

HS CS598 — Progress Report

Course: Foundations of Data Curation

Student: Hard Shah

Date: October 27, 2025

1 Introduction

This progress report documents the current status of my data-curation project, which focuses on enhancing the quality, reusability, and transparency of the **AI4I 2020 Predictive Maintenance Dataset**. The goal is to design an end-to-end curation workflow that improves reproducibility and ensures the dataset can be responsibly reused in research and teaching contexts.

The report describes progress since the proposal stage, how instructor feedback has been incorporated, challenges faced, evidence of progress through deliverables, and the remaining work planned before final submission.

2 Incorporation of Proposal Feedback

Feedback from the proposal (14 / 15 points) highlighted two improvement areas:

1. Linking the project explicitly to a recognized data-curation lifecycle model.
2. Examining dataset trustworthiness and potential bias, given the synthetic origin of the AI4I data.

2.1 Lifecycle Model Integration

I adopted the **Digital Curation Centre (DCC) Curation Lifecycle Model** as the guiding framework. The model supports sequential activities (conceptualization, ingest, preservation) as well as continuous actions (description, access, reuse). All project tasks have been mapped to corresponding DCC stages to ensure systematic progress.

2.2 Dataset Trustworthiness and Bias

Although the dataset is synthetic, it is hosted by the **UCI Machine Learning Repository** and referenced in peer-reviewed literature. The generation process, based on physics-inspired failure models, provides reasonable confidence in its conceptual validity.

Bias documentation has been expanded to include:

- **Class imbalance bias** (3.39 % failure cases).
- **Synthetic data bias**, as simulated conditions may omit environmental and operational variability.
- **Feature selection bias**, since only five variables were modeled for failure prediction.

All these limitations are clearly stated in the metadata and README so that future users can interpret results responsibly.

3 Progress Overview

Progress aligns closely with the proposed timeline and objectives.

3.1 Data Acquisition and Ethics (Completed)

The dataset was obtained from the UCI repository under a **CC BY 4.0** license. Provenance documentation now includes source, DOI, authorship, and acquisition date. Ethical review confirmed no personal or sensitive data are involved.

3.2 Data Cleaning and Transformation (In Progress)

Data validation confirmed correct data types and complete records. Outliers were detected and retained for integrity. Class distribution and feature correlations were analyzed and recorded. A modular Python pipeline handles normalization and quality control to ensure reproducibility.

3.3 Metadata and Documentation (Substantial Progress)

Metadata were prepared following the **DataCite 4.4 schema**, including:

- Title, creators, and license information
- Feature-level details with units and ranges
- Subject classifications and intended use
- Links to related identifiers and publications

A detailed **README** explains dataset structure, intended use, and limitations.

3.4 Workflow Automation and Provenance (Started)

A modular folder structure and version-controlled scripts now capture transformation logs, validation checks, and environment specifications. A provenance record tracks every modification, ensuring transparency and reproducibility.

3.5 Packaging and Preservation (Planned)

Final packaging will use open formats (CSV and JSON) and standard metadata. The curated dataset will be deposited in an open repository such as **Zenodo**, ensuring long-term accessibility and DOI assignment.

4 Evidence of Progress and Artifacts

All deliverables are organized in a structured directory and included in the ZIP submission. They demonstrate measurable progress at each stage of the curation process.

Folder / File	Description	Category
data/AI4I_2020.csv	Original dataset from UCI repository	Input Data
output/AI4I_2020_curated.csv	Cleaned and validated dataset	Primary Deliverable
output/AI4I_2020_curated.json	Machine-readable curated version	Derived Data
output/data_dictionary.json	Variable-level metadata	Metadata Deliverable
output/summary_statistics.csv	Quantitative data quality summary	Quality Assessment
metadata/metadata.json	Dataset-level metadata	Metadata Deliverable
metadata/provenance_record.json	Provenance log tracking transformations	Provenance Deliverable
scripts/	Python modules for data loading, transformation, validation	Workflow Code
notebooks/AI4I_Data_Curation.ipynb	Interactive notebook showing full pipeline	Documentation Artifact
docs/README.md	Overview and usage instructions	Project Documentation

docs/Progress_Report.md	Narrative report version	Report Artifact
Project_Summary.md	Summary of goals, methods, and results	Narrative Deliverable

These materials collectively demonstrate substantial progress, organization, and readiness for final integration.

5 Challenges and Scope Adjustments

Synthetic Data Transparency: The lack of published generation code limited verification. Mitigated through detailed documentation of known generation principles and full statistical profiling.

Class Imbalance: Rather than altering distribution, documentation now includes clear guidance for downstream handling to preserve dataset authenticity.

Metadata Standard Adaptation: Some ML-specific parameters did not align perfectly with DataCite. A layered documentation approach was adopted, combining structured metadata with supplemental README context.

Timeline Adjustment: Addressing dataset bias and metadata completeness required a one-week shift in workflow-automation tasks. The final completion target remains December 2025.

6 Next Steps

November Tasks:

- Finalize automation scripts and provenance capture.
- Complete quality-assurance checks and metadata fields.
- Prepare usage examples and validation documentation.

December Tasks:

- Package and deposit curated dataset.
- Generate final report summarizing full workflow, lessons learned, and recommendations.

These steps will complete the project within schedule and deliver a transparent, reusable curated dataset.

7 Reflection

The project demonstrates that even a synthetic dataset benefits greatly from structured curation. Applying the DCC Lifecycle Model ensured all critical actions—documentation, provenance, and preservation—were systematically addressed.

By clarifying dataset limitations and recording all curation decisions, this work increases trust, reproducibility, and educational value. The methods developed here can serve as a general template for curating similar machine-learning datasets in academic and industrial contexts.

References

- Digital Curation Centre (2023). *DCC Curation Lifecycle Model*. <https://www.dcc.ac.uk/guidance/curation-lifecycle-model>
- Matzka, S. (2020). *Explainable Artificial Intelligence for Predictive Maintenance Applications*. IEEE AI4I Conference, 69–74.
- DataCite Metadata Working Group (2021). *DataCite Metadata Schema v4.4*. <https://doi.org/10.14454/3w3z-sa82>
- UCI Machine Learning Repository (2020). *AI4I 2020 Predictive Maintenance Dataset*. <https://doi.org/10.24432/C5HS5C>