[**Project status report (Week 8)**](https://canvas.lms.unimelb.edu.au/courses/215540/assignments/542637) **-Group 21**

**Title: GenAI for Smarter Data Quality Assurance**

**Frequency of Meetings with Client: Fortnightly Meetings**

**Frequency of Meetings with Project group meetings: at least once a week**

**Week 5: Dataset Exploration**

We explored various public datasets related to urban infrastructure and spatial data, prioritizing those with real-world inconsistencies or missing information. Our aim was to identify data suitable for automated quality assessment.

**Week 6: Literature Review**

This week focused on reviewing academic literature on data quality frameworks, anomaly detection, metadata validation, and FAIR principles. We identified key models and clarified how ML and NLP can be used to improve data reliability.

**Week 7: Dataset Selection and Acquisition**

We faced challenges finding messy enough datasets, as many public sources were well-curated. Eventually, we selected a dataset from the AURIN platform that contains urban point-of-interest data with practical quality issues like missing fields and inconsistent labels—ideal for our objectives.

**Week 8: Preliminary Data Analysis and Client Meeting**

We analyzed attribute distributions, identified missing values, and began defining potential quality indicators. Planning for data preprocessing and custom quality metrics also began.

We held a meeting with the client to present our progress, including dataset selection and literature insights. The meeting confirmed the dataset’s technical suitability and helped align next steps for integrating our quality-checking pipeline.

**Rating for each member: Everyone in the group completed their task perfectly**

* **60%：Haoran Guo & Yang Jin & Yushi Wang(Each person contributes equally):** Select the dataset and conduct a preliminary analysis of the data
* **40%：Pranav Pai & Chaoge Zhu(Each person contributes equally):** Collect research literature and conduct a preliminary analysis of the dataset