**CS598 Course Project Progress Report: End-to-End Data Curation Workflow**

**Team Members:**   
  
Cesar Mancillas  
*UIN: 651066529  
email:* [*cam37@illinois.edu*](mailto:cam37@illinois.edu)  
Aristofanes Cruz  
*UIN: 655558479  
mail:* [*ac163@illinois.edu*](mailto:ac163@illinois.edu)  
Cesar Nava   
*UIN: 654326785d  
mail:* [*Can14@illinois.edu*](mailto:Can14@illinois.edu)

Repo: <https://github.com/ceasarm7/cs598-e2e-datacuration>

**Status of Deliverables and Progress against Proposed Plan**

Our project, End-to-End Data Curation Workflow for Climate Change Impact Data, remains on schedule and aligns closely with our proposed timeline.  
We have completed the Data Acquisition & Modeling and Quality Assessment phases and finalized the initial cleaning and transformation phase using both OpenRefine and Python/Pandas.  
All deliverables for the progress report milestone have been met and documented in the GitHub repository under the /artifacts directory.

| **Milestone** | **Proposed Date** | **Status** | **Evidence of Progress** |
| --- | --- | --- | --- |
| **Proposal** | 9/15 | Completed | Proposal PDF submitted. |
| **Data Acquisition & Modeling** | 9/20-9/30 | Completed | Datasets collected, primary dataset identified as "Annual Temperature Anomalies" from Our World in Data. Relational schema defined. |
| **Quality Assessment** | 10/1-10/10 | Completed | Initial data profiling performed; key data quality issues identified (Missing Data, Inconsistent Formats, Duplicates, Incorrect Data Entries/Outliers). |
| **Cleaning & Transformation** | 10/11-10/20 | Completed | OpenRefine + Python initial cleaning artifacts |
| **Progress Report** | 10/27 | On Schedule | This report submitted. |

**Data Acquisition and Modeling**

We successfully acquired the **Climate Change Impacts Data (Annual Temperature Anomalies)** from Our World in Data, which is publicly available under the CC BY 4.0 license, ensuring legal compliance.

Our relational model defines a core table **TemperatureAnomalies** with *country, country\_code, year,* and *temperature\_anomaly*, optimized for longitudinal analysis.  
This model directly supports our primary use case (U1): analyzing long-term regional and global temperature trends to understand climate change impacts*.*

**Quality Assessment**

In the quality assessment phase we identified several critical data quality problems:

* **Missing Data**: Especially for early 20th-century records or low-coverage regions.
* **Inconsistent Formats**: Variations in measurement baselines (e.g., 1951–1980 reference).
* **Duplicates**: Repeated (country, year) entries across multiple versions.
* **Outliers**: Implausible anomaly values (> ±5 °C) requiring verification.

These findings informed the cleaning plan and guided the construction of both OpenRefine transformations and Python validation scripts.

**Cleaning & Transformation**

This phase was divided into two complementary workflows:

**OpenRefine Operations (v3.9.5):**   
Used for interactive profiling and normalization.  
Key operations (documented in openrefine\_operations.json) include:

* Case and whitespace normalization for country names and codes.
* Numeric type coercion for year and temperature\_anomaly.
* Composite key generation (key\_code\_year) to detect duplicates.
* Clustering (fingerprint and n-gram fingerprint) to unify name variants.
* Annotation flags for missing or extreme values.

The cleaned result was exported as temperature\_anomalies\_refine\_clean\_base.csv.

**Python Initial Cleaning Script:**   
Implemented in artifacts/initial\_cleaning\_script.py, this script automates the initial reproducible cleaning pipeline, performing:

* Column name standardization and schema enforcement.
* Country code normalization (ISO-3 alignment with extended mappings).
* Type coercion for numeric and string fields.
* Deduplication of (country\_code, year) records.
* Lightweight data-quality (DQ) report summarizing all missing values and duplicates.

Outputs:

* temperature\_anomalies\_initial\_clean.csv: standardized dataset snapshot.
* temperature\_anomalies\_initial\_dq\_report.csv: QA summary report.
* schema\_definition.sql: relational schema reflecting cleaned structure.

**Justification of Changes in Scope or Deliverables**

As noted in our proposal feedback, we refined the initial scope from the broader *Climate Change Impacts Data* to focus on the *Annual Temperature Anomalies* subset.  
This adjustment increases focus and feasibility while maintaining alignment with the course learning goals::

* **Focus and Manageability:** Narrowing to one key climate indicator enables deeper quality control and reproducibility.
* **Analytical Value:** Temperature anomalies remain a central metric for climate analysis, directly tied to our use case.
* **Deliverable Alignment:** This scope still supports all required concepts (data lifecycle, metadata, and reproducibility).

**Challenges Identified and Scope Adjustments:**

* **Missing Data:** Selecting appropriate imputation methods without biasing long-term trends remains a key challenge.
* **Baseline Standardization:** Reconciling different reference baselines may become necessary if auxiliary datasets (e.g., emissions or population) are integrated.

To mitigate these issues, we will prioritize achieving high internal consistency within the primary dataset before extending integration to any secondary sources. The project remains feasible, and we are confident in meeting the final submission due on **12/10**.

**Evidence of Progress through Corresponding Artifacts**

The following artifacts provide verifiable evidence of progress and are included in the repository:

* artifacts/openrefine\_operations.json: reproducible transformation sequence.
* artifacts/schema\_definition.sql: schema for relational integration.
* artifacts/initial\_cleaning\_script.py: automated Python cleaning.
* artifacts/temperature\_anomalies\_refine\_clean\_base.csv: OpenRefine export.
* artifacts/temperature\_anomalies\_initial\_clean.csv: standardized dataset.
* artifacts/temperature\_anomalies\_initial\_dq\_report.csv: QA metrics summary.

**References**

Our World in Data. (2025). *Climate change data*. <https://ourworldindata.org/climate-change>.