

# Synthetic CCS–MRV Dataset for Carbon Storage Reporting (SCCS-MRV)

Authors: Sreekanth Muktevi, Yogesh Nagpal, Rajesh Leela Thotakura, Jyotsna Muktevi

Version: v1.0 | DOI: <https://doi.org/10.5281/zenodo.17003094>

## Abstract

This white paper introduces the SCCS-MRV dataset (v1.0), the first publicly available synthetic dataset for Carbon Capture and Storage (CCS) Measurement, Reporting, and Verification (MRV). It is explicitly aligned with international standards including EPA 40 CFR Part 98 Subpart RR, ISO 27916, and OGMP 2.0, and mapped to the Open Footprint (OFP) and Open Subsurface Data Universe (OSDU) data models. The dataset provides facility-level annual summaries, monthly and daily injection records, schema definitions, and QA rules. It enables interoperability testing, regulatory prototyping, academic research, and ESG reporting validation while protecting confidentiality by using synthetic but realistic values.

## 1. Background & Motivation

Carbon Capture and Storage (CCS) is a cornerstone of global decarbonization strategies. To ensure transparency and credibility, operators are required to report activities under regulatory frameworks such as EPA Subpart RR, ISO 27916, and OGMP 2.0. However, real CCS facility data is rarely shared publicly due to confidentiality. The SCCS-MRV dataset addresses this gap by providing a synthetic but realistic dataset that supports open research, prototyping, and interoperability validation.

## 2. Methods & Data Generation

The dataset was generated using Python (NumPy, Pandas, Faker, GeoPy) with explicit QA rules to ensure realism. Synthetic coordinates were bounded within U.S. limits and flagged as non-real. Key rules include:

- Mass balance enforced:  $\text{Injected} - \text{Produced} - \text{Transport Loss} - \text{Leak Mass} = \text{Net Stored CO}$
- Reservoir logic: Basalt/Saline negligible production; Depleted Gas 0.1–5% recycling
- Capture energy intensity: validated against MEA, MDEA, Selexol, PSA, Amine technology ranges
- Additivity check: daily totals reconcile with monthly and annual
- Leak events: 0.1 tonnes if events > 0, else 0

These rules ensure consistency with MRV standards while keeping values entirely synthetic.

## 3. Schema Alignment

The dataset is mapped to OSDU Well-Known Schemas (WKS) and Entities (WKE) and OFP emissions model:

- Facility: facility\_id, country\_code, lat, lon, reservoir\_type
- Asset: capture\_tech, pipeline\_length\_km, transport\_mode
- Wellbore/Well: well\_id, injection\_start\_date, injection\_end\_date, reservoir pressure/temp
- ProductionData: co2\_injected\_tonnes, co2\_produced\_tonnes
- Measurement: co2\_captured\_tonnes, capture\_energy\_MWh, transport\_loss\_tonnes, ch4\_emissions\_tonnes
- Event/Method: leak\_events\_count, leak\_mass\_tonnes, mmv\_methods, ogmp\_source\_category
- Extension: co2\_net\_stored\_tonnes

This alignment ensures compatibility with industry data models and regulatory schemas.

## 4. Dataset Contents

The SCCS-MRV package (v1.0) includes:

- ccs\_full\_dataset\_v1.0.csv – Facility-level annual summary (10 facilities)
- ccs\_injection\_daily\_v1.0.csv – Daily injection timeseries

- ccs\_injection\_monthly\_v1.0.csv – Monthly roll-up from daily
- schema/schema.yaml – Schema definition
- schema/schema\_crosswalk.csv – Field mapping to OFP/OSDU/MRV
- docs/NOTES.md – Methodology and QA rules
- LICENSE.txt – License (CC BY 4.0)
- CITATION.cff – Citation metadata
- README.md – Documentation and authorship

## 5. Usage & Applications

This dataset is intended for:

- Prototyping CCS MRV reporting systems
- Testing OFP–OSDU interoperability
- Benchmarking analytics pipelines for storage monitoring
- Academic research, teaching, and training
- Machine learning model development for energy sustainability

## 6. Licensing & Citation

This dataset is licensed under Creative Commons CC-BY 4.0. Users may share and adapt the dataset with appropriate attribution.

Citation:

Muktevi, S., Nagpal, Y., Thotakura, R. L., & Muktevi, J. (2025). SCCS–MRV Synthetic Dataset (v1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.17003094>

## 7. Disclaimer

This dataset is entirely synthetic. It does not represent any real facility, reservoir, or company data. It is intended solely for educational, research, and development use.