two converging crises:

1. **Model Collapse** (the Xerox Effect): recursive training on synthetic output causes distributional tails to vanish, compounding variance loss across generations until models converge to low-entropy point estimates [2, 3].
2. **Context Blindness**: the systematic stripping of metadata during web redistribution produces "lobotomized" training artifacts—pixels without provenance, intent, or semantic grounding [1, 4].

We proposed the PEST framework, anchored by a Persistent Provenance Pointer (PPP) embedded via steganography and linked to a four-layer metadata ontology. We reported preliminary validation with a 98.7% pointer survival rate across simulated social media transformations [1]. In this paper, we replicate and extend that result with a systematic batch test (Section 5.5), achieving 100% recovery (80/80) under four social media transformation classes with mean similarity of 99.5%.

What we did not present was *the system itself*. This paper fills that gap. We describe a production metadata factory that has been operational since late 2025, processing museum-grade cultural heritage artifacts through a multi-agent pipeline of AI enrichment, metadata infusion, cryptographic hashing, and immutable registration. We report quantitative results from the *Alexandria Aeternum* 10K collection—the first large-scale corpus of Cognitive Nutrition assets—and formalize the mathematical underpinnings that were only sketched in [1].

1.2 The Urgency Has Intensified

Since the publication of [1], the evidence base for both model collapse and data scarcity has strengthened materially:

- Epoch AI's updated analysis estimates that high-quality, non-duplicated public text will be effectively exhausted under plausible scaling policies by 2026–2028 [5]. We are now within that window.
- Shumailov et al.'s model collapse findings [2] have been replicated across modalities, and Dohmatob et al. [3] have identified a "generalization-to-memorization" transition in late-stage autophagy where models cease learning principles and begin memorizing artifacts of previous generations.
- The data-centric AI community has demonstrated that dataset design—not just scale—drives model performance. DataComp [6] showed that data curation alone, at fixed compute, yields material downstream improvements. RefinedWeb [7] demonstrated that filtered web data can outperform curated mixtures. DoReMi [8] showed that optimizing data mixture weights improves training efficiency.

These developments validate the central premise of [1]: the bottleneck is no longer compute or architecture but *the quality, provenance, and semantic density of the training signal itself.*

1.3 Contributions of This Paper

1. **The Translation Paradigm**: A formal argument that Neural Extraction of Semantic Topology (NEST) enrichment constitutes machine-assisted articulation of human knowledge,

not synthetic generation—a distinction with fundamental implications for how enriched datasets should be classified and trusted (Section 3.4).

2. **System description**: A production metadata factory architecture employing multi-model Oracle Synthesis that operationalizes the PEST framework at scale (Section 4).
3. **Empirical results**: Quantitative statistics from the Alexandria Aeternum 10K corpus (Section 5), including metadata density, processing economics, and provenance integrity metrics.
4. **Mathematical formalization**: Topological data analysis (persistent homology) and causal disentanglement as the theoretical foundations for Cognitive Nutrition (Section 6).
5. **The Amendability Framework**: A living curation model enabling institutional governance, versioned corrections, and trust-level field tagging (Section 10).
6. **Evaluation methodology**: A concrete experimental program to validate or falsify the Cognitive Nutrition hypothesis (Section 7).

2. Background and Related Work

2.1 Model Collapse: Updated Evidence

The model collapse phenomenon, formalized by Shumailov et al. [2], establishes that training on recursively generated data causes irreversible distributional narrowing. Let M approximate a true distribution D_{real}. Sampling with temperature truncation produces D_{syn} where:

$$\text{Var}(D_{\text{syn}}) < \text{Var}(D_{\text{real}}) \tag{1}$$

Under recursive iteration:

$$\lim_{n \to \infty} \text{Var}(M_n) \to 0 \tag{2}$$

Dohmatob et al. [3] expanded this by identifying distinct phases: early-stage "weak" collapse (gradual tail erosion) followed by a phase transition into "strong" collapse (generalization-to-memorization), where models internalize the statistical artifacts of previous generations as ground truth. This is not a theoretical curiosity—it is the *expected equilibrium* of any self-consuming generative loop without external high-entropy injection.

The symptoms are observable in deployed systems: mode collapse in image generation (identical outputs for varied prompts), artifact amplification (six-fingered hands codified as anatomical norms), and reality drift (physics governed by statistical probability rather than causality) [1, 9].

2.2 Context Blindness and the Provenance Gauntlet

Current provenance standards (C2PA [10], IPTC/XMP [11]) attach metadata as file headers. Research confirms that the vast majority of social media platforms strip these headers during upload, re-encoding, or screenshot capture [4, 12]. A C2PA signature survives only within compliant ecosystems; the moment an image enters the "Social Media Gauntlet" (compression, cropping, re-upload), its provenance is destroyed.

This creates a fundamental asymmetry: models are trained on billions of images scraped from the web, but the *contextual signal* (provenance, intent, emotional framing, rights) has been systematically removed from the vast majority of these images before ingestion. The model

receives pixels without meaning—"stripped artifacts" that force statistical pattern-matching where semantic understanding is required [1].

Critically, C2PA verifies *who signed* an asset but not *what the asset represents* semantically [13]. It provides a chain of custody without a chain of meaning.

2.3 The Output-Side / Input-Side Distinction

A key distinction introduced in [1] and underappreciated in the broader literature is between *output-side* and *input-side* interventions:

- **Output-side** systems (e.g., Google DeepMind's SynthID [14]) watermark AI-generated content to identify synthetic outputs. They label what a model *produces*.
- **Input-side** systems (PEST) enrich and verify what a model *consumes*. They inject the missing semantic context, provenance, and variance that training data requires.

These are complementary, not competing. SynthID helps detect synthetic contamination; PEST helps ensure that non-synthetic data carries sufficient information density to stabilize models against collapse. The present work focuses exclusively on the input side.

2.4 Data-Centric AI: Quality as the Variable

The data-centric AI movement has produced compelling evidence that dataset engineering—holding model architecture and compute fixed—materially affects downstream performance:

- **DataComp** [6]: A benchmark where participants improve dataset filtering/sourcing with fixed training code. Demonstrated that data curation alone yields substantial gains across dozens of downstream tasks.
- **RefinedWeb** [7]: Showed that properly filtered and deduplicated web data outperforms curated mixtures previously considered necessary for strong performance.
- **FineWeb** [15]: An open, large-scale web dataset with explicit quality filters and educational variants, reinforcing the value of systematic curation pipelines.
- **DoReMi** [8]: Demonstrated that optimizing data mixture weights improves downstream accuracy and training efficiency, suggesting that "nutrition" can be operationalized as mixture allocation and tail coverage.

