

Data · pinned from sggs-data

Attribution: who composed each line

🕒 3 min read · Updated 27 October 2026

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

Pinned copy. The canonical file is [sggs-data/docs/architecture/attribution.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`.

`lines.author` names the voice (Guru, Bhagat, Bhatt or Gursikh) of every line. It is metadata *about* the scripture and never changes a character of it. This page explains where the value comes from and how it is proven.

Why a dedicated pass

`build_corpus.py` reads a composition's author from its heading and carries it forward until the next heading that names someone. Several kinds of printed evidence were not read, so authorship “bled” from one composition into the next. A line-level audit (2026-10-23) against an independent edition found 1,047 lines affected, for example:

- Baba Sheikh Farid Ji's saloks (Angs 1377–1384) tagged as Guru Amar Das Ji's or Guru Arjan Dev Ji's, because a

ਮ: ੩ ॥

label inside his collection was never closed;

- Bhagats Dhanna, Sain, Pipa, Sadhna and Parmanand, and Baba Sundar Ji (Ramkali Sadu), absent from the author list because their headings use inflected names (

ਧੰਨੇ ਜੀ ਕੀ

,

ਸਧਨੇ ਕੀ

) or bare labels (

ਪੀਪਾ ॥

,

ਸ੍ਰੀ ਮੈਣ ॥

);

- Japji with no author at all (no Mahalla is printed over it).

The rules (`pipeline/attribution.py`, post-pass 4 of `build_corpus.py`)

Each rule is tied to something printed in the Granth:

Rule	Evidence
P	A numbered pauri label names its Mahalla: ਪਉੜੀ ੫ ॥ (Guru Arjan Dev Ji's pauris in Guru Ram Das Ji's Gauri Vaar).
L	A heading or bare label names the voice of the next shabad/salok: ਆਸਾ ਬਾਣੀ ਭਗਤ ਧੰਨੇ ਜੀ ਕੀ , ਧੰਨਾ ॥ , ਮ: ੫ ॥ . A label covers one shabad or salok.
C	A shabad under an unnamed heading (ਬਾਣੀ ਭਗਤਾਂ ਕੀ , ਸਾਰੰਗ ॥) is signed in its final stanza with one Bhagat's pen-name and never ਨਾਨਕ . Words that are also ordinary vocabulary (ਨਾਮੋ “in the Name”, ਸੈਣੁ “friend”, ਧੰਨਾ “blessed”, ਪਰਮਾਨੰਦ “bliss”) count only in a strict signing position (ਕਹਿ , ਬਦਤਿ , ਪ੍ਰਣਵੈ , “ ਭਣੈ).
N	An unsigned shabad takes the signer of its numbered series: the running count (॥ ੨॥੧॥ , ॥ ੨॥੨॥ , ॥ ੪॥੩॥) continues with no heading between.

Rule	Evidence
S	In a Bhagat's salok collection (ਸਲੋਕ ਸੇਖ ਫਰੀਦ ਕੇ) every salok is the collection's, except one introduced by a ਮ: N ॥ label, which ends at its own number.
H	Otherwise a section heading's named Bhagat repairs bleed from a Guru or a missing tag. It never moves a line from one Bhagat to another.
V	A bare ਸਲੋਕੁ ॥ in a Guru's Vaar that opens with ਕਬੀਰ / ਫਰੀਦਾ is that Bhagat's salok.
E	Approved editorial decisions in <code>pipeline/attribution_ledger.json</code>.

A heading that names a Mahalla always wins over a pen-name in the verse (Ang 326:

ਗਉੜੀ ਕਬੀਰ ਜੀ ਕੀ ਨਾਲਿ ਰਲਾਇ ਲਿਖਿਆ ਮਹਲਾ ੫
stays Guru Arjan Dev Ji's).

The ledger (`pipeline/attribution_ledger.json`)

Status	Meaning
approved	Applied by rule E (Japji → Guru Nanak Dev Ji; Ramkali Sadu → Baba Sundar Ji).
proposed	Recorded for scholarly review, changes nothing: Bhai Mardana's saloks (Ang 553), Raagmala, the per-Bhatt splits, Surdas's single line (Ang 1253), saloks signed by other Bhagats inside Kabir's collection.
keep	Our tag stands on printed evidence although the independent edition differs (Angs 326, 1136, 1253).

The gate (`pipeline/verify_attribution.py`)

```
python3 pipeline/verify_attribution.py <pre.jsonl> corpus/sggs.jsonl --oracle
database.sqlite \
--review-dir validation/attribution-review
```

1. Only author differs between the previous and the new corpus (text, comp_id, line_no, is_header, markers are identical).
2. The pass is idempotent.
3. Every line that still disagrees with the independent edition lies inside a ledger entry.

The review pack (`validation/attribution-review/README.md`) lists each re-attributed composition with its printed heading and its first and last verse, verbatim, for human review.

The independent edition (`database.sqlite`, pinned by `pipeline/shabados.SHA256SUMS`) is a cross-check only: it is never copied into the corpus, never shipped and never cited. `pipeline/attribution_audit.py` aligns it line by line to produce the diff.

Links

1. <https://github.com/Algorythmos-AI/siggs-data/blob/c36e3172bdf38a61070849bb604b38da89ca5d8b/docs/architecture/attribution.md>

Sri Guru Granth Sahib Ji — Knowledge Base 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/attribution/>