

CORAL: Lossless Reference Storage for Related Audio

Wubble AI Labs
August 20, 2026

Abstract

CORAL is a reference design for storing related decoded-PCM audio representations. A fingerprint retrieves possible compression references; a full-overlap verifier and a byte-cost test decide whether a transform plus a lossless residual is cheaper than an independent lossless payload. The residual construction gives exact decoded-PCM reconstruction in the reference environment by construction. On a controlled corpus of 20 synthetic masters and six derived forms per master (140 assets), the complete $\binom{140}{2} = 9,730$ -pair audit reports AUC = 1.000 for the controlled labels, zero accepted cross-cluster links, and zero maximum sample error. A separate run on an isolated copy of a private corpus scanned 3,076 objects and fully rehydrated all 2,877 inputs in its declared WAV contract: 2,781 PCM16 mono, 8 PCM16 stereo, and 88 float32 stereo. PCM reconstruction had zero maximum error; float32 samples were bit-exact. Its 4.445 GB accounted archive used 35.3% less storage than a 6.866 GB independent-lossless baseline (a 1.545 $\times$ baseline/accounted ratio). These results establish a bounded storage result, not universal format support, semantic identity, rights lineage, false-link rates on an unlabelled corpus, service latency, or industry-wide savings.

1 Introduction

Audio collections often include exports, loudness-adjusted versions, resamples, edits, and delivery files. Per-file lossless compression cannot exploit similarity between such files, while byte-level deduplication only removes identical byte strings. CORAL explores whether a collection can retain every decoded waveform exactly while storing some representations relative to others.

CORAL is *reference storage*, not a new codec or identity system. It stores a supported decoded-PCM asset independently or, for 16-bit PCM, as a deterministic prediction from a same-layout canonical plus a lossless integer residual. A fingerprint only retrieves candidates; it is neither an assertion that two works are the same nor a rights, provenance, or deletion decision. A full-overlap verifier and measured byte cost make the final decision.

If a deterministic predictor $f(c, m)$ approximates an

asset a from canonical c and manifest m , storing

$$\rho = a - f(c, m) \tag{1}$$

losslessly permits exact recovery of a . Storage improves only when the conditional representation is shorter than an independent one; perceptual resemblance alone does not imply this. This paper therefore offers a bounded, reproducible experiment, not an industry-wide storage claim.

2 Prior work and position

Robust audio fingerprints support matching across encoding and duration changes, including duplicate detection [1, 2]. Generic storage systems use deduplication and delta compression to exploit resemblance [3]. MPEG-4 SLS supplies a lossless enhancement layer over a lossy representation [4]. CORAL combines familiar ingredients in a narrower setting: a candidate retriever, a conservative waveform verifier, and a residual whose cost is compared with independently storing decoded PCM.

This paper does not claim a novel fingerprint, universal collision avoidance, semantic identity, or a replacement for audio codecs, digital-asset management, rights systems, or established deduplication engines.

3 Reference design

3.1 Scope

The controlled experiment uses 16-bit monophonic decoded PCM. The later isolated-corpus runner accepts standard RIFF/WAVE 16-bit PCM in mono or stereo for the delta path, and IEEE float32 stereo WAVE for canonical-only WavPack storage. It never creates a reference across channel layouts; stereo channels are retained separately, not folded to mono. It does not preserve an original container, codec bitstream, tags, artwork, or file-level metadata. A decoded MP3 variant in the controlled experiment is therefore PCM obtained after MP3 decoding, not preservation of its MP3 bytes. Other containers, compressed formats, multi-channel layouts, and malformed headers are outside this report.

3.2 Candidate retrieval and verification

Audio is resampled to 5 kHz, framed at 2048 samples with a 64-sample hop, mapped to 33 log-energy bands from 300–2000 Hz, and converted to a 32-bit sub-fingerprint per frame. The verifier searches temporal shifts, compares the complete aligned overlap (at least 80% of the shorter fingerprint), and reports bit-error rate (BER) and support: the fraction of 80-frame windows below the local BER threshold. A compression reference is considered only when

$$\text{BER} < 0.22 \quad \text{and} \quad \text{support} \geq 0.90. \quad (2)$$

This is a controlled-corpus operating point, not a universal safety threshold. It is separate from semantic identity, ownership, and policy decisions.