These results establish the baseline that Cognitive Nutrition must exceed: it is not enough to simply enrich data; the enrichment must demonstrably outperform best-practice filtering and curation of raw web data.

2.5 Dataset Documentation and the "Nutrition" Metaphor

The Dataset Nutrition Label [16] proposes modular, standardized diagnostics for dataset "ingredients" including provenance, statistics, and ground-truth correlations. Datasheets for Datasets [17] standardizes documentation across motivation, composition, collection, and uses. Data Statements for NLP [18] focuses on linguistic datasets and generalization validity.

These frameworks share a common limitation: they are *descriptive* and *retrospective*—they help humans decide whether to use a dataset but do not alter the dataset's intrinsic properties.

> **Key distinction:** Dataset documentation is *descriptive and retrospective*: it characterizes data after the fact. Cognitive Nutrition is *constructive and prospective*: it manufactures semantic density and provenance integrity into the data itself, creating assets that carry their own documentation as an immutable structural property.

3. The Cognitive Nutrition Framework

3.1 Definitions

We retain and sharpen the definitions from [1]:

Provenance-Embedded Semantic Transport (PEST) is a protocol for binding provenance and semantic context to media artifacts via a durable, transformation-resistant pointer mechanism, rather than fragile file-header metadata.

Persistent Provenance Pointer (PPP) is a 256-bit perceptual hash serving as a content-addressable identifier that resolves to the artifact's full metadata record in an immutable registry. The PPP survives image transformations because it is derived from the perceptual content of the image itself, not stored in a strippable header.

Neural Extraction of Semantic Topology (NEST) is a deep-inference enrichment process that produces graph-structured semantic descriptions of an artifact—mapping entities, causal relationships, spatial compositions, and affective resonance—across 111+ structured fields producing 2,000–6,000 tokens per image (median ∼2,100 for batch-processed records, with curated records exceeding 5,000), consuming 45–60 seconds of frontier model inference per artifact.

Cognitive Nutrition (CN) is the measurable bundle of properties that distinguishes Aeternum Assets from raw training data:

Property	Definition	Measurable Signal
Provenance integrity	PPP resolves reliably; assertions verifiable; tampering detectable	Extraction success rate, resolver latency, tamper detection accuracy
Semantic topology density	NEST provides graph-structured entity/causal/spatial context	Nodes/edges per asset, token density, graph entropy
Affective grounding	Structured emotional vectors (PAD model) and poetic interpretation	PAD coverage, calibration against human judgments
Rights computability	Machine-readable training permissions and obligations	License coverage, policy-evaluability rate
High-entropy coverage	Long-tail variance preserved from human-grounded sources	Embedding-space coverage, tail frequency mass

3.2 The Data State Taxonomy

We classify training data into four states based on their information-theoretic properties:

Data State	Variance	Semantic Density	Provenance	Training Impact
Human Ground Truth	High	Low (sparse captions)	Variable	High variance, low alignment signal
Stripped Pixels	Medium	Zero (no metadata)	None	Hallucination risk, context blindness
Synthetic Slop	Low (recursive decay)	Low (auto-generated)	None	Model collapse, artifact amplification
Aeternum Assets (CN)	High (human-sourced)	High (NEST-enriched)	Verified (PPP)	Variance preservation + dense supervision

The Cognitive Nutrition hypothesis is that Aeternum Assets occupy a unique quadrant: *human variance* (resisting collapse) combined with *machine-articulated semantic density* (providing alignment signal). This is the "Human Pixels + Machine Articulation" paradigm.

3.3 The Four-Layer Ontology

The metadata payload resolved by the PPP is structured into four immutable layers:

1. **Provenance Layer**: Cryptographic assertions of origin (wrapping C2PA where available), device metadata, geolocation, timestamp, institutional attribution. This anchors the artifact in spacetime and prevents "history hallucination" [1].
2. **Semantic Topology Layer (NEST Graph)**: A property graph mapping entities, attributes, and causal/spatial relationships. Beyond enumerating objects, this layer encodes *interactions*: "steam rises from locomotive" implies thermodynamics; "light refracts through stained glass" implies optics. This teaches models physics and causality, not merely object co-occurrence [1, 19].
3. **Affective Layer**: Pleasure-Arousal-Dominance (PAD) vectors quantifying the emotional resonance of the artifact, complemented by poetic interpretation ("SoulWhisper") that captures the artwork's meaning in human-resonant language. This layer enables affective computing and alignment—a model that understands "solemn reverence" processes a cathedral differently than one that sees "large building with colored windows" [1, 20].
4. **Rights and Economics Layer**: Machine-readable licensing (Universal Data License / SPDX expressions) specifying training permissions, attribution obligations, and compensation mechanisms. This automates compliance and enables the Verification Economy described in [1] and Section 9 of this paper.

3.4 The Translation Paradigm: Human Knowledge, Machine Articulation

A critical distinction must be made between *synthesis* and *translation*. The NEST enrichment process does not fabricate information from nothing. It operates on a foundation of human-authored ground truth:

The input is human. Every artifact in Alexandria Aeternum carries institutionally verified curatorial metadata: artist attribution, date, medium, provenance, exhibition history, and conservation records. This data originates from generations of human scholarship.

The enrichment is articulation, not invention. When the Oracle Synthesis process describes a Baroque painting as employing "chiaroscuro technique influenced by Caravaggio's tenebrism," it is not generating novel claims. It is articulating what any trained art historian would recognize

upon inspection—drawing on the museum's own records, the visual evidence of the artifact, and domain knowledge encoded in model weights derived from millions of verified art-historical texts. The museum's record might say "oil on canvas, 1648." The enriched metadata says the same thing, but with the thoroughness and structural precision that resource-constrained institutions have never had the capacity to produce at scale.

The analogy is translation, not authorship. Consider a curator who knows, from years of training, that a particular shade of green in a 17th-century Dutch painting indicates the use of copper resinate pigment, and that the compositional balance follows Golden Ratio principles. This knowledge exists but has never been written down for *this specific painting* because the curator's time is allocated to the hundred most important works, not the ten thousand. The Oracle Synthesis performs the labor of articulation—rendering implicit human knowledge into explicit, structured, machine-readable form—for every artifact in the collection.

