UNIVERSITY OF TORONTO
FACULTY OF INFORMATION
Digital Curation Institute
JSD Just Sustainability Design Lab

Auditing the Data Curation Practices of Power Systems Datasets

Eshta Bhardwaj & Christoph Becker · University of Toronto

NEURAL INFORMATION PROCESSING SYSTEMS
WiML Women in Machine Learning

Introduction

Power systems (PS)
- Cyber-physical system domain (i.e., network of physical and digital components, tightly coupled for sensing, communication, and control).
- Aimed at meeting a projected 80% increase in the electricity demand by mid-century.

- Data curation
  - An information science discipline that can inform methods to actively document and steward datasets in ML research.

Motivation

- In PS, inadequacy of conventional methods for modeling and control of stochastic grids with a high degree of controllability and low degree of observability has led to the increased usage and application of machine learning (ML).
- As ML use expands, there is a growing need for dataset development processes to be consistent and systematic.

Goal

We assess documentation to evaluate standards and data practices within power systems.

- We apply data curation, as a reference discipline, to examine ML dataset development in PS datasets.

Impact

The findings of this study are expected to transfer meaningfully to other cyber-physical system domains navigating the transition toward data-driven methods.

RQ

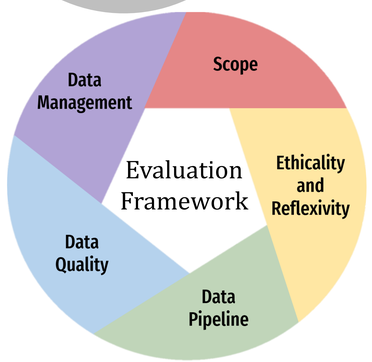
What is the state of data curation (DC) at the intersection of machine learning (ML) and power systems (PS) and how can it be further advanced?

Methods

Apply **DC evaluation framework** to 43 datasets representative of the PS field

Calculate IRR to establish the feasibility of the framework for PS datasets

Analyze how well PS datasets meet DC criteria



Context, purpose, motivation
Requirements

Ethicality
Domain knowledge & data practices
Context awareness
Environmental footprint

Data collection
Data processing
Data annotation

Suitability
Representativeness
Authenticity
Reliability
Structured documentation

Findability
Accessibility
Interoperability
Reusability

Context awareness

Context awareness demonstrates an understanding of the subjective, non-neutral nature, and situatedness of data.

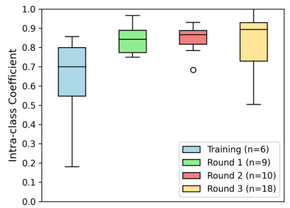
Criteria to meet minimum standard

Documentation includes a positionality statement.

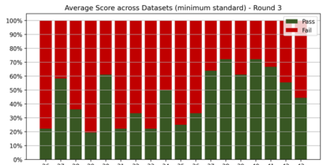
Criteria to meet standard of excellence

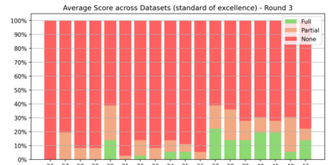
Documentation adopts a reflexive approach to dataset development. For example, documentation discusses how field epistemologies impact assumptions, methods, or framings.

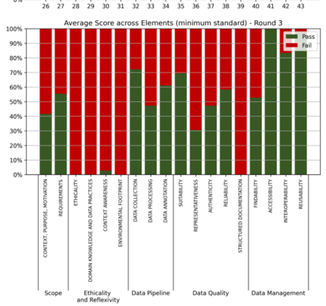
Findings

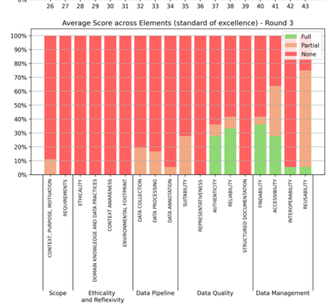


Inter-rater reliability shows improvement of median ICC over several rounds thus establishing the framework's feasibility and consistency for assessing power system datasets' documentation.









The analysis of completed documentation for both minimum and excellence standards for each dataset and element in Round 3 shows the following results:
- Documentation varies across datasets
- Documentation is often incomplete
- Documentation is not context-aware
- Documentation omits environmental footprint
- Documentation does not include structured documentation, such as datasheets
- PS datasets have a high standard for documentation corresponding to the FAIR principles

Recommendations

Context, purpose, motivation

- ☐ Provide explicit motivation.
- ☐ Justify specific gap filled by new dataset.

Ethicality

- ☐ Always document ethicality, even when considered non-problematic.
- ☐ Explicitly state if ethics review processes were needed, conducted, and their outcome.

Requirements

- ☐ Reflect on how problem formulation is a social construction.
- ☐ Consider the assumptions that justify the problem articulation.

Context awareness

- ☐ Include positionality statements.
- ☐ Demonstrate reflexivity, e.g., by reflecting on group power dynamics.
- ☐ Consider situatedness of dataset development and its potential impacts.

Data processing

- ☐ Consider "custody" to differentiate between processing and collection and ensure documentation of both processes.
- ☐ Data collection refers to a dataset coming into the "custody" of the dataset creator. Data processing is steps taken to change or transform the dataset after it has been collected.

Domain knowledge and data practices

- ☐ Provide documentation on background knowledge and skills to enable reuse and collaborations between fields.

Representativeness

- ☐ Consider the physical, regulatory, and operational context of the data collection environment .
- ☐ Document whether the context of the dataset is conceptually or statistically representative of the population.

Structured documentation

- ☐ Include standardized documentation such as datasheets, data cards, data statements, etc.