Data Management Plan

Carrow Morris-Wiltshire

June 24, 2024

1 Collected Data

1.1 Sources

The data used in this project will be predominantly from secondary