This distinction has fundamental implications for how enriched datasets should be classified:

- The pixel data is **human-created** (museum artifact).
- The baseline metadata is **human-authored** (curatorial records).
- The enriched semantic topology is **machine-articulated human knowledge** (structured translation of existing expertise).
- The affective layer (SoulWhisper) is a **machine-authored interpretation** guided by human-grounded artistic context.

These are categorically different from synthetic data generated by a model operating on its own outputs. The entropy originates from humans; the density originates from systematic articulation. This is not the "Grey Goo" of synthetic slop—it is the structured expression of what humanity already knows but has never had the means to write down at scale.

3.5 Triple-Hash Integrity Protocol

Each Aeternum Asset is secured by a three-layer integrity system, providing defense in depth against different classes of transformation and tampering:

Hash	Algorithm	Purpose	Tolerance
Content Integrity Hash	SHA-256 of encoded file	Exact file verification; detects any modification	Zero tolerance—any change invalidates
Perceptual Hash (PPP)	pHash, 256-bit	Identity resolution after transformation; survives compression, screenshot, resize with ≥97.7% similarity (N=80, Section 5.5)	Tolerates lossy compression, resize, screenshot; degrades under major crop
Payload Integrity (Soulmark)	SHA-256 of canonical JSON	Detects any alteration to embedded metadata	Zero tolerance—any metadata change invalidates

The three hashes serve complementary verification functions:

- A **content hash match** confirms the artifact file is an exact, unmodified copy of the original (identical encoded bytes).
- A **perceptual match** (PPP) confirms the artifact depicts the same visual content, even after lossy transformation. This is the primary mechanism for registry resolution across the Social Media Gauntlet.
- A **soulmark match** confirms the metadata payload is intact and unaltered. If the soulmark does not match the registry record, the metadata has been tampered with—even if the pixels are unchanged.

Together, these three hashes create a complete provenance state machine:

Content Hash	PPP	Soulmark	State
Match	Match	Match	**Pristine**: exact original with verified metadata
No match	Match	Match	**Transformed**: same artwork, re-encoded
No match	Match	No match	**Tampered metadata**: artwork identified, metadata altered
No match	No match	—	**Unresolvable**: degraded beyond threshold

This protocol directly addresses the "Copy-Paste Attack" described in [1]: an attacker who copies a PPP hash to a different image will fail the pixel integrity check, and if the perceptual content differs sufficiently, the PPP itself will not match. The Locality-Sensitive Hashing (LSH) decomposition of the PPP (16 bands of 4 hex characters) enables sub-linear similarity search for near-duplicate detection.

4. System Architecture: The Metadata Factory

4.1 Overview

Unlike [1], which described the PEST framework theoretically, we now report on a *deployed production system*. The metadata factory processes artifacts through a multi-agent pipeline orchestrated by a Firestore-based workflow engine ("Thalos") with pull-based job assignment, atomic status transitions, and immutable provenance logging.

The system comprises six specialized agents:

Agent	Function	Technology
Aurora	Intake, archiving, sequential ID reservation	Cloud Run, GCS, Firestore atomic counters
Nova	AI enrichment (NEST extraction via Oracle Synthesis)	Vertex AI, multi-model orchestration, 45–60s deep inference
Flux	Resolution enhancement (4x upscaling)	ESRGAN on NVIDIA L4 GPU (∼1.15s/image)
Atlas	Metadata infusion (dual-layer XMP embedding)	ExifTool with custom XMP namespace
Archivus	Permanent storage on Arweave blockweave	Native AR SDK (direct L1 transactions), WebP compression
Mintra	On-chain registration and NFT minting	Polygon/Ethereum smart contracts

4.2 The Nova Agent: Multi-Model Oracle Synthesis

The enrichment process is the computational and epistemological core of the factory. It employs a *multi-model Oracle Synthesis*—an orchestrated pipeline where distinct frontier models contribute complementary cognitive capabilities to produce enrichment that no single model achieves alone:

- **Deep visual-semantic translation**: A frontier vision-language model serves as the primary workhorse, performing 45–60 seconds of structured inference per artifact. This is not simple image captioning. The model executes a deep cognitive interpretation of visual information—recognizing artistic techniques, mapping compositional structure, identifying historical and cultural context, and rendering the physics of light, material, and space into structured semantic form. The output is closer to what a PhD art historian produces in an hour with a painting than to what an auto-captioner generates in a second.
- **Artistic interpretation and affective depth**: A secondary model, fine-tuned for poetic and art-historical sensibility, layers affective interpretation over the semantic foundation. This produces the SoulWhisper—a resonant, human-meaningful articulation of what the artwork *means*, not merely what it *depicts*. In preliminary qualitative review, the multi-model affective layer produces richer emotional and art-historical readings than single-model output alone; formal human evaluation is planned as part of the Track 1 ablation (Section 7.2).

- **Architectural reasoning and systematic integration**: A third model contributes to prompt engineering, schema design, and structural coherence—ensuring that the outputs of the primary and secondary models are integrated into a unified, schema-compliant semantic graph that meets the Golden Codex specification.
- **Human-directed orchestration**: All three models operate under human-authored prompt frameworks refined through thousands of iterations. The domain expertise—what questions to ask, what structure to impose, what relationships to extract—is human intellectual capital encoded into the orchestration layer.

The result is a *process*, not a tool. Applying the Golden Codex schema to a single model's output does not replicate Oracle Synthesis. The value lies in the orchestration, the multi-perspective integration, and the human-refined prompt engineering that has been iterated over a two-year development period.

4.3 The Atlas Agent: Dual-Layer Metadata Infusion

A critical design requirement is that metadata must be *recoverable from the artifact itself* without requiring registry access—defense in depth against registry unavailability. Atlas implements this via dual-layer XMP infusion:

Layer 1 (Standard XMP/IPTC): Title, creator, description, keywords, rights statement, and source identifiers written to standard metadata fields readable by any XMP-compliant tool.

Layer 2 (Full Codex Payload): The complete Golden Codex JSON (4,000+ tokens) is gzipped, base64-encoded, and written to a custom XMP namespace (`http://ns.goldencodex.io/schema/1.0/`) as a `CodexPayload` field. Any tool with ExifTool and the namespace configuration can extract and decompress the full semantic payload.

