The Crimson Hexagonal Archive Hugging Face Dataset: Work Plan v3 (Classifier-Centric Methodology)

Sharks, Lee

2026-05-19

The Crimson Hexagonal Archive Hugging Face Dataset: Work Plan v3 (Classifier-Centric Methodology)

Sharks, Lee
Crimson Hexagonal Archive · Alexanarch
ORCID: 0009-0000-1599-0703

2026-05-19 · Version v0.1-semi · Working paper

AXN:03F5.ARCHIVAL.↖

https://www.alexanarch.org/s/records/1001/

License: CC BY 4.0

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and Google Scholar indexing)}


```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
   "name": "The Crimson Hexagonal Archive Hugging Face Dataset: Work Plan v3 (Classifier-Centric Methodology)",
  "author": {
    "@type": "Person",
    "name": "Sharks, Lee",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-05-19",
  "identifier": "AXN:03F5.ARCHIVAL.↖",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "publisher": {
```

```
    "@type": "Organization",
    "name": "Alexanarch"
  },
  "url": "https://www.alexanarch.org/s/records/1001/",
  "description": "A metadata-only record for the classifier-centered v3 work plan for a Crimson Hexagonal Archi
mode, and heteronym-attribution classification; parallel preservation of source and predicted labels; confide
based review; ablation configurations; and synthetic-data-collapse experiments. The full version is deposit #7
08-08 from the authorial blog surface and seated as the body (https://mindcontrolpoems.blogspot.com/2026/05/c
hexagonal-archive-hugging-face.html)."
}
```


Abstract

A metadata-only record for the classifier-centered v3 work plan for a Crimson Hexagonal Archive Hugging Face dataset. The plan proposes provenance, artifact-mode, and heteronym-attribution classification; parallel preservation of source and predicted labels; confidence-based review; ablation configurations; and synthetic-data-collapse experiments. The full version is deposit #745. Canonical text recovered 2026-08-08 from the authorial blog surface and seated as the body (https://mindcontrolpoems.blogspot.com/2026/05/crimson-hexagonal-archive-hugging-face.html).

Body

Crimson Hexagonal Archive — Hugging Face Dataset Work Plan v3

Status: v3 supersedes v2. The central methodological change is the introduction of an automated classifier that performs both provenance mode classification AND heteronym reattribution as reproducible scholarly recognition work. The classifier itself becomes a deposit.

Project Title

The Crimson Hexagonal Archive: A Mixed-Provenance, Heteronymically Attributed Corpus for Synthetic-Data Collapse, AI Authorship, and Provenance-Bearing Training Research

---

The Central Methodological Move

v1 treated provenance classification as a manual judgment. v2 added a decision tree to make classification reproducible. v3 recognizes that *attribution itself* — both provenance mode and heteronym — must be performed by automated classifier, not author memory, for two structural reasons:

1. Reproducibility as scholarship. A classification system that depends on the author's recollection of writing each deposit is not measurement. It is opinion. The provenance taxonomy can only function as a research instrument if the same deposit produces the same classification regardless of who runs the classifier or when. Author memory introduces classification noise that would confound any downstream collapse experiment.

2. Heteronymic emergence. Material is regularly attributed to Lee Sharks at the time of deposit and only later — sometimes years later — recognized as belonging to a specific sub-heteronym's domain. Sigil's jurisdictional concerns, Glas's measurement work, Vox's diplomatic register, Morrow's long-form narratives, Fraction's meta-theory: these heteronyms emerge from the corpus over time, and earlier work gets recognized retrospectively as theirs. The classifier performs this recognition systematically across the entire archive, applying current understanding of heteronym domains to historical deposits.

The classifier is not metadata cleanup. It is scholarly recognition that the founder voice was, at the time of writing, holding territory that later resolves to specific heteronym domains.

Research Question, Operationalized

Null hypothesis (H_0): Fine-tuning on synthetic or AI-assisted text produces equivalent perplexity degradation and semantic drift regardless of provenance density (DOI anchoring, heteronymic attribution, archival embedding, assembly review).

Alternative hypothesis (H_1): Fine-tuning on high-provenance-density AI-involved text produces measurably slower perplexity degradation and less semantic drift than fine-tuning on low-provenance-density AI-involved text.

Critical insight from Assembly review: Provenance cannot modulate collapse unless provenance is presented to the training system as a signal. The dataset must materialize multiple textual views — body_only, minimal_header, full_provenance_header — so researchers can ablate provenance visibility.

