

Data

Data & pipeline

Updated 27 September 2026 · Verified against [fe1ff7d](#) on 2026-09-25

Where the scripture data comes from, how it is proven, how it reaches this platform by pin, and which pages explain each part — with the canonical sggs-data documents published beside them.

<https://docs.gurbanisoul.com/data/>

Everything the API serves comes from one artifact: a SQLite database of **60,658 line records over Angs 1–1430**, built by a reproducible pipeline from the source Bir PDF and proven, character for character, to contain exactly the printed text. That artifact is built, gated and published in the data repository, [Algorythmos-AI/sggs-data](#)¹; this platform never edits it and never commits a copy — it **pins** a commit and a checksum (`dataset.lock.json`) and installs exactly that object.

Prime directive. The Gurmukhi is sacred and verbatim. Nothing in this platform, and nothing described on these pages, rewrites it. The two sanctioned kinds of transform — decoding the PDF font’s glyph order into logical Unicode, and a handful of scholar-reviewed repairs of impossible sequences the font emitted — are registered in the [editorial ledger](#) and enforced by a gate in sggs-data.

The pages

Anatomy of a line record

The 23 columns of the `lines` table, how a page becomes rows, what search reads (Poster 03 walkthrough · a live Ang explorer against the real API)

Corpus pipeline and gates

`rebuild_all.sh` stage by stage, and the gates that stop it (Poster 04 walkthrough · code read from sggs-data at the pinned commit)

Data dictionary

Every table in the pinned database: its layer, columns, row count and the API context that reads it (Generated from the database; CI fails if it drifts)

The dataset pin

How this platform obtains the database, what it verifies, how a new dataset arrives (The lock, the fetcher’s guarantees, a pin bump end to end)

The editorial ledger

What may change in the text, what never does, and how a fidelity concern travels (The rules and their applications, the review flow)

The English translation layer

Where the English comes from, how it is matched to lines, how it stays separate, English search, and its licence status (Two sources, one translator)

The integrity chain

One line of scripture from the source PDF to a screen, and the gate that proves it at each stage (Six stages, each read against the code that enforces it)

Pinned from sggs-data

The canonical data documents live in the data repository. The wiki publishes them here **by pin** (`docs-site/sources.lock.json` records the commit and the sha256 of every file), with a banner that says where the canonical file is; edit them there.

Document	Canonical file
Database schema and the <code>comp_id</code> model	docs/architecture/database-schema.md
Scripture integrity: the guarantee chain, the proof tools, the hash chain	docs/architecture/scripture-integrity.md
Runbook: rebuild the database from the PDF	docs/process/runbooks/rebuild-db.md
Runbook: restore the data from backups	docs/process/runbooks/data-restore.md
ADR-0002 · keep-last <code>comp_id</code> regroup	docs/adr/0002-comp-id-keep-last-regroup.md
ADR-0003 · integrity proofs	docs/adr/0003-integrity-proofs.md
ADR-0004 · backups and restore	docs/adr/0004-backups-and-restore.md
The Answer Protocol: how Gurbani questions are answered	Answer-Protocol.md

The numbers that never drift

Fact	Value	Proven by
Line records	60,658	<code>data_quality.py</code> (sggs-data), <code>/api/health</code>
Angs	1–1430, gap-free	<code>golden_test.py</code> , <code>/api/health</code>

Fact	Value	Proven by
Compositions (<code>comp_id</code>)	4,527 distinct; <code>max(comp_id)</code> 5,376; gaps permanent (read from the pinned database)	<code>verify_regroup.py --invariants</code>
Vaars	22, detected by title header	<code>build_vaars.py</code>
Editorial rules / applications	4 / 11, all reviewed	<code>ledger_check.py</code>
Dataset served	<code>dataset.lock.json</code> → commit, sha256, size	<code>make dataset-check</code> , the <code>scripture-integrity</code> workflow

Read next: [Anatomy of a line record.](#)

Links

1. <https://github.com/Algorythmos-AI/sggs-data>

Sri Guru Granth Sahib Ji — Knowledge Base wiki. Generated from commit c3d9360. Updated 27 September 2026. Verified against fe1ff7d on 2026-09-25.
This copy may be out of date; the page is at <https://docs.gurbanisoul.com/data/>