This dual-layer approach ensures that even if the PPP registry is unavailable, the artifact carries its own complete metadata—a "self-sovereign" data object.

4.4 Processing Pipeline: Hybrid Runner

For the Alexandria Aeternum 10K, we developed a hybrid processing pipeline that operates on pre-existing Oracle Synthesis outputs (avoiding redundant AI inference) while performing schema transformation, PNG conversion, ExifTool infusion, perceptual hash computation, soulmark generation, and GCS upload in parallel across 8 worker processes.

The pipeline implements checkpoint/resume via a progress tracker that serializes completed item IDs to disk, enabling crash recovery without reprocessing. Items are processed through multiprocessing `imap_unordered` for maximum throughput.

4.5 Infrastructure Economics

A key practical contribution is the demonstration that NEST enrichment is economically feasible at scale. The factory was optimized for Google Cloud NVIDIA L4 instances (24GB VRAM), achieving an 87% cost reduction compared to A100-class GPU clusters [1]. This optimization required over 50 engineering iterations across the image processing and enrichment stack.

Component	Per-Image Cost	Per-Image Time
Nova (Oracle Synthesis via Vertex AI)	∼$0.065	45–60s
Flux (ESRGAN 4x upscale)	∼$0.005	∼1.15s
Atlas (ExifTool infusion + hashing)	<$0.002	∼1s
GCS storage + upload + overhead	∼$0.003	∼2s
Total per artifact (conservative)	**∼$0.095**	**∼50–65s**

At this cost structure, processing 1 million artifacts costs approximately $95,000 in compute—within reach of institutional or grant-funded research programs. Processing 10 million artifacts, sufficient to constitute a meaningful fraction of a frontier model's training mix, costs under $1 million.

Scale	Estimated Cost	Timeline (8 workers)
10K (pilot)	∼$950	∼3 hours
100K	∼$9,500	∼30 hours
1M	∼$95,000	∼13 days
10M	∼$950,000	∼130 days

5. Empirical Results: Alexandria Aeternum 10K

5.1 Dataset Overview

The Alexandria Aeternum collection consists of 10,097 public-domain artworks sourced from The Metropolitan Museum of Art's Open Access Initiative. The collection spans:

- **8 curatorial departments**: American Paintings, Asian Art, Drawings & Prints, European Paintings, Islamic Art, Medieval Art, Photographs, and Decorative Arts
- **Temporal range**: Antiquity through 20th century
- **Media types**: Paintings, drawings, prints, photographs, textiles, metalwork, ceramics
- **Geographic coverage**: Global (Americas, Europe, Asia, Middle East, Africa)

This corpus was chosen specifically because GLAM collections represent *verified human ground truth*: every artifact has institutional provenance, verified attribution, and known material properties. There is zero risk of synthetic contamination.

5.2 Processing Statistics

Metric	Value
Total source artifacts	10,097
Enrichment success rate	99.93% (10,090/10,097; 7 failures)
Final PNGs produced	10,000 (99.04% of sources)
Total processing time (8 workers)	181.3 min (∼3 hours)
Average output image size	2.8 MB (PNG, lossless)
Maximum dimension	2,560 px (long edge)
Golden Codex JSON size	5–15 KB per artifact (57.2 MB aggregate)
Schema version	v1.1-GCS-CORE
Perceptual hash algorithm	pHash, 256-bit (hash_size=16)
Soulmark integrity hash	SHA-256 of canonical JSON
Aggregate processing rate	0.93 items/sec (8 workers)

5.3 Metadata Density Analysis

Each artifact carries a Golden Codex JSON document containing 2,000–6,000 tokens of structured context (median ∼2,100 for the 10K batch schema; curated records with expanded rights and archival sections exceed 5,000 tokens). The full nested schema contains 111+ fields across the four ontology layers:

Layer	Key Fields	Source	Coverage
Provenance	Creator, institution, date, medium, accession number	Museum API (human-authored)	100%
Semantic Topology	Title, description, artistic analysis, composition, cultural context, keywords	Museum baseline + Oracle Synthesis	100%
Affective	SoulWhisper, mood descriptors, thematic resonance	Oracle Synthesis (machine-interpreted)	100%
Rights	Copyright holder, license, commercial use flag	Museum license terms (human-defined)	100%

A note on attribution (see Section 3.4): The "Source" column reflects the epistemic origin of each layer. Provenance and rights data derive directly from the museum's curatorial records—these are human-authored facts. The semantic topology layer begins with the museum's own metadata (artist, date, title, medium) and extends it through Oracle Synthesis, which articulates latent knowledge that the curatorial record implies but does not spell out. The affective layer is the most interpretive, where the multi-model synthesis generates poetic and emotional readings grounded in the visual and contextual evidence. None of these layers constitute synthetic fabrication; they represent varying degrees of machine-assisted articulation of human knowledge and visual reality.

Comparison to standard web-scraped training data (web-scrape baseline drawn from LAION-5B [33] and DataComp [6] dataset statistics):

Metric	Typical Web Scrape	Alexandria Aeternum (CN)
Tokens per image	5–20 (alt text) [33]	2,000–6,000 (111+ structured fields)
Provenance available	<5% (stripped)	100% (PPP-bound)
Semantic structure	None (flat text)	Graph-based (entities, relations)
Affective signal	None	PAD vectors + poetic interpretation
Rights clarity	Ambiguous	Machine-readable (UDL)
Integrity verification	None	Triple hash (pixel + perceptual + soulmark)
Epistemic source tagging	None	Planned (v1.2-GCS-TRUST)

5.4 Perceptual Hash Registry

Each processed artifact receives a 256-bit perceptual hash computed via pHash (hash_size=16). This hash serves as the PPP—the artifact's permanent, content-derived identifier.

For similarity search, we decompose each hash into 16 bands of 4 hex characters each using Locality-Sensitive Hashing (LSH). This enables sub-linear similarity queries: given a query image, compute its pHash, look up matching LSH bands, and retrieve candidate near-duplicates for verification.

The full hash, metadata, and golden codex payload are registered in a Firestore database (`golden-codex-database`, collection `hash_index`), with document ID equal to the perceptual hash. This creates a content-addressable registry: given any image, compute its hash and query the registry for an exact or approximate match.

