

Data

Data dictionary

🕒 8 min read · Updated 28 October 2026

Every table in the pinned database — its layer, what it holds, its columns, its row count and which API context reads it — generated from the database itself.

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

Every table in the pinned database: 40 tables, of which 4 are full-text indexes (their 19 internal shadow tables are not listed). The page is generated from the database by `tools/gen_data_dictionary.py`, and CI fails if it differs from the database the platform pins, so the counts and columns below are those of the dataset in the lock.

The pinned dataset	
Built by	Algorythmos-AI/sggs-data ¹ at <code>efbf74e4</code>
DATASET_VERSION (the lock)	1.3.0
Database build version (<code>meta.version</code>)	1.3.0, built 2026-10-26
SHA-256	4e6777c54140d7c4...
Size	103.6 MiB
Lines · Angs	60,658 · 1,430

Which version is which, and how a new dataset arrives, is on [the dataset pin](#). The database is opened read-only by the API; **Read by** names the bounded contexts that declare the table (TABLES in each module, enforced by `webapp/tests/test_declared_tables.py`).

Scripture and structure

The verbatim text and the structure derived from it. Everything else refers to these rows.

Table	Rows	Read by	What it holds
authors	36	insights, reader	Each contributor with the Ang span and number of lines attributed.
lines	60,658	insights, reader, search, verify	One row per printed line: the verbatim Gurmukhi with its Ang, raag, section, author and composition; the source of truth (anatomy of a line record).
meta	9	reader	Facts about the build as key/value pairs: version, build date, edition, totals.
raags	31	insights, knowledge, reader	Each raag with its Ang span, line and shabad counts, Roman name and order in the Granth.

Table	Rows	Read by	What it holds
<u>sections</u>	15	reader	Each section of the Granth with its Ang span and line count.
<u>word_freq</u>	29,312	search	Every distinct Gurmukhi word and how often it occurs.

authors

name TEXT · first_ang INT · last_ang INT · n_lines INT

lines

id INTEGER · ang INT · pdf_page INT · raag TEXT · section TEXT · author TEXT ·
 comp_type TEXT · ghar TEXT · comp_id INT · line_no INT · is_rahao INT · is_header
 INT · markers TEXT · gurmukhi TEXT · text TEXT · translit TEXT · translit_norm
 TEXT · fl_g TEXT · fl_r TEXT · skeleton TEXT · stanza_index INT · pada_total INT ·
 source_category TEXT

meta

key TEXT · value TEXT

raags

name TEXT · first_ang INT · last_ang INT · n_lines INT · n_shabads INT · roman TEXT ·
 seq INT

sections

name TEXT · first_ang INT · last_ang INT · n_lines INT

word_freq

word TEXT · n INT

Search indexes

Built from the scripture tables at rebuild time; the search and verify contexts read them.

Table	Rows	Read by	What it holds
<u>canon_tokens</u>	24,676	search	Every distinct token of the transliteration, used to tell a known Roman word from a misspelling.
<u>fts</u>	60,658	reader, search, verify	Full-text index over lines : the text, transliteration, Roman fold, first letters and consonant skeleton.
<u>fts_shabad</u>	4,528	search	Full-text index over each composition's Roman consonant skeleton, for queries that span lines.
<u>fts_tri</u>	55,508	—	Trigram index over the transliteration, for partial and misspelt Roman queries.

Table	Rows	Read by	What it holds
<u>variants</u>	79,840	search	Spelling and phonetic variants, each mapped to the Gurmukhi word it stands for, with frequency, score and type.

canon_tokens

token TEXT

fts

FTS5 index; tokenizer unicode61, with 18 Gurmukhi signs counted as part of a token.

text · translit · translit_norm · fl_g · fl_r · skeleton

fts_shabad

FTS5 index; tokenizer unicode61 (the default).

tnorm · comp_id

fts_tri

FTS5 index; tokenizer trigram.

translit · line_id

variants

variant TEXT · gurmukhi TEXT · translit TEXT · freq INT · score REAL · rtype TEXT

The English translation

A separate, labelled layer. It never replaces or mixes with the Gurmukhi.

