**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)

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

Our project, the "End-to-End Data Curation Workflow" for climate change impact data, has followed the proposed timeline and is currently on schedule. We have successfully completed the first two major milestones outlined in the project plan, and we are nearing completion of the "Cleaning & Transformation" phase.

| **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 | In Progress | Initial cleaning planned for OpenRefine and Python/Pandas. Initial scripts developed to address Missing Data and Duplicates (Artifacts: Initial Cleaning Script - Python/Pandas). |
| **Progress Report** | 10/27 | On Schedule | This report submitted7. |

**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 data modeling focused on the primary time-series nature of the dataset, defining a clear relational schema for **TemperatureAnomalies** with *country, country\_code, year,* and *temperature\_anomaly* as key fields. The **use case (U1)** driving our curation is the analysis of global and regional temperature anomaly trends over the past century to understand climate change impacts in different regions.

**Quality Assessment**

The quality assessment phase, led by Aristofanes Cruz, identified several critical data quality problems:

* **Missing Data:** Specifically for earlier decades or less-reported regions.
* **Inconsistent Formats:** Potential variations in temperature units or baseline reference periods (e.g., the dataset uses a 1951-1980 baseline).
* **Duplicates:** Repeated records for the same region and year.
* **Incorrect Entries/Outliers:** Implausible temperature anomaly values needing flagging or correction.

**Cleaning & Transformation**

We have initiated the **Cleaning & Transformation** phase. The plan is to use **OpenRefine** for initial, high-level data profiling and deduplication, and **Python (Pandas)** for advanced statistical cleaning, such as filling missing values with logical methods and identifying outliers. Initial Python scripts have been developed (Artifacts: *initial\_cleaning\_scripts.py*) to handle deduplication and standardize country codes, addressing two of the identified issues.

**Justification of Changes in Scope or Deliverables**

Our original proposal used the broad "Climate Change Impacts Data". In Phase I, we **narrowed the scope** to focus on the **Annual Temperature Anomalies** subset of this data. This is a justified change because:

1. **Increased Focus and Feasibility:** Focusing on one specific, critical climate indicator (temperature anomalies) makes the end-to-end curation task more manageable and feasible within the project timeline, while still demonstrating all required course concepts (data cleaning, integration, metadata.
2. **Clearer Use Case:** The specific temperature anomaly dataset directly supports the main analysis question (U1) regarding long-term regional temperature trends.

No other changes to the overall deliverables or subsequent milestones have been made.

**New Challenges Identified and Scope Adjustments**

**Challenges and Risks**

We have encountered a few potential challenges that align with our initial constraints.

1. **Handling Missing Data:** The decision on the **best method to fill missing values** (e.g., regional mean, interpolation, or flagging) is challenging, as the wrong approach could introduce spurious trends. This requires careful justification to ensure the cleaned data (D') is verifiable against the use case (U1).
2. **Standardization of Baselines:** The dataset uses a **1951-1980 baseline**. If we integrate an auxiliary dataset (e.g., socioeconomic indicators) that uses a different baseline for its temperature data, reconciling these inconsistent formats will require a robust standardization procedure.

**Scope Adjustments**

To mitigate the **Standardization of Baselines** risk, we are making a minor, proactive adjustment:

* **Focus on Internal Consistency:** We will prioritize achieving high internal consistency within the *primary* dataset (Temperature Anomalies) before attempting to integrate any **secondary datasets** (e.g., Population or Emissions data). The use of secondary datasets for enrichment will be revisited *only* if the timeline allows after the core curation workflow is complete, ensuring the central deliverable remains feasible.

The project remains feasible, and we are confident in meeting the Final Submission deadline of 12/10

**Evidence of Progress through Corresponding Artifacts**

The submitted artifacts demonstrate progress in the **Data Acquisition**, **Modeling**, and **Quality Assessment** phases, and the start of **Cleaning & Transformation**. These are available in the accompanying zip file/GitHub repository (link/file name placeholder).

* *raw\_temperature\_anomalies.csv*: The initial dataset acquired from Our World in Data.
* *schema\_definition.sql*: The basic SQL schema for the *TemperatureAnomalies* table.
* *quality\_profile\_report.pdf*: The initial profiling report (generated via OpenRefine or a Python notebook) highlighting the identified issues (Missing Values, Duplicates, Outliers).
* *initial\_cleaning\_scripts.py*: Python/Pandas script showing initial steps for deduplication, country code standardization, and a function for identifying potential outliers

**References**

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