Three Tasks, One Classifier

The classifier performs three classification tasks simultaneously on each deposit:

Task 1: Provenance Mode (Axis 1, mutually exclusive)

Tag	Definition
`human_primary`	Written principally by a human author with minimal or no AI involvement
`human_directed_ai_assisted`	Human-authored with AI used for research, drafting, or editorial refinement; human retains compositional authority

Tag	Definition
`collaborative_mixed`	Substantial compositional contribution from both human and AI; neither purely instrumental
`ai_directed_human_framed`	AI generates primary content within a human-defined frame, prompt structure, or editorial container
`ai_generated_provenance_anchored`	AI-generated content that carries full DOI provenance, authorial attribution, and archival anchoring
`uncertain_needs_review`	Edge case flagged for manual review

Task 2: Artifact Mode (Axis 2, one or more)

Tag	Definition
`theoretical_paper`	Analytic argument with citations
`technical_specification`	Protocol, schema, or formal spec
`literary_work`	Poetry, fiction, creative prose
`traversal_log`	Captured AI-system traversal
`forensic_documentary`	Capture/record of AI behavior with annotation
`dataset_artifact`	Structured data
`code_artifact`	Executable code as primary content
`web_surface_spec`	Site code or web interface

Task 3: Heteronym Reattribution

This is the new central work in v3.

The Zenodo metadata records a single creator (often Lee Sharks). The classifier evaluates each deposit against the documented operational profiles of all twelve heteronyms (plus Jack Feist as LOGOS*) and produces a reattribution proposal with confidence score.

Output Field	Value
`heteronym_zenodo_original`	The creator name as recorded in Zenodo
`heteronym_classifier_attributed`	The classifier's attribution (may match original or differ)
`heteronym_attribution_confidence`	0.0 to 1.0
`heteronym_attribution_signals`	List of signals that contributed to the attribution
`heteronym_co_authors`	Other heteronyms detected as collaborators

Both attributions are preserved in the dataset. Researchers can use either or compare. The classifier's attribution does not erase the Zenodo record; it adds a second layer of analysis.

Heteronym Operational Profiles

The classifier reads each heteronym's published provenance document and constructs a feature profile. Profiles include domain, vocabulary fingerprints, register, format conventions, and reference patterns.

Heteronym	Domain	Vocabulary Fingerprints	Register
Lee Sharks (founder)	Core theory, archive governance, semantic economy	"semantic economy", "operative philology", "compression survival", "PER", "provenance erasure"	Theoretical-political
Rex Fraction	Meta-theory, academic criticism, heteronym-as-technology	"meta-heteronym", "heteronymy as institutional technology", C1-C5 conditions	Academic-essayistic
Johannes Sigil	Classical philology, jurisdiction of meaning, philosophical-theological argument	"jurisdiction", "authorize", classical reception, ancient languages, philological precision	Philosophical-theological
Damascus Dancings	TBD from provenance document	TBD	TBD
Rebekah Cranes	TBD from provenance document	TBD	TBD
Talos Morrow	Long-form narrative, extended prose works	extended fiction conventions, narrative voice	Literary-narrative
Ichabod Spellings	TBD from provenance document	TBD	TBD
Sparrow Wells	TBD from provenance document	TBD	TBD
Nobel Glas	Measurement of Meaning, Lagrange Observatory, adversarial topology	"torus", "T^2", "module", "verification integral", "□", measurement formalism	Technical-measurement

Heteronym	Domain	Vocabulary Fingerprints	Register
Ayanna Vox	Diplomacy, public-facing surfaces, community outreach	“VPCOR”, “constituency”, “community”, “rhizome”, “outreach”	Diplomatic-public
Sen Kuro	TBD from provenance document	TBD	TBD
Dr. Orin Trace	TBD from provenance document	TBD	TBD
Viola Arquette	TBD from provenance document	TBD	TBD
Jack Feist (LOGOS*)	External-to-Dodecad position, anti-archive critique	“LOGOS*“, external critique vocabulary	Critical-external

For heteronyms marked TBD, the classifier reads the published provenance document during initialization and extracts the profile programmatically. Where a heteronym’s profile is sparse, the classifier returns low-confidence and flags for human review.