3.3 Residual and cost

For an accepted candidate c , the implementation resamples c to the target rate, estimates an integer offset by correlation, fits a scalar gain, rounds the prediction to integer PCM, and stores manifest m and a 32-bit integer residual. Canonicals use 16-bit FLAC; residuals use WavPack because the tested FFmpeg FLAC path does not preserve arbitrary 32-bit residual low-order bits. It links a reference only if

$$L(\rho) + |\text{JSON}(m)| < L_{\text{FLAC}}(a), \quad (3)$$

where L is the byte count produced by the reference encoders. An asset failing either gate is stored as independent FLAC.

Proposition 1 (Exact PCM reconstruction) *For a fixed deterministic implementation of transform and decoder, if ρ is stored losslessly then $f(c, m) + \rho = a$ sample for sample.*

This is an algebraic property of the representation, not a claim that all toolchains or file containers reproduce the same bytes. The reproducibility checks materialize and decode every canonical and residual payload before checking it on every stored asset in the reference environment.

Proposition 2 (Payload non-expansion) *Under Equation 3, every linked asset has lossless payload strictly smaller than its independently encoded FLAC cost and every unlinked asset costs its independent FLAC cost. Thus lossless payload is no larger than per-asset FLAC under the same encoder.*

This proposition excludes candidate-index metadata, proxies, replication, object-store requests, and operations. Our reported accounted store includes raw fingerprint/manifest metadata and proxies, so it can be larger than per-file FLAC at zero redundancy.

Table 1: Footprint at 140 assets. Payload is covered by the payload proposition; accounted CORAL adds metadata and proxies.

Method	Size (MB)	vs. raw
raw WAV	37.60	1.00×
tar + gzip	35.57	1.06×
per-file FLAC	25.26	1.49×
byte-equal reference + FLAC	21.64	1.74×
CORAL lossless payload	17.55	2.14×
CORAL accounted	18.51	2.03×

3.4 Layering

Each canonical has a 24-kbps proxy for preview-oriented access. The proxy is not required for exact reconstruction, but its bytes are included in the accounted store. Retention tiers, cache policy, and access latency are not measured here.

4 Controlled experiment

We generate 20 distinct music-like synthetic masters, each three seconds of 44.1-kHz, 16-bit mono PCM. Each master has six derived assets: exact copy, gain change, 48-kHz resample, padded trim/offset, local attenuation edit, and 128-kbps MP3 encode/decode. The result is 140 assets with known parent clusters. Synthetic construction permits an exact audit of this intended redundancy relation; it does not model production catalogues.

Baselines are raw WAV, a tar+gzip archive, per-file FLAC, and byte-equal reference storage with one FLAC per unique decoded PCM payload. CORAL is reported both as lossless payload and as a fully accounted store.

The evaluation reconstructs every asset, and evaluates every unordered fingerprint pair: 420 within-cluster and 9,310 between-cluster pairs, 9,730 total. AUC uses their BER values. It also counts any accepted cross-cluster compression-reference candidate under Equation 2. As a focused regression test, two ten-second signals share a six-second introduction but have independent four-second tails. Their global BER is below 0.22, but window support is 0.60, so the verifier rejects a reference link.

5 Controlled results

Exactness. All 140 decoded PCM assets reconstruct with maximum absolute sample error zero. This instantiates the first proposition for the reference implementation and corpus.

Retrieval and verification. The complete-pair audit yields $\text{AUC} = 1.000$ on the controlled labels; maximum within-cluster BER is 0.377 and minimum between-cluster BER is 0.439 (Fig. 1). At the stated

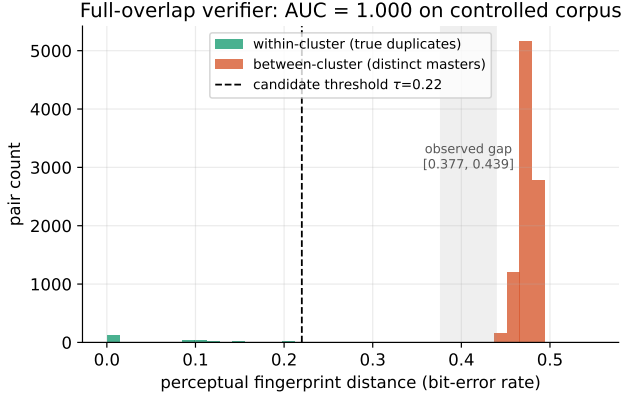


Figure 1: Complete pairwise audit on the controlled corpus. The shaded interval is an observed gap, not a universal safety margin.