Table	Rows	Read by	What it holds
<u>fts_en</u>	58,039	search	Full-text index over the English translation.
<u>sources</u>	2	—	Where each translation comes from: attribution, licence terms and the date ingested.
<u>translations</u>	58,039	insights, reader, search	The English translation, one row per translated line, with its source and match quality.

fts_en

FTS5 index; tokenizer unicode61 (the default).

text · line_id

sources

source_id TEXT · kind TEXT · attribution TEXT · license TEXT · ingest_date TEXT

translations

line_id INT · lang TEXT · source TEXT · text TEXT · match_quality TEXT

Banis (Nitnem)

The daily prayers as ordered references to lines, plus the few lines that come from outside the Granth.

Table	Rows	Read by	What it holds
<u>bani_lines</u>	7,910	reader	The ordered lines of each bani, each a reference to a Granth line or to an extra line.
<u>banis</u>	23	reader	Each bani and variant: titles, category, order, length and where it comes from.
<u>extra_lines</u>	1,191	reader	Lines of banis that are not in the Granth, with their source and review status.

bani_lines

bani_id INTEGER · seq INTEGER · line_group INTEGER · line_id INTEGER · extra_id INTEGER

banis

bani_id INTEGER · key TEXT · variant TEXT · is_default INTEGER · title_gm TEXT · title_en TEXT · category TEXT · order_no INTEGER · n_lines INTEGER · n_groups INTEGER · has_extra INTEGER · estimated_minutes INTEGER · description_en TEXT · source_label TEXT

extra_lines

extra_id INTEGER · source TEXT · panna INTEGER · gurmukhi TEXT · translit TEXT · is_header INTEGER · shabados_line_id TEXT · review_status TEXT

Insights (precomputed analytics)

Descriptive statistics over structure, the theme tags and the English. See [the insights context](#).

Table	Rows	Read by	What it holds
<u>author_analytics</u>	36	insights	Per-author counts and stylometry of the English translation, with a reliability flag.
<u>author_dialogues</u>	17	insights	A Guru's salok set directly after a Bhagat's salok in the Kabir and Farid collections, by line range.

Table	Rows	Read by	What it holds
<u>author_distinctive_terms</u>	195	insights	The Gurmukhi words most characteristic of each voice (log-odds; pen-names and function words excluded).
<u>author_facts</u>	36	insights	Per-voice facts from the Gurmukhi: one verse count, Angs present (bit mask), raags, raags no other voice uses, stylometry.
<u>author_mentions</u>	249	insights	Verses in which one voice names another: Bhagats named together, the Gurus praised by the Bhatts.
<u>author_resonance</u>	679	insights	How often one voice's lines land nearest another's, as lift.
<u>concept_lines</u>	62,221	insights, reader, search	Which lines carry which theme.
<u>concepts</u>	54	insights, reader, search	The curated themes.
<u>line_neighbors</u>	532,212	insights	For each line, the lines whose English is closest by TF-IDF cosine (0.30 and above).
<u>raag_analytics</u>	31	insights	Per-raag counts and theme emphasis.
<u>shabad_neighbors</u>	43,840	insights	For each composition, the ten with the most similar theme profile.
<u>theme_fingerprint</u>	2,316	insights	How strongly each author or raag carries each theme, as lift over the whole Granth.
<u>theme_network</u>	2,528	insights	Theme pairs that occur in the same shabads, with PPMI and Jaccard.
<u>vaar_units</u>	1,443	insights	Each Vaar's saloks and pauris in reading order.
<u>vaars</u>	22	insights	The Vaars: span, pauri and salok counts, and who wrote each part.

author_analytics

author TEXT · n_lines INTEGER · n_shabads INTEGER · n_raags INTEGER · n_tokens
 INTEGER · mattr_100 REAL · hapax_pct REAL · avg_words_line REAL ·
 avg_lines_shabad REAL · top_themes TEXT · is_reliable INTEGER

author_dialogues

seq INTEGER · bhagat TEXT · guru TEXT · ang INTEGER · bhagat_first_id INTEGER ·
 bhagat_last_id INTEGER · guru_first_id INTEGER · guru_last_id INTEGER

author_distinctive_terms

author TEXT · term TEXT · z_score REAL · count_author INTEGER · count_rest
 INTEGER · rank INTEGER