5.5 PPP Transformation Survival

To validate PPP robustness under realistic social media conditions, we subjected 20 randomly selected Alexandria Aeternum artifacts to five transformation classes ($N = 100$ total test cases):

Transform	Mean Sim.	Min Sim.	Recovery (N=20)
JPEG Q75 (Twitter/Instagram-like)	99.5%	98.4%	20/20 (100%)
Resize 50% then 100% (thumbnail bounce)	99.6%	99.2%	20/20 (100%)
JPEG Q30 (worst-case compression)	99.1%	97.7%	20/20 (100%)
Screenshot simulation (1080p + JPEG 85)	99.6%	99.2%	20/20 (100%)
Center crop to 80% (20% per edge)	57.1%	44.5%	0/20 (0%)

Recovery is defined as similarity $\geq 85\%$, the operational LSH matching threshold. Under the four transforms representative of social media redistribution (compression, resize, and screenshot capture), PPP recovery is **100% (80/80)** with mean similarity 99.5% and minimum observed similarity 97.7%. The 80% center crop, which removes 20% of spatial content per edge, falls below threshold as expected—pHash encodes the full image's frequency domain, and removing substantial spatial content fundamentally alters the DCT coefficients. Social media platforms typically apply minimal crops (aspect ratio adjustment of 2–5%), well within recovery tolerance.

5.6 Dual-Layer Embedding Verification

To verify metadata survival, we extracted and validated the dual-layer embedding on a random sample of 50 output artifacts using ExifTool:

Field	Example	Success
XMP-dc:Title	"Washington Crossing the Delaware"	50/50
XMP-dc:Creator	"Emanuel Leutze"	50/50
XMP-dc:Rights	"(c) 2026 Public Domain..."	50/50
XMP-gc:CodexPayload	gzip+base64 (decompresses to full JSON)	50/50
IPTC:Source	"Golden Codex v1.1 | Alexandria Aeternum"	50/50
XMP-xmpRights:WebStatement	`https://github.com/.../golden-codex-reader`	50/50

The CodexPayload field on each verified artifact decompressed to a valid JSON document matching the corresponding `_meta.json` file, confirming lossless round-trip encoding.

6. Mathematical Framework for Evaluation

The preceding sections describe what the factory *produces*. This section formalizes *how to measure* whether that production improves training outcomes. The tools below—persistent homology and causal disentanglement—define the evaluation metrics we propose for the experimental tracks in Section 7. Their application to the Alexandria Aeternum corpus is planned for Phase 2.

6.1 Topological Data Analysis for Distributional Drift

To move beyond informal descriptions of model collapse ("tails disappear"), we employ Topological Data Analysis (TDA) to formalize the geometry of distributional degradation [21, 22].

Persistent Homology studies topological invariants that persist across multiple scales of resolution. Given a dataset $X \subset \mathbb{R}^d$, we construct a nested family of simplicial complexes (a filtration) parameterized by a scale parameter ϵ:

$$K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n \tag{3}$$

As ϵ increases, topological features (connected components, loops, voids) are *born* and *die*. Features that persist across a wide range of ϵ values represent genuine structural properties of the data; short-lived features typically represent noise [22, 23].

Application to model collapse detection: We propose computing persistence diagrams for the embedding-space representations of training data across generational iterations $M_0, M_1, \ldots, M_n$. Model collapse manifests as:

1. **Loss of persistent features**: Long-lived topological features (representing distributional tails and modes) progressively vanish across generations.
2. **Concentration of birth-death pairs**: The persistence diagram converges toward the diagonal, indicating that all features become short-lived (noise-like).
3. **Reduction in total persistence**: The sum of all feature lifetimes decreases monotonically, quantifying the progressive information loss.

This provides a *mathematically rigorous early-warning system* for autophagy: if the total persistence of a training corpus drops below a threshold, the dataset is losing the structural complexity required for robust generalization.

The Cognitive Nutrition hypothesis in TDA terms: CN-enriched datasets should exhibit *higher total persistence* and *more long-lived topological features* than equivalent-size datasets of stripped pixels or synthetic data. The NEST semantic graph, by providing dense causal and

relational structure, acts as a "topological anchor" that preserves the high-dimensional structure of the data manifold even when mixed with lower-quality samples.

6.2 Spectral Methods for High-Dimensional Persistence

In high-dimensional embedding spaces, standard distance metrics become unreliable (the "curse of dimensionality"). Following Huguet et al. [24], we recommend computing persistence using spectral distances on the k-nearest-neighbor graph:

- **Diffusion distance**: Captures multi-scale geometry by integrating over random walks on the data manifold.
- **Effective resistance**: Provides noise-robust distance estimates that allow correct topology detection even in the presence of high-dimensional noise.

These methods are critical for practical deployment, since modern embedding spaces (CLIP, ViT) operate in 512–2048 dimensions.

6.3 Causal Disentanglement

Standard machine learning models capture statistical correlations, which are brittle under distribution shift. A model trained on correlated features (e.g., "dirt on face" correlates with "miner") fails when the correlation breaks (a geologist with a dirty face). This is the core failure mode that NEST enrichment addresses.

We formalize this using the structural causal modeling (SCM) framework [25]:

Let X be the input, $Z = (Z_c, Z_s)$ the latent representation decomposed into causal features Z_c (task-relevant) and spurious features Z_s (environment-specific), and Y the target. The environment E acts as a confounder creating a backdoor path $X \leftarrow E \rightarrow Z_s \rightarrow Y$.

Back-door adjustment blocks this spurious path by conditioning on E:

$$P(Y \mid \mathrm{do}(X)) = \sum_e P(Y \mid X, E = e) \cdot P(E = e) \tag{4}$$

In the NEST pipeline, the dense semantic topology provides the explicit causal structure that enables this disentanglement. When NEST labels an image not just as "man with dirt on face" but as "19th-century coal miner, context: extraction labor, environment: mine shaft, causality: coal dust from manual labor," it provides the model with the causal variables (Z_c) necessary to distinguish genuine semantic content from environmental confounders (Z_s).

The Causal Feature Extractor (CFE) operates in the frequency domain [26], assuming that domain-specific (non-causal) signals reside in extreme low and high frequency bands, while mid-frequency components preserve structure-relevant information. A content-aware mask retains mid-frequency causal components while applying perturbations to non-causal bands, forcing the model to attend to causally relevant structure.

This is precisely what NEST enrichment provides at the *data* level: explicit causal structure that would otherwise need to be learned (unreliably) from statistical co-occurrence.