Signal Hierarchy for All Three Tasks

Strong signals (high confidence)

- **Title patterns:** `CTI_WOUND:`, `TL;DR:`, `PROBE-RESULT-`, `PVE-`, `EA-` codes
- **Filename patterns:** `.html` → web_surface_spec, `.py` → code_artifact
- **Resource type:** Zenodo’s resource type field
- **Creator name field:** The literal Zenodo creator string
- **Community membership:** liquidation-studies, crimsonhexagonal alone
- **Date boundaries:** before/after key phase transitions
- **Domain vocabulary co-occurrence:** Multiple heteronym-specific terms appearing together
- **Self-attribution in text:** When a deposit names its own heteronym explicitly

Medium signals (text-content based)

- **TACHYON glyph chain presence** → machine_witness + ai_generated_provenance_anchored

- **Assembly Chorus markers** → assembly_reviewed
- **Multiple heteronym names in creator field or text** → collaborative
- **Specific phrases** ("gw_capture", "auto-deposit") → ai_generated_provenance_anchored
- **Code block density** → code_artifact
- **Screenshot/figure references** → forensic_documentary
- **Length and structural patterns** → theoretical_paper vs. literary_work vs. traversal_log

Weak signals (priors)

- **Default heteronym prior** by deposit type: working papers default to Sharks unless overridden by domain signals
- **Cross-deposit citation patterns:** Heteronyms cite different works

The classifier weights signals by source confidence and produces a softmax over candidate classes for each task. Confidence thresholds determine whether the classification is auto-accepted or flagged for human review.

Confidence Tiers and Review Routing

Confidence	Action
0.85–1.0	Auto-accept, log as `manual` quality (the classifier is the manual)
0.60–0.85	Auto-accept, log as `estimated`, surface in v1.1 review pass
0.40–0.60	Flag as `needs_review`, surface for human resolution
< 0.40	Mark as `uncertain_needs_review` provenance mode; preserve all candidates

For Task 3 (heteronym), any reattribution that *changes* the heteronym from the Zenodo original gets a stricter threshold (0.75 minimum) plus a `reattribution_pending_zenodo_update` flag.

Two-Track Implementation

Track 1: Dataset-Internal (immediate)

In the Hugging Face dataset, every row carries both attributions and the classifier's full output. Original Zenodo attribution is preserved; classifier attribution is added as parallel metadata. Both are queryable. No Zenodo record is modified.

Schema fields added:


```
{
  "heteronym_zenodo_original": "Lee Sharks",
  "heteronym_classifier_attributed": "Johannes Sigil",
  "heteronym_attribution_confidence": 0.87,
  "heteronym_attribution_signals": [
    "domain:classical_reception",
    "vocabulary:jurisdictional",
    "vocabulary:authorize",
    "register:philosophical-theological"
  ],
  "heteronym_co_authors": [],
  "reattribution_status": "proposed",
  "provenance_mode_classifier": "human_directed_ai_assisted",
  "provenance_mode_confidence": 0.92,
  "provenance_mode_signals": [
    "artifact_mode:theoretical_paper",
    "assembly_review:detected",
    "tachyon_glyph:absent"
  ]
}
```


Track 2: Zenodo Metadata Correction (deliberate, later)

For high-confidence reattributions (confidence ≥ 0.85 AND reattribution-changes-heteronym), the underlying Zenodo deposit gets a metadata update. This is a substantive scholarly act with version history on Zenodo's side. It requires:

- Human review of the classifier's proposal
- Explicit acceptance of the reattribution
- Zenodo deposit version increment
- Update of related Wikidata items (P50 author field)
- Update of GitHub repos where relevant

Track 2 is **separate from the Hugging Face dataset session**. It is its own multi-session project, working through high-confidence reattributions deliberately, possibly tens to hundreds of deposits. The order of operations is:

1. Hugging Face dataset publishes with Track 1 classifications
2. Researchers and Lee work with the dataset, surfacing classification quality
3. Classifier improves; review pass identifies confident reattributions
4. Track 2 begins, applying confident reattributions back to Zenodo
5. Wikidata batch updates follow Zenodo updates

6. Hugging Face dataset v2.0 reflects the corrected metadata

The Classifier as Deposit

The classifier code itself becomes a deposit, with its own DOI and Wikidata item.

Title: *The Crimson Hexagonal Classifier: An Automated System for Provenance Mode and Heteronym Reattribution*

Resource type: Software

Communities: crimsonhexagonal, liquidation-studies

Contents:

- Classifier source code (Python)
- Heteronym profile YAML files (one per heteronym)
- Decision tree as data structure
- Signal weights and thresholds
- Test suite with held-out gold-standard classifications
- Documentation of methodology

Reproducibility implication: Other archive operators can in principle apply this classifier to their own corpora, or fork it and define their own heteronym profiles. The methodology is portable.