author_facts

author TEXT · n_verses INTEGER · n_compositions INTEGER · n_rahao INTEGER ·
n_angs INTEGER · first_ang INTEGER · last_ang INTEGER · first_line_id INTEGER ·
ang_mask TEXT · raags_json TEXT · signature_raags_json TEXT ·
outside_raag_verses INTEGER · chhap_count INTEGER · n_tokens INTEGER ·
mattr_100 REAL · words_per_verse REAL

author_mentions

line_id INTEGER · ang INTEGER · src_author TEXT · mentioned TEXT · form TEXT

author_resonance

src_author TEXT · dst_author TEXT · edges INTEGER · mean_score REAL · lift REAL

concept_lines

concept TEXT · line_id INT · term TEXT

concepts

concept TEXT · gurmukhi_terms TEXT · description TEXT

line_neighbors

line_id INTEGER · neighbor_id INTEGER · score REAL

raag_analytics

raag TEXT · n_lines INTEGER · n_shabads INTEGER · n_authors INTEGER ·
dominant_author TEXT · dominant_author_pct REAL · n_themes INTEGER · top_themes
TEXT

shabad_neighbors

comp_id INTEGER · neighbor_comp_id INTEGER · rank INTEGER · score REAL

theme_fingerprint

entity_type TEXT · entity_id TEXT · concept TEXT · n_tagged INTEGER ·
entity_total INTEGER · entity_rate REAL · corpus_rate REAL · lift REAL ·
is_reliable INTEGER

theme_network

source TEXT · target TEXT · shabad_count INTEGER · source_n INTEGER · target_n
INTEGER · ppmi REAL · jaccard REAL

vaar_units

vaar_id INT · seq INT · kind TEXT · comp_id INT · author TEXT · n_lines INT · pauri_no INT · first_line_id INT · ang INT · theme TEXT

vaars

vaar_id INTEGER · raag TEXT · roman TEXT · first_ang INT · last_ang INT · n_pauris INT · n_saloks INT · pauri_author TEXT · salok_authors TEXT · cross_author INT · title TEXT

Knowledge (timing and forms)

Attributed scholarship about the text. See [the knowledge context](#).

Table	Rows	Read by	What it holds
<u>raag_timing_claims</u>	46	knowledge	When each raag is sung according to a cited source: pahar, season or occasion, with confidence.
<u>shabd_musical_markers</u>	4,531	knowledge	Ghar, partaal, rahao and other musical markers stated in a composition's heading.
<u>shabd_poetic_genre</u>	57	knowledge	A composition's named genre when its heading states it.
<u>shabd_raag_map</u>	4,531	knowledge	Each composition's raag, first Ang and first line.
<u>shabd_structural_form</u>	4,531	knowledge	A composition's form when its heading states it; otherwise NULL, never guessed.
<u>timing_sources</u>	5	knowledge, reader	The sources behind the timing claims and their tradition.

raag_timing_claims

id INTEGER · raag_name TEXT · claim_type TEXT · pahar INTEGER · time_start TEXT · time_end TEXT · season TEXT · occasion TEXT · source_id INTEGER · confidence TEXT · notes TEXT

shabd_musical_markers

comp_id INTEGER · ghar INTEGER · partaal INTEGER · has_rahao INTEGER · has_rahao_dooja INTEGER · dhunni TEXT · jati TEXT · source_label TEXT

shabd_poetic_genre

comp_id INTEGER · genre TEXT · source_label TEXT

shabd_raag_map

comp_id INTEGER · raag_name TEXT · first_ang INTEGER · first_line_id INTEGER

shabd_structural_form

comp_id INTEGER · form TEXT · pada_count INTEGER · source_label TEXT

timing_sources

id INTEGER · name TEXT · tradition TEXT · url TEXT · notes TEXT

Build records

Bookkeeping written by the rebuild; no route serves it except as metadata.

Table	Rows	Read by	What it holds
<u>analytics_meta</u>	12	insights	How the analytics were built: methods, framing, build dates.
<u>timing_migrations</u>	1	—	The timing layer's applied migrations and whether the scripture baseline was verified.

analytics_meta

key TEXT · value TEXT

timing_migrations

id INTEGER · name TEXT · applied_at TEXT · baseline_verified INTEGER

Links

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