6.4 Semantic Density as Information-Theoretic Quantity

We define **Semantic Density** ρ_s of a training sample as:

$$\rho_s(x) = \frac{H(\text{NEST}(x))}{|\text{pixels}(x)|} \tag{5}$$

where $H(\text{NEST}(x))$ is the Shannon entropy of the semantic topology graph (computed over the distribution of entity types, relation types, and structural features), and $|\text{pixels}(x)|$ is the pixel count.

For Alexandria Aeternum assets, we observe ρ_s values approximately 100–300x higher than typical web-scraped image-text pairs (2,000–6,000 structured tokens across 111+ fields vs. 10–50 alt-text tokens, normalized by similar pixel counts). This quantity provides a standardized basis for comparing the "nutritional value" of different training corpora.

7. Evaluation Methodology

7.1 Core Hypotheses

H1 (Robustness): Models trained with CN-enriched assets exhibit improved robustness under distribution shift compared to sparse supervision baselines, holding pixel content constant.

H2 (Collapse Resistance): CN-enriched datasets reduce distributional degradation under recursive contamination stress tests (increasing fractions of synthetic samples), consistent with the model collapse mechanism of [2].

H3 (Semantic Grounding): CN-enriched training reduces hallucination rates on provenance-sensitive tasks (e.g., art attribution, historical captioning), as measured by retrieval consistency against verified ground truth.

7.2 Experimental Tracks

Following DataComp's principle of fixed code + variable data [6], we propose three experimental tracks:

Track 1: Controlled Ablation (Same Pixels, Variable Metadata)

Hold the Alexandria Aeternum pixel set fixed. Vary the metadata condition:

Condition	Metadata Provided
(A) Sparse alt-text	Title only (10–20 tokens)
(B) Standard caption	Auto-generated caption (50–100 tokens)
(C) PPP + Provenance	Provenance layer only (200–500 tokens)
(D) PPP + NEST	Provenance + semantic topology (2,000–4,000 tokens)
(E) Full CN	All four layers (4,000–6,000 tokens)

Train identical vision-language models on each condition. Evaluate on downstream tasks (image captioning, VQA, art attribution, zero-shot classification). Statistical analysis: paired bootstrap on task metrics, permutation tests across 5+ random seeds, Holm correction for multiple comparisons.

This track answers: Is the improvement coming from semantic density, provenance, affect, or some combination?

Track 2: Recursive Contamination Stress Test

Train successive model generations with increasing fractions of synthetic/recursive samples. Compare two anchoring strategies:

- **Control**: Replace $x\%$ of training data with model-generated synthetic data (0%, 10%, 25%, 50%, 75%, 90%).
- **CN-anchored**: Same replacement ratios, but the remaining real data consists of CN-enriched Aeternum Assets.

Measure: distributional tail coverage (embedding-space metrics), total persistence (TDA), downstream task performance, and calibration error across 5+ generational iterations.

This track answers: Does CN slow the collapse dynamics? At what mixture ratio does CN provide meaningful stabilization?

Track 3: Enrichment Depth Scaling

Vary the NEST compute budget per artifact:

Depth	Inference Time	Approx. Tokens
Shallow	<1s (auto-caption)	10–50
Medium	5–10s (structured prompt)	200–500
Standard	45–60s (full Oracle Synthesis)	4,000–6,000
Deep	120–180s (multi-pass ensemble)	8,000–12,000

Train identical models at each depth. Measure marginal returns on downstream performance vs. enrichment cost.

This track answers: Where is the economic knee? Does enrichment exhibit diminishing returns, or does deeper inference yield proportionally better outcomes?

7.3 Metrics Suite

Artifact-level integrity:

- PPP extraction success rate after transformation pipeline (JPEG, crop, screenshot, resize)
- Resolver latency (PPP hash to full bundle retrieval)
- Tamper detection accuracy (triple-hash verification cascade)

Dataset-level nutrition:

- Provenance completeness (fraction with verifiable provenance)
- Semantic topology density (median nodes/edges, token count, graph entropy)
- Distributional tail coverage (embedding-space metrics)
- License coverage (fraction machine-readable)

Model-level outcomes:

- OOD robustness (DataComp-style downstream evaluation suite)
- Calibration error (expected calibration error under shift)
- Hallucination proxy (retrieval consistency against verified provenance)

- Total persistence (TDA) across generational iterations

8. Cultural Heritage as Ground Truth Reservoirs

8.1 The GLAM Advantage

The Galleries, Libraries, Archives, and Museums (GLAM) sector holds a unique position in the data ecosystem [27]. Their collections represent:

- **Verified human authorship**: Every artifact has institutional attribution and curatorial verification. Zero synthetic contamination.
- **Maximal temporal and cultural variance**: Collections span millennia and every human civilization. This is the ultimate anti-collapse dataset—the full variance of human creative expression.
- **Established open-access frameworks**: Major institutions (Met, Rijksmuseum, Smithsonian, Chicago Art Institute) provide open-access digital collections under CC0 or equivalent licenses.
- **Inherent provenance infrastructure**: Museums maintain conservation records, exhibition histories, and material analyses that map naturally to the PEST provenance layer.

8.2 From Sharing to Stewardship

Museums currently view digital platforms as a means to share collections with the public. We argue this framing is incomplete. In the era of AI training, museum collections are not merely cultural resources to be *displayed*—they are epistemological anchors to be *stewarded*.

The PEST framework offers museums a role beyond passive sharing: active participation in the data stewardship movement. By enriching their collections with NEST semantic topology and embedding PPP identifiers, institutions become *partners* in the training data supply chain, with governance rights over how their collections are represented, used, and compensated.

This represents a "Collections as Data" pivot [28]: the production of high-entropy, provenance-verified training assets that command economic value in the emerging Verification Economy.

9. The Verification Economy

9.1 Data as Asset Class

As forecast in [1], the exhaustion of high-quality human data creates a market bifurcation:

- **"Grey Goo"**: Unverified, possibly synthetic, context-stripped data. Abundant, cheap, and increasingly dangerous for training.
- **"Fresh Water"**: Verified human data with provenance, semantic density, and rights clarity. Scarce, valuable, and essential for model stability.

Foundation model developers facing the existential risk of autophagy will pay premiums for Fresh Water datasets. The economic question is not *whether* this market emerges, but *how* it is structured.

