DATASET

Inhaltsverzeichnis

docs/DATASET.md

[← Back to README](#)

Summary

This project uses a local Postgres database (`esg_db`) that is initialized by Docker on first startup.
The assistant queries this database via SQL.

The database is populated by importing CSV files from `./data` (mounted into the dataset container as `/data`).

Where to obtain the dataset

The dataset is publicly available via OSF:

- https://osf.io/q2jpv/
- Files: https://osf.io/q2jpv/files/osfstorage

If you cannot access it, request the CSVs from the author (see README contact).

Academic source (APA 7)

Forster, K., Keil, L., Wagner, V., Müller, M. A., Sellhorn, T., & Feuerriegel, S. (2025). *Assessing Corporate Sustainability with Large Language Models: Evidence from Europe* (TRR 266 Accounting for Transparency Working Paper Series No. 202). https://doi.org/10.2139/ssrn.5361703

Required CSV files (minimum)

These files must exist locally BEFORE starting Docker (because init SQL uses COPY):

- data/raw/companies.csv
- data/raw/indicator_metadata.csv
- data/processed/esg_indicators_postprocessed.csv

The exact CSV → table mapping is documented in:

- data/dataset-manifest.json

Import mechanism

Docker Compose mounts:

- ./data → /data (read-only)
- ./db/esg/init → /docker-entrypoint-initdb.d (read-only)

On first container startup, Postgres runs:

1) db/esg/init/00_schema.sql
   Creates schema and tables.
2) db/esg/init/10_load.sql
   Imports CSV data using COPY statements (from `/data/...`).

Database tables (3)

1) esg.companies (raw)

Company master data (per year).
Typical columns: firm UUID, name, isin, country, primary sector, year.

2) esg.indicator_metadata (raw)

Metadata describing ESG indicators (topic, label/specification, type_standard, etc.).

3) esg.esg_indicators_postprocessed (processed)

Main fact table with extracted observations:

- company_id, year, company_year
- data_point_id, data_point_description
- model_output, model_output_valid
- value_final, unit_final

Notes on missing values

Missing values do not necessarily imply errors. Often an indicator was not reported by a company in a given year, or it was not applicable/material for that company-year context.

← Back to README