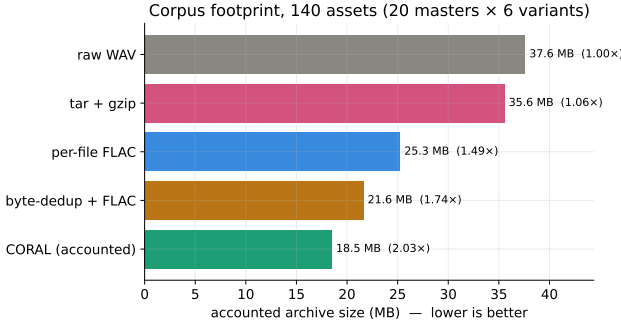


Figure 2: Accounted archive size, including CORAL metadata and proxies.

acceptance rule, no cross-cluster candidate is accepted. This is controlled evidence about this corpus, not a measured false-link rate on real music, speech, sound effects, or adversarial inputs.

Storage. Accounted CORAL is 18.51 MB: 2.03 $\times$ smaller than raw WAV and 1.36 $\times$ smaller than per-file FLAC (Table 1, Fig. 2). Lossless payload is 17.55 MB; metadata contributes 0.38 MB and proxies 0.59 MB (Fig. 4). The experiment creates 56 independent canonicals and 84 cost-saving reference links, recovering 70.0% of the 120 derived assets as links. The remaining derivatives are stored independently because the verifier or cost decision declines them.

Scaling. The fully accounted store is larger than per-file FLAC at zero derived variants because fixed metadata/proxy costs dominate. Savings appear as controlled redundancy is introduced and reach 1.36 $\times$ versus FLAC at six derived forms per master (Fig. 3). This is a sensitivity result, not evidence that real archives have a particular redundancy level.

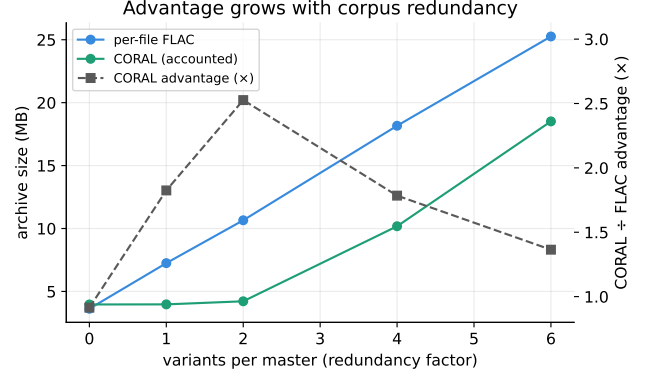


Figure 3: Controlled redundancy sweep. Fully accounted CORAL need not match FLAC at zero redundancy because it includes fixed metadata and proxies.

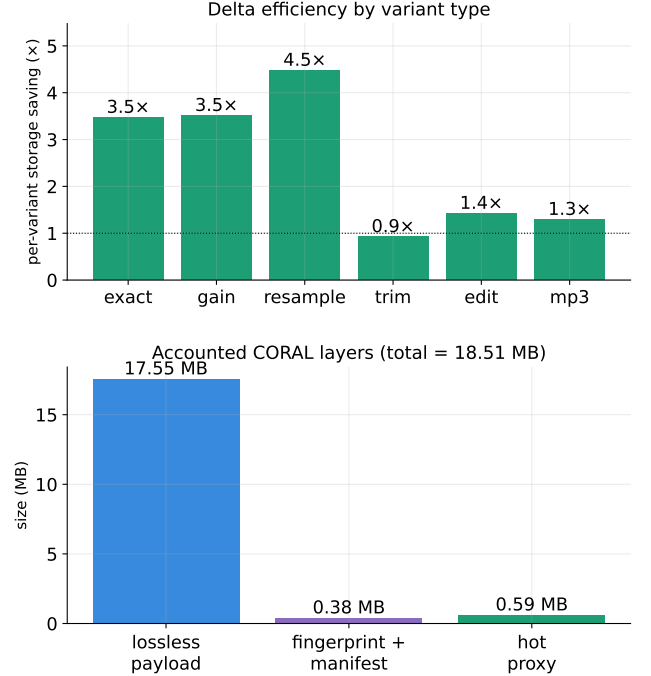


Figure 4: Accepted-link efficiency and byte layers in the accounted total. Type ratios describe accepted links only.

Table 2: Isolated copied-corpus run. GB is decimal. The baseline uses FLAC for PCM16 and WavPack for float32, independently per accepted asset.