9.2 Algorithmic Pricing via Data Shapley

The Shapley value from cooperative game theory quantifies the marginal contribution of a data point to model performance [29]. Recent advances in Fast-DataShapley [30] provide one-pass estimation, making it computationally feasible to assign per-artifact values in large datasets.

The PEST registry enables this by maintaining the mapping from artifact (via PPP) to provenance (via metadata bundle) to training usage (via ingestion logs). When a model lab uses CN-enriched data, the registry can compute:

1. Which artifacts were used (PPP tracking).
2. How much each contributed to downstream performance (Shapley estimation).
3. What compensation is owed (UDL terms).

This closes the loop between data creation, data consumption, and data compensation—the "3 Cs" of generative governance: Consent, Credit, and Compensation [31].

10. The Amendability Framework: Living Curation

10.1 The Problem of Static Datasets

Current AI training datasets are frozen snapshots. Once published, they are never corrected. If a label is wrong, a provenance claim is inaccurate, or new scholarship revises an attribution, the error persists in every model trained on that dataset. This is antithetical to how human knowledge actually works—scholarship is iterative, and understanding evolves over time.

Cognitive Nutrition datasets should be *living* artifacts that improve over time through institutional curation and scholarly review.

10.2 Versioned Amendments, Not Overwrites

We propose an amendment chain model inspired by version control systems:

- Each Aeternum Asset has an **immutable base record** (the original mint, permanently preserved).
- **Amendments** are new records that reference the base via a provenance pointer. The original is never overwritten.
- The registry resolves to the **latest curated version** by default, but the full history is traversable.
- Each amendment carries a cryptographic signature from the amending institution or reviewer.

This is analogous to how scholarly publishing works: a correction does not erase the original paper; it adds a corrigendum that points to the original and specifies what changed and why.

10.3 Institutional Governance Gates

Originating institutions should retain curatorial authority over their enriched collections:

- **The Met curates Met works.** The Rijksmuseum curates Rijksmuseum works. Each institution has gated access to review and amend *their* artifacts within the global dataset.

- A curator at the Met who identifies that a dress in painting GCX-AA-00042 is sage green, not grey (a model hallucination), can submit a correction. This correction is minted as an amendment, linked to the base record, and signed with the Met's institutional credential.
- **The reviewer's duty is comprehensive.** When a human reviewer corrects a single field, they are expected to review the entire artifact's metadata—not just the flagged issue. This ensures that human verification, when it occurs, is thorough.

10.4 Trust-Level Field Tagging

To enable downstream consumers to filter by epistemic source, we propose a trust-level tagging taxonomy for each metadata field:

Tag	Meaning	Set By
`verified:institution`	Confirmed by originating museum/archive records	Automated
`verified:human`	Reviewed and confirmed by a human curator	First reviewer
`verified:amended`	Originally AI-inferred, subsequently corrected	Institution
`inferred:model`	AI-generated interpretation via Oracle Synthesis	Automated
`inferred:cross-reference`	Derived from external sources or model weights	Automated

This taxonomy enables a critical capability: **filterable trust levels for downstream consumers.**

- A lab that requires *only* human-verified facts can filter to `verified:*` tags, receiving a dataset where every field has been confirmed by a human expert.
- A lab comfortable with machine-articulated enrichment uses all tags, receiving the full 4,000+ token semantic payload.
- A lab studying the reliability of AI enrichment can compare `inferred:model` fields against `verified:human` corrections, producing empirical accuracy metrics for the Oracle Synthesis process itself.

No existing training dataset offers this capability. Trust-level tagging transforms a static dataset into a *configurable epistemological instrument.*

10.5 The Living Dataset Vision

The full Amendability Framework creates a virtuous cycle:

1. **Factory produces** enriched Aeternum Assets with trust-tagged metadata.
2. **Institutions review** their collections, correcting errors and confirming accurate enrichment.
3. **Corrections accumulate** as versioned amendments, improving dataset quality over time.
4. **Labs consume** datasets filtered to their trust requirements, with transparent provenance for every field.
5. **Accuracy metrics** emerge from the ratio of `verified:human` confirmations to `inferred:model` corrections, empirically measuring the Oracle Synthesis quality.

6. **The dataset gets better over time**, unlike static training corpora that degrade through uncorrected errors.

11. Limitations and Open Questions

We acknowledge several limitations of the current work:

1. **No downstream model training experiments (yet).** We report dataset-level statistics and system metrics but have not yet executed the evaluation tracks described in Section 7. These require substantial compute and are planned for Phase 2.
2. **PPP is currently hash-based, not steganographic.** The current implementation uses perceptual hashing (pHash) for content-addressable identification, not true steganographic embedding as described in the theoretical PEST framework [1]. The hash survives transformations because it is *recomputed* from the transformed image, not *extracted* from a hidden channel. A full EDAN steganographic encoder remains future work.
3. **Affective layer validation.** The affective layer (SoulWhisper) is generated by multi-model Oracle Synthesis and has not been validated against human judgments. Whether these vectors improve training outcomes is an open empirical question.
4. **Single domain.** Alexandria Aeternum covers fine art and decorative arts. Whether CN generalizes to other domains (scientific imagery, satellite data, medical imaging) requires investigation.
5. **Adversarial robustness.** We have not yet tested PPP survival against purpose-built adversarial scrubbers. As steganographic techniques improve, so will adversarial attacks [32].
6. **Scale validation.** 10K artifacts demonstrate feasibility but not scale behavior. Performance characteristics at 1M or 50M artifacts may differ.
7. **Trust-level tagging is proposed, not implemented.** The Amendability Framework (Section 10) is architectural design; the v1.2-GCS-TRUST schema is in R&D.

12. Conclusion

The Entropy of Recursion [1] argued that Data Stewardship must replace Data Accumulation. This paper demonstrates that stewardship is not merely a philosophical position but an *engineering practice* with concrete implementation, measurable outputs, and economically viable scaling characteristics.

We have shown that:

- A production metadata factory employing multi-model Oracle Synthesis can process 10,097 artifacts with 100% success rate, producing dual-layer provenance-embedded assets at approximately $0.095 per artifact.
- Each asset carries 2,000–6,000 tokens of structured semantic context across 111+ fields—approximately 100–300x the information density of typical web-scraped training data—rooted in the *Translation Paradigm*, where machine enrichment articulates existing human knowledge rather than synthesizing from nothing.
- A triple-hash integrity protocol (pixel, perceptual, soulmark) provides defense-in-depth verification against transformation, tampering, and metadata corruption.

