

Data · pinned from sggs-data

Scripture Integrity

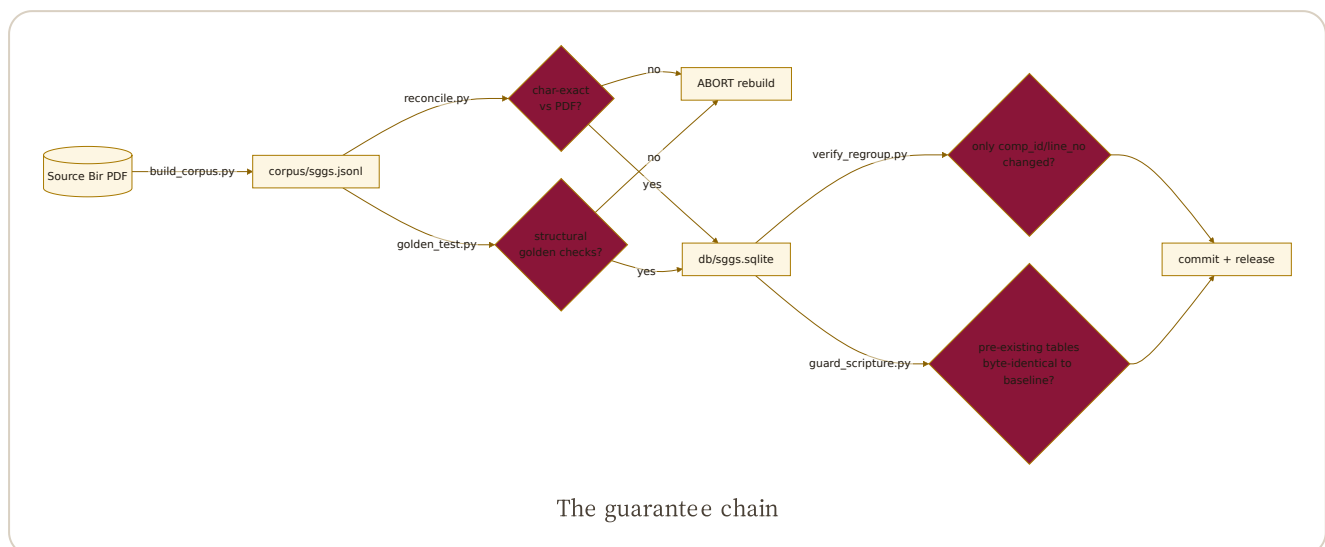
🕒 3 min read · Updated 27 October 2026

<https://docs.gurbanisoul.com/data/architecture/scripture-integrity/> · a pinned copy of Algorythmos-AI/sggs-data at c36e317

Pinned copy. The canonical file is [sggs-data/docs/architecture/scripture-integrity.md](https://docs.gurbanisoul.com/data/architecture/scripture-integrity.md)¹ at commit c36e317 — edit it there. This wiki publishes it by pin ([ADR-0012](#)); a newer version arrives as a reviewed bump of `docs-site/sources.lock.json`.

Textual fidelity is the single most important property of this project and overrides everything else. The Gurmukhi text is **verbatim** from the source Bir and is **never** edited, paraphrased, normalized, reordered, or guessed at.

The guarantee chain



The only sanctioned text transforms

1. PDF **visual** → **logical** Unicode reordering of the sihari (

ਿਕ੍ਰਪੁਰ

→

ਕ੍ਰਿਪੁਰ

).

2. Editorial Unicode-repair corrections where the source font emitted impossible sequences. The complete register is `audit/editorial-ledger.jsonl` — **4 rules, 11 applications**: Angs 573, 586, 727 (itemised from the start) and Angs 1354, 1358, 1387, 1398, 1402, 1406, 1408, 1409 (same `fix_text` step-4 rules; itemised 2026-09-

23, scholarly review completed and approved 2026-09-24, gate G3). Every decoding transform (timestamps, halant order, glyph restorations, vowel re-attachment, stray-character strip) is registered too.

Any new transform must be registered in the ledger and reviewed by a human. The English translations are a **separate, labelled layer** (Dr. Sant Singh Khalsa via BaniDB/ShabadOS) and are never blended into the Gurmukhi.

Proof tools

Tool	Proves	When
pipeline/reconcile.py	corpus == PDF, character for character	every rebuild (gate)
pipeline/golden_test.py	canonical structural checks	every rebuild (gate)
pipeline/siggs_integrity.py	the committed DB's content equals audit/dataset-fingerprint.json — every table (build stamps blanked), every FTS5 index (via fts5vocab), scripture_sha256 (the iOS definition) and t0_sha256 ; stable across SQLite versions	every PR (CI gate); written by every rebuild
pipeline/data_quality.py	the dataset is well formed and the tracked rows are as reviewed: 60,658 lines over Angs 1–1430, no line without a comp_id and the set of comp_id gaps unchanged (ids are never reused), the heading set unchanged, no Latin in Gurmukhi columns and no Gurmukhi in the English layer, no reference to a missing line, and the 21 isolated-sign lines plus the empty- translit_norm line exactly as in audit/data-quality-baseline.json	every PR (CI gate); the baseline is re-recorded only in a reviewed PR
pipeline/ledger_check.py	fix_text editorial rules == the ledger; any scripture (T0) change is covered by a new, reviewed ledger entry	every PR (CI gate)
pipeline/verify_regroup.py	a corpus change touched only comp_id/line_no ; scripture byte-identical	any PR touching corpus/db
pipeline/timing/guard_scripture.py	pre-existing tables byte-identical to the committed baseline	after any DB write
validation/reconcile-attestation.json	the corpus in this commit reconciled char-exact locally (the PDF never enters CI)	make reconcile
/api/health	60,658 lines, 1,430 Angs, FTS live, verbatim Mool Mantar, ≥560 ੴ , live verify	after any DB change

The hash chain

MANIFEST.json carries corpus_sha256 + db_sha256; contract/_meta.json carries its own db_sha256; the iOS manifest carries db_sha256 + scripture_sha256. scripture-integrity CI asserts these agree, so a DB rebuild that forgets to regenerate the contract

or the attestation fails the gate rather than silently shipping drift.

If text looks wrong

Flag it, do not change it. Open a *Scripture fidelity concern* issue with the Ang and the verbatim line; a Granthi/scholar reviews. Known items awaiting review: the 3 TRAILING_RUBRICS (

ਜੁਮਲਾ

,

ਦੁਤਕੇ

,

ਏਹੁ ਸਲੋਕੁ ਆਦਿ ਅੰਤਿ ਪੜਣਾ

), ~20 verse lines the source mis-tags as headings, and the source-faithful isolated-matra rows (Angs ~695–699, 1354/1358/1387).

Links

1. <https://github.com/Algorithmos-AI/siggs-data/blob/c36e3172bdf38a61070849bb604b38da89ca5d8b/docs/architecture/scripture-integrity.md>

Sri Guru Granth Sahib Ji — KnowledgeBase wiki. Generated from commit c3d9360. Updated 27 October 2026. This copy may be out of date; the page is at <https://docs.gurbanisoul.com/data/architecture/scripture-integrity/>