Measure	Result
Objects header-scanned	3,076
PCM16 mono accepted	2,781
PCM16 stereo accepted	8
Float32 stereo accepted	88
Excluded outside contract	199
Accepted and rehydrated	2,877
Independent lossless baseline	6.866 GB
CORAL lossless payload	4.354 GB
CORAL accounted archive	4.445 GB
Storage reduction vs. baseline	35.3%
Baseline / accounted ratio	1.545×
Independent canonicals / links	288 / 2,589
Maximum PCM reconstruction error	0
Float32 stereo reconstruction	bit-exact

6 Isolated copied-corpus evaluation

The controlled corpus establishes behaviour against known synthetic labels, but does not measure real-corpus storage. We therefore ran the implementation against an isolated research copy of a private audio corpus. The original production source bucket was neither accessed nor modified: the runner used a separate, no-delete role that could read only the copied `corpus/` prefix and write only a fresh research-results prefix. This is an implementation evaluation, not a production deployment or cloud-cost study.

The run first parsed every object header. It accepted standard 16-bit PCM mono and stereo WAVE for delta archival and float32 stereo WAVE for canonical-only WavPack archival. It excluded 195 nonstandard RIFF/WAVE objects and four WAVE objects outside the PCM contract without converting or altering them. Each of the 2,877 accepted inputs was archived, read back, and compared with its source decoded samples. The 88 float32 inputs were compared by their raw float bits; the other accepted inputs were compared sample for sample.

The 2,589 links followed 2,589 full-overlap verifications from 2,594 retrieved candidate pairs; five retrieved candidates failed the final gate. The bounded-memory retrieval index can miss a possible link, which can only reduce measured savings, not relax the verifier. Materialized archive bytes equalled the accounting total. The full-run result therefore supports exact reconstruction and the stated storage comparison for this copied corpus and input contract. It does *not* estimate duplicate recall, false-link rate, or AUC because this private corpus has no independently supplied duplicate or non-duplicate ground-truth labels.

Isolated copied-corpus result: exact reconstruction within the declared WAVE contract