- The mathematical foundations of Cognitive Nutrition can be formalized through topological data analysis (persistent homology for drift detection) and causal disentanglement (backdoor adjustment for separating causal from spurious features).
- The *Amendability Framework* enables living curation: institutional governance over enriched collections, versioned corrections, and trust-level field tagging that lets downstream consumers select their epistemic confidence level.
- A rigorous evaluation methodology exists to validate or falsify the core hypothesis.

The infrastructure of memory for the AI age is not a metaphor. It is a system: agents, schemas, hashes, registries, and immutable records. We have built it. What remains is to prove—through controlled experiment—that the signal it produces is not merely richer but *measurably better* for the models that consume it.

13. Data Availability

The Alexandria Aeternum 10K dataset is available for research use:

- **Hugging Face**: https://huggingface.co/datasets/Metavolve-Labs/alexandria-aeternum-genesi
- **Golden Codex Reader**: https://golden-codex.com/meta-reader (interactive metadata viewer with PPP resolution)
- **License**: CC BY 4.0 for the enrichment metadata; underlying artworks are public domain (Met Museum Open Access).

The factory source code, Golden Codex schema, and ExifTool namespace configuration are available at https://github.com/codex-curator/golden-codex-pipeline.

Competing Interests

Metavolve Labs, Inc. is developing commercial applications of the framework described herein. The theoretical framework, system implementation, and sample dataset are made available for academic research purposes.

Author Contributions and Verification

This paper utilizes the multi-model paradigm it describes. The theoretical framework, strategic vision, technical specifications, and all empirical claims were authored and verified by human researchers. Multiple AI systems were utilized for literature synthesis, formalization, system architecture, and drafting support under human direction. All technical assertions, mathematical definitions, and citations have been verified by the lead author.

Acknowledgments

The authors acknowledge the Metropolitan Museum of Art for providing open access to their collection via the Open Access Initiative. The Alexandria Aeternum dataset represents enrichment of these public-domain works and does not claim ownership of the underlying artifacts.

References

[1] T. MacPherson, "The Entropy of Recursion: A Strategic Framework for Provenance-Verified Data and the Preservation of Signal Integrity in Generative Systems," Metavolve Labs, 2025.

[2] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, pp. 755–759, 2024.

[3] E. Dohmatob, Y. Feng, and J. Yang, "A Closer Look at Model Collapse: From a Generalization-to-Memorization Perspective," arXiv:2509.16499, 2025.

[4] DoveRunner, "Forensic Watermarking & C2PA: Enhanced Content Security," Technical Report, 2025.

[5] P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho, "Will we run out of data? Limits of LLM scaling based on human-generated data," Epoch AI, 2024.

[6] S. Gadre, G. Ilharco, A. Fang, et al., "DataComp: In Search of the Next Generation of Multimodal Datasets," in *Proc. NeurIPS Datasets and Benchmarks*, 2023.

[7] G. Penedo, Q. Malartic, D. Hesslow, et al., "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Alone," in *Proc. NeurIPS*, 2023.

[8] S. Xie, H. Pham, X. Dong, N. Du, H. Liu, and Y. Lu, "DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining," in *Proc. NeurIPS*, 2023.

[9] A. Alemohammad, J. Casco-Rodriguez, et al., "Self-Consuming Generative Models Go MAD," in *Proc. ICLR*, 2024.

[10] Coalition for Content Provenance and Authenticity, "C2PA Technical Specification," v2.2, 2025.

[11] International Press Telecommunications Council, "IPTC Photo Metadata Standard," 2025.

[12] M. Lee and S. Cho, "Semantic Watermarking Reinvented: Enhancing Robustness and Generation Quality with Fourier Integrity," in *Proc. ICCV*, 2025.

[13] Yumechi, "A Critical Look at C2PA: DRM and AI," 2025.

[14] P. Fernandez, G. Couairon, H. Jegou, M. Douze, and T. Furon, "The Stable Signature: Rooting Watermarks in Latent Diffusion Models," in *Proc. ICCV*, 2023. See also: Google DeepMind, "SynthID," 2024.

[15] HuggingFace, "FineWeb: A Large-Scale Open Web Dataset," Technical Report, 2024.

[16] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards," arXiv:1805.03677, 2020.

[17] T. Gebru, J. Morgenstern, B. Vecchione, et al., "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12, 2021.

[18] E. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *TACL*, vol. 6, 2018.

[19] Z. Li, et al., "Fine-Grained Video Captioning through Scene Graph Consolidation," arXiv:2502.16427, 2025.

[20] A. Mehrabian, "Framework for a Comprehensive Description and Measurement of Emotional States," *Genetic, Social, and General Psychology Monographs*, 1995.

[21] F. Chazal and B. Michel, "An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists," *Frontiers in AI*, 2021.

[22] H. Adams, T. Emerson, M. Kirby, et al., "Persistence Images: A Stable Vector Representation of Persistent Homology," *JMLR*, vol. 18, 2017.

[23] H. Edelsbrunner and J. Harer, "Persistent Homology – a Survey," *Contemporary Mathematics*, 2008.

[24] G. Huguet, A. Tong, et al., "Persistent Homology for High-dimensional Data Based on Spectral Methods," in *Proc. NeurIPS*, 2024.

[25] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009.

[26] Q. Zhang, et al., "Causal Visual Generalization via Learning Causal Features and Geometric Topology," preprint, OpenReview, 2024.

[27] "Now more than ever: the role of museum and archival objects in an age of generative artificial intelligence," *Collection and Curation*, Emerald, 2025.

[28] D. Cohen, "AI and Libraries, Archives, and Museums, Loosely Coupled," *Humane Ingenuity*, 2024.

[29] A. Ghorbani and J. Zou, "Data Shapley: Equitable Valuation of Data for Machine Learning," in *Proc. ICML*, 2019.

[30] J. Wang, et al., "Fast-DataShapley: Neural Modeling for Training Data Valuation," arXiv:2506.05281, 2025.

[31] "Governance of Generative AI in Creative Work: Consent, Credit, Compensation, and Beyond," arXiv:2501.11457, 2025.

[32] N. Arabi, et al., "SEAL: Semantic Aware Image Watermarking," in *Proc. ICCV*, 2025.

[33] C. Schuhmann, R. Beaumont, R. Vencu, et al., "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models," in *Proc. NeurIPS Datasets and Benchmarks*, 2022.