

Data · Runbooks · pinned from sggs-data

Runbook: restore the dataset from the backups (and the quarterly drill)

🕒 3 min read · Updated 27 October 2026

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

Pinned copy. The canonical file is [sggs-data/docs/process/runbooks/data-restore.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.

The dataset can always be rebuilt from its sources alone. This runbook restores it, and the same procedure is the **quarterly disaster-recovery drill** that proves it still works.

Where everything lives (3-2-1)

Asset	Copy 1	Copy 2	Copy 3
Source PDF (pdf_sha256 in validation/reconcile-attestation.json)	private Algorythmos-AI/sggs-source, release source-pdf-v1	R2 bucket sggs-source-backup, prefix source-pdf-v1/	encrypted offline drive (pending)
English translation sources (pipeline/translations.SHA256SUMS)	sggs-source, release translations-en-v1	R2 sggs-source-backup, prefix translations-en-v1/	encrypted offline drive (pending)
ShabadOS database 4.8.7 (pipeline/shabados.SHA256SUMS)	sggs-source, release shabados-database-4.8.7	R2 sggs-source-backup, prefix shabados-database-4.8.7/	encrypted offline drive (pending)
Corpus, database, ledger, fingerprints, history	this repository (git + LFS)	sggs-source release data-v1.0.0	every clone

Nothing in sggs-source is ever replaced or deleted: a new edition or release gets a new release.

Copy 2 is a Cloudflare R2 bucket, sggs-source-backup, in the company Cloudflare account. It holds each sggs-source release under a prefix named after the release, with the release’s own checksum files. Every object was downloaded back and verified by sha256 when it was stored (2026-09-25). The bucket lock rule keep-forever retains every object indefinitely, so nothing stored there can be deleted or overwritten. A new release gets a new prefix. To restore from it when GitHub is unavailable (Cloudflare access needed, e.g. npx wrangler login):

```
npx wrangler r2 object get sggs-source-backup/source-pdf-v1/SHA256SUMS --file SHA256SUMS
--remote
npx wrangler r2 object get "sggs-source-backup/source-pdf-v1/Siri-Guru-Granth-Sahib-in-
Gurmukhi-with-Index.pdf" --file Siri-Guru-Granth-Sahib-in-Gurmukhi-with-Index.pdf --remote
shasum -a 256 -c SHA256SUMS
```

The same pattern restores translations-en-v1/ and shabados-database-4.8.7/. Verify every file against its checksum file (and the PDF against pdf_sha256) before using it.

Restore (or drill) — one command

```
bash scripts/data/dr_drill.sh # fresh clone of main, restore,
rebuild, compare
bash scripts/data/dr_drill.sh --ref data-v1.0.0 # any past commit or tag
```

It needs gh with read access to sggs-source, ~5 GB free disk, and the pinned toolchain: make toolchain builds toolchain/Dockerfile (base image pinned by digest, the libraries in pipeline/requirements-toolchain.txt, gh pinned by checksum), and docker run --rm -e GH_TOKEN -v "\$PWD:/repo" sggs-toolchain bash scripts/data/dr_drill.sh runs the drill inside it — exactly what the quarterly workflow does. It clones this repository fresh (nothing from your checkout is used), restores the PDF, the translation sources and the ShabadOS database from sggs-source — each verified against its committed hash — runs the full fail-hard rebuild (reconcile character, golden checks, every gate), then requires:

- the rebuilt corpus/sggs.jsonl byte-identical to the committed one, and
- the rebuilt database's fingerprint equal to audit/dataset-fingerprint.json: scripture (scripture_sha256, t0_sha256), every table and every FTS index.

On a different platform from the one that built the release (the Linux CI drill vs a macOS release build), tables holding computed floating-point values may differ in the last digit; --allow-float-tables reports those as WARN. Scripture, structure, text and index differences always fail.

A restored database is then installed by the normal path: a reviewed PR, then the platform's dataset.lock.json bump. Never copy a database into production by hand.

The quarterly drill

The dr-drill workflow runs on a clean GitHub runner on 1 January, April, July and October (and on demand). It needs the repository secret SGGs_SOURCE_TOKEN: a fine-grained token with read-only **Contents** on Algorythmos-AI/sggs-source. After each run, add its RESULT line to [audit/dr-drills.md](#)². A failed drill is a P1: the dataset is not recoverable as released until it passes again.

Links

1. <https://github.com/Algorythmos-AI/sggs-data/blob/c36e3172bdf38a61070849bb604b38da89ca5d8b/docs/process/runbooks/data-restore.md>

2. <https://github.com/Algorythmos-AI/siggs-data/blob/c36e3172bdf38a61070849bb604b38da89ca5d8b/audit/dr-drills.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/process/runbooks/data-restore/>