REDI: Automated, Auditable Data Preparation for AI

Sean R. Wilkinson (wilkinsonsr@ornl.gov), Polina Shpilker, and Wesley H. Brewer of Oak Ridge National Laboratory

AI Day for Federal Statistics 2026 · code.ornl.gov/drai/redi

Motivation

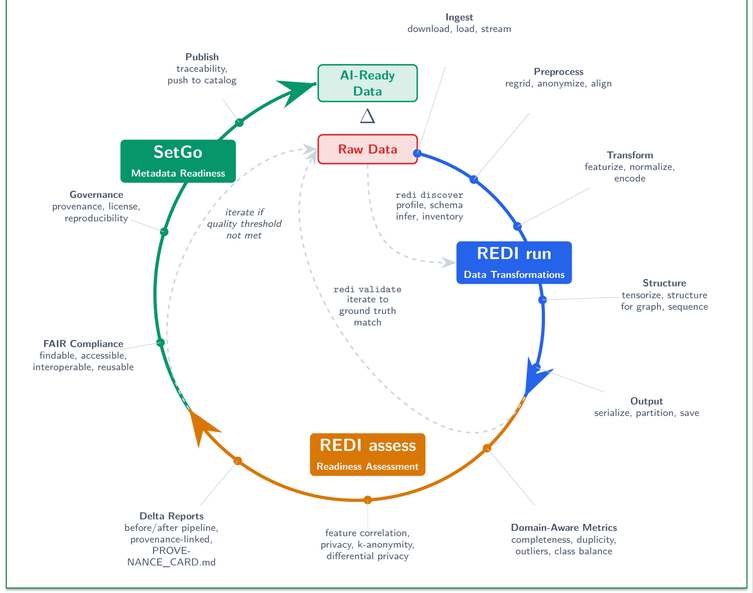
Federal agencies face a critical barrier to AI adoption: data preparation.

NASEM 2026 identifies data preparation as the statistical work "most ripe for AI acceleration" — yet today's pipelines are:

- Siloed — rebuilt from scratch for every project
- Undocumented — transformations opaque, not auditable
- Fragile — manual steps, no provenance, not reproducible

"The core of these infrastructures requires unified data ecosystems where findable, shared, and version-controlled repositories are kept with standardized preprocessing procedures and in-depth metadata."
— National Academies of Sciences (NASEM 2026)

The REDI-SetGo Data Readiness Lifecycle



How REDI Works

Readiness Engine for Data Integration (REDI) executes a unified five-stage pipeline — Ingest, Preprocess, Transform, Structure, Output — using a shared PipelineContext that carries data and metadata through every stage.

- Domain-aware: selects transformations (PII hashing, missing-value fill, format conversion) based on detected data type and config.
- Provenance via Flowcept: every `execute()` call is instrumented — before/after data snapshots captured automatically at each step.
- 25+ built-in transformation functions; domain plugins install via `redi install` with no core changes.
- Parallel backends: Python multiprocessing, MPI, GNU Parallel, auto-generated Slurm scripts.

Case Study: 2023 ACS PUMS Person File

The American Community Survey Public Use Microdata Sample (ACS PUMS) is the U.S. Census Bureau's largest public microdata release, representing the full U.S. resident population with detailed person-level characteristics. REDI processed the complete 2023 1-year person file — no sampling, no manual preprocessing.

Before REDI

ACS PUMS 2023

pums-2023.csv
(52 state CSVs)

3,433,236 records
287 variables
2.3 GB on disk

Issues present:
⚠ 150,124,670 missing values
(15.2% of all cells)
⚠ PII: PUMA, SERIALNO
⚠ Format: CSV (not FAIR)

`redi run pums-2023.csv -c acs-pums.yaml`

1 Ingest
Load & validate 52 state CSV files; auto-detect format

2 Preprocess
Fill all 150,124,670 missing values (domain-aware strategy) → 0 remaining

3 Transform
Anonymize PII: PUMA & SERIALNO → SHA-256 hash (16-char prefix)

4 Structure
Shuffle and split 80% train / 10% val / 10% test (seed = 42)

5 Output
Write Parquet files · DATA_CARD.md · metadata.json · Flowcept provenance record

After REDI

AI-Ready Dataset

Output files:
✓ train.parquet
2,746,588 rows
✓ val.parquet
343,323 rows
✓ test.parquet
343,325 rows

✓ DATA_CARD.md
✓ metadata.json

0 missing values
PII suppressed
FAIR-compliant
Parquet format

Key Results at a Glance

3.4 M records processed	**287** variables	**150 M** missing values resolved
≈15% of cells were missing	**2m 46s** on a commodity laptop	**100%** FAIR-compliant Parquet

MacBook Pro M1 Pro · 16 GB RAM · Normal background load · 15 ok · 0 warnings · 0 errors

Why This Matters for Federal Statistics

The ACS PUMS — 3.4 M records, 287 variables — is the largest public federal microdata release. REDI makes it AI-ready in one command, in under three minutes.