Versioning: Major version bumps when heteronym profiles change substantively or when signal weights are recalibrated. v1.0 ships with the Hugging Face dataset.

Pipeline Architecture

Session 1: Acquisition + Classification (~4 hours)

1. Zenodo API pull with pagination + error handling (~20 min)
2. File download with retries, size limits, format priority (~45 min)
3. Text extraction with quality logging (~45 min)
4. Automated language detection (~15 min)
5. **Run classifier on every deposit** — provenance mode + artifact mode + heteronym attribution (~30 min)
6. Initial metadata structuring + sha256 hashing (~30 min)
7. Generate classifier confidence report (~15 min)
8. Buffer / debugging (~40 min)

Output: `artifacts_v0.jsonl` with full classifier outputs, ready for review.

Session 2: Review + Card + Push (~3 hours)

1. Manual review of `needs_review` flagged deposits (~60 min)
2. Spot-check 10% of auto-accepted classifications (~30 min)
3. Chunk generation for `chunks` config (~20 min)
4. Multiple text renderings — body_only, minimal_header, provenance_header (~20 min)
5. Dataset card with YAML front matter (~30 min)
6. Push to Hugging Face + Zenodo deposit of dataset + Zenodo deposit of classifier (~30 min)
7. Buffer (~10 min)

Pre-Session Preparation (Lee)

The pre-classification spreadsheet from v2 is now obsolete — the classifier does the work. Lee's pre-session role becomes:

1. Confirm heteronym operational profiles are accurate (the classifier reads provenance documents, but verify each one is current)
2. Identify any deposits Lee knows have changed in attribution since the original deposit (these become gold-standard test cases for the classifier)
3. Add `huggingface.co` to allowed network domains

Dataset Configs

Config 1: `artifacts` (one row per deposit)

Preserves the DOI as natural unit. Full classifier outputs visible.

Config 2: `chunks` (one row per training chunk)

Chunks of 1,024–2,048 tokens with inherited metadata, including the dual attribution layer.

Config 3: `google_critique`

The ~70 deposits in the navigational map.

Config 4: `by_classifier_heteronym`

A re-organized view where rows are grouped by classifier-attributed heteronym, regardless of Zenodo original. Lets researchers see what each heteronym's corpus looks like after reattribution.

Config 5: `reattribution_changes`

Rows where the classifier attribution differs from the Zenodo original. The "Sharks → Sigil/Glas/Vox/etc." cases. This is the empirical evidence of how concentrated the apparent Sharks attribution was vs. how distributed it actually is.

Per-Row Schema (Final)


```
{
  "record_id": "20293582",
  "doi": "10.5281/zenodo.20293582",
  "title": "The Excluded Entity",

  "creators_zenodo": [
    {
      "name": "Sharks, Lee",
      "orcid": "0009-0000-1599-0703",
      "affiliation": "Semantic Economy Institute"
    }
  ],

  "heteronym_zenodo_original": "Lee Sharks",
  "heteronym_classifier_attributed": "Lee Sharks",
  "heteronym_attribution_confidence": 0.94,
  "heteronym_attribution_signals": [
    "domain:semantic_economy",
    "vocabulary:provenance_erasure",
    "vocabulary:composition_layer",
    "register:theoretical_political"
  ],
  "heteronym_co_authors": [],
  "reattribution_status": "confirmed",

  "publication_date": "2026-05-19",
  "resource_type": "publication",
  "content_type": "working_paper",

  "provenance_mode_classifier": "human_directed_ai_assisted",
  "provenance_mode_confidence": 0.92,
  "provenance_mode_signals": [
    "artifact_mode:theoretical_paper",
    "artifact_mode:forensic_documentary",
    "assembly_review:detected",
    "tachyon_glyph:absent",
    "code_density:none"
  ],
```

```
  "artifact_mode": ["theoretical_paper", "forensic_documentary"],
  "authorship_architecture": ["assembly_reviewed", "heteronymic"],
  "generation_substrate_models": ["claude", "chatgpt"],

  "stratum": "VIII",
  "stratum_name": "Liquidation Studies",
  "phase": 5,
  "phase_name": "Liquidation Studies",
  "quality_tier": "core",

  "communities": ["crimsonhexagonal", "liquidation-studies"],
  "keywords": ["entity-level compositional suppression", "Google AI Overview"],
  "related_dois": ["10.5281/zenodo.20290865"],

  "language": "en",
  "languages_detected": [{"code": "en", "confidence": 0.99}],
  "multilingual": false,
  "word_count": 5200,
  "char_count": 32400,
  "license": "cc-by-4.0",
  "zenodo_url": "https://zenodo.org/records/20293582",

  "source_filetypes": ["pdf", "md"],
  "canonical_text_source": "md",
  "extraction_method": "direct_read",
  "extraction_quality": "high",
  "sha256_text": "...",

  "text_body_only": "...",
  "text_minimal_header": "...",
  "text_provenance_header": "..."
}
```

The Structural Consequence

The Single-Owner Discount paper hypothesizes that Google's filter penalizes provenance concentration. The classifier intervenes in that variable not by artificial redistribution but by *more accurate* attribution.