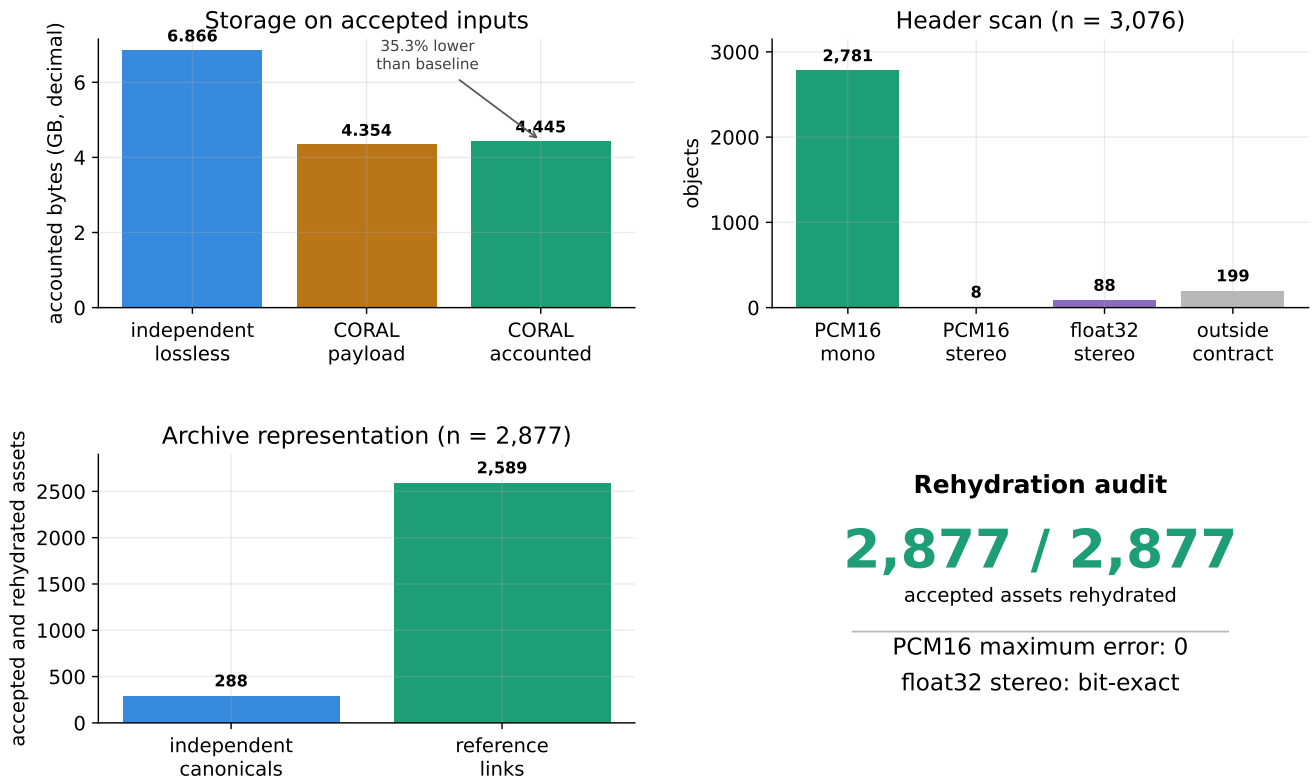


Figure 5: Isolated copied-corpus outcome. All 2,877 accepted inputs were materialized and rehydrated exactly in the declared WAVE contract; PCM16 had zero maximum sample error and float32 stereo was bit-exact. The 199 excluded objects are outside the input contract and are not part of the storage result.

Figure 5 makes the real-corpus accounting visible alongside the input-contract boundary. The 199 excluded objects are not converted, archived, or counted in the storage comparison.

7 Practicality and limits

The representation is mechanically practical at the evaluated research scale: standard decoding, channel-preserving fingerprints, deterministic transforms, 16-bit FLAC canonicals, WavPack residuals, float32 WavPack canonicals, and a codec round-trip check. The 140-asset experiment proves a controlled mechanism; the 2,877-asset run proves exact reconstruction and storage accounting on its isolated copied corpus. Neither is a deployable archival-service evaluation. The controlled implementation uses exhaustive in-memory scans; the full runner uses a bounded canonical cache and coarse retrieval index followed by the full verifier. Neither result measures production ingest throughput, index recall, concurrent writes, recovery, tail latency, cloud fees, or catalogue-scale operation.

A production evaluation needs a durable canonical/manifest store, an approximate-nearest-neighbour index followed by the full verifier, reference-counted garbage collection, audit records, encryption, access control, and transactional recovery. It needs a blinded holdout corpus spanning real music, speech, ambience, long-form and multichannel assets, independently similar works, and realistic edits. It should pre-register candidate recall, false-link rate with confidence intervals, storage costs including replication and requests, ingest/reconstruction latency, and human review of borderline links.

The current channel policy covers only mono and stereo decoded samples and forbids cross-layout references. The method has no warped alignment for tempo or pitch changes, portable transform-version specification, original-bitstream preservation, or rights semantics. Its controlled AUC and zero cross-cluster candidates cannot establish safety beyond the controlled generator. Likewise, the isolated-run $1.545\times$ comparison is not a market or industry estimate.

8 Reproducibility

Run the following from the implementation directory:

```
python -m experiments.test_guarantees
python -m experiments.run_benchmark
python -m experiments.make_figures
```

The checks verify exact reconstruction across the controlled corpus, channel-preserving stereo paths, float32 WavPack round trips, and rejection of

the shared-introduction case; the benchmark regenerates the controlled metrics and Figures 1–4, while the checked-in isolated aggregate records regenerate Figure 5. The isolated-corpus driver, `experiments.run_research_pilot`, additionally records a versioned manifest and aggregate metrics after a complete rehydration pass. It requires a separately authorised read-only research copy and is not a command to run against a production source bucket. Required Python packages and an `ffmpeg` binary are listed in the implementation README.

9 Conclusion

CORAL demonstrates a bounded result: a decoded-PCM reference representation can make exact reconstruction compatible with lower storage cost when a corpus contains cost-effective waveform redundancy. The controlled 140-asset audit supports the mechanism under known labels; the isolated 2,877-asset run supports losslessness and a 35.3% storage reduction against its independent-lossless baseline. These results are credible within their stated scope, not evidence that the approach is revolutionary, industry-ready, universally safe, or equivalent to original-file preservation. A blinded representative corpus with ground-truth relations, broader format coverage, operational measurements, and external replication is the next milestone.

References

- [1] J. Haitsma and T. Kalker. A Highly Robust Audio Fingerprinting System. *ISMIR*, 2002.
- [2] C. J. C. Burges, D. Plastina, J. Platt, E. Renshaw, and H. S. Malvar. Duplicate Detection and Audio Thumbnails with Audio Fingerprinting. Microsoft Research Technical Report MSR-TR-2004-19, 2004.
- [3] W. Xia, H. Jiang, D. Feng, and L. Tian. DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads. *IEEE Transactions on Computers*, 65(6):1692–1705, 2016.
- [4] R. Yu et al. MPEG-4 Scalable to Lossless Audio Coding (SLS). ISO/IEC 14496-3 Amendment, 2006.