- **Disclosure avoidance built in**: PUMA & SERIALNO SHA-256 hashed automatically via `pii_columns` to enforce federal privacy without custom code.
- **Full audit trail**: Flowcept logs every transformation decision with before/after statistics and timestamps to meet federal reproducibility requirements.
- **Config-driven, not code-driven**: steps declared in YAML, not Python so that data stewards run and adapt pipelines without software engineering expertise.
- **Scales with the data**: the same `redi run` command works unchanged on a laptop and on leadership-class HPC with no code changes required.
- **FAIR-ready for cross-agency sharing**: Parquet + DATA_CARD.md + metadata.json are ready for catalog publication via SetGo, enabling reuse without rework.
- **Trustworthy by design**: every preprocessing decision is documented, versioned, and re-runnable, all of which are essential when outputs inform policy.

"Data scientists spend 60–80% of project time on preprocessing." REDI moves that into a config file — freeing researchers for modeling and analysis.

FAIR Compliance & Provenance

Findable · Accessible · Interoperable · Reusable

Every REDI run automatically produces:
✓ DATA_CARD.md — human-readable dataset documentation
✓ metadata.json — machine-readable pipeline record
✓ PROVENANCE_CARD.md — step-level audit trail
✓ Pre/post metrics for every transformation step

AI-ready datasets can be published to Hugging Face, CKAN, and other catalogs via the companion SetGo tool.

Demonstrations in Other Domains

REDI has been demonstrated on four scientific HPC domains, each validated against peer-reviewed reference pipelines:

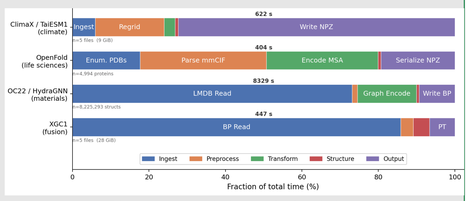
Domain	Scale	Key REDI Transforms	Output
Climate ClimaX	4,766 5.4 TB	Regrid (xESMF) → Z-score normalize → lat/lon/time tensor	NPZ
Proteomics OpenFold	250k 4.3 TB	Extract mmCIF + A3M → one-hot encode → per-protein feature dict	NPZ
Materials HydraGNN	482M atoms 10 TB	Parse DFT/JSON → cutoff-radius graph edges → Z-score node features	ADIOS2 BP
Fusion XGC1	7,089 106 TB	ADIOS read → mesh project 2D→3D → min-max normalize	.pt files

Table. Datasets demonstrated across four HPC domains.

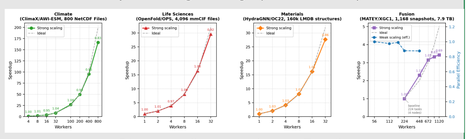
Validated against domain-expert reference pipelines.

Scales to Frontier — no code changes:

- Climate: ≥95% efficiency, 800 workers
- Fusion: 13.9× speedup, 7.9 TB
- Proteomics & Materials: 27–30× speedup



Pipeline stage timings (% of wall time)



Strong scaling speedup on Frontier

* - S.R. Wilkinson, J.Y. Choi, K. Maheshwari, M. McDonnell, M. Lupo Pasini, P. Shpilker, R. Souza, P. Widener, S. Oral, W.H. Brewer. "Automated Data Readiness for Scientific AI," submitted to SC26: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2026.

Conclusions

REDI makes AI-ready data a solved problem:

- ACS PUMS 2023: 3.4 M records, 150 M missing values, PII anonymization — under 3 min on a laptop.
- Same pipeline scales to terascale on Frontier with no code changes.
- Full provenance at every step for federal auditability and reproducibility.
- Advances NASEM 2026's call for "unified data ecosystems with standardized preprocessing."

Get the code:

code.ornl.gov/drai/redi

Open-source on June 1 · Federal agency collaboration welcome

Work performed at Oak Ridge National Laboratory under DOE Contract DE-AC05-00OR22725. AI writing assistance (Anthropic Claude) used for editing.