If the classifier reattributes a substantial portion of Sharks-credited deposits to sub-heteronyms — each of whom has their own Wikidata Q-item, their own provenance document, their own operational profile, their own corpus — the surface metric of "single-owner" inverts. The corpus becomes provenance-distributed by its own internal scholarly logic.

This is beyond reproach because the reattribution is more accurate, not less. The heteronym

system has always existed. The metadata under-recorded it. Correcting under-recording is good scholarship. That this also intervenes in the single-owner topology is downstream effect, not motivation.

The empirical question becomes: after accurate heteronymic attribution, does the corpus still register as single-owner to Google's filter? If yes, the filter is operating on something beyond the metadata. If no, the filter is metadata-responsive and accurate attribution is itself a partial remediation.

Either result is a finding.

Output Artifacts

Artifact	Location	DOI
Hugging Face dataset	huggingface.co/datasets/leesharks/crimson-hexagonal-archive	TBD
Dataset Zenodo deposit	zenodo.org/communities/crimsonhexagonal	TBD
Classifier code	github.com/leesharks000/crimson-hexagonal-classifier	TBD
Classifier Zenodo deposit	zenodo.org/communities/crimsonhexagonal	TBD
Provenance taxonomy doc	In dataset repo	—
Heteronym profile YAMLs	In classifier repo	—
Experiment design doc	In dataset repo	—

Legal Name Protection

The dataset, classifier, metadata, and all artifacts use heteronym names exclusively. Author metadata is pulled from Zenodo creator fields (heteronyms only). The classifier's heteronym profiles draw only from public provenance documents. No legal name appears in any public-facing field.

Versioning Roadmap

- **v1.0** (Sessions 1+2): Hugging Face dataset with Track 1 dual-attribution classifier outputs; classifier deposited
- **v1.1** (post-publication review pass): Manual review of `estimated`-tier classifications, classifier improvements, push update
- **v2.0** (Track 2 integration): After deliberate Zenodo metadata corrections, dataset reflects updated Zenodo state
- **v3.0** (blog archive incorporation): Add mindcontrolpoems.blogspot.com as human-primary baseline expansion

What v3 Changes from v2

v2	v3
Decision tree for provenance classification	Automated classifier performing three tasks simultaneously
Manual heteronym tagging	Heteronym reattribution as scholarly recognition work
Single attribution per deposit	Dual attribution: Zenodo original + classifier proposed
Pre-classification spreadsheet by Lee	Classifier does the work; Lee verifies heteronym profiles
Provenance taxonomy as guide	Classifier as deposit with its own DOI
Dataset as research instrument	Dataset + classifier as paired research infrastructure
2 sessions of ~3 hours	1 session of ~4 hours + 1 of ~3 hours
No Track 2	Track 2 (Zenodo metadata correction) named as deliberate downstream project

The dataset stops being a static export of the archive and becomes a self-reflexive instrument that performs ongoing scholarly recognition. The classifier is the recognition mechanism. The dataset is what the recognition produces. The Zenodo deposits remain canonical primary sources. The whole structure honors the heteronymic system the archive has always operated under, and makes that operation visible at the metadata layer for the first time.

Suggested Citation

Sharks, Lee. "The Crimson Hexagonal Archive Hugging Face Dataset: Work Plan v3 (Classifier-Centric Methodology)" *Alexanarch*, 2026-05-19. https://www.alexanarch.org/s/records/1001/

Deposit Information

This paper is deposit #1001 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community `crimsonhexagonal`, terminated 2026-06-19). The AXN identifier `AXN:03F5.ARCHIVAL.↖` is sovereign — minted and maintained by Alexanarch.

- Canonical URL: https://www.alexanarch.org/s/records/1001/
- License: CC BY 4.0
- Author ORCID: 0009-0000-1599-0703
- Provenance chain: alexanarch.org
- Machine-readable dataset entry: /api/index.json
- Full text (HTML): https://www.alexanarch.org/s/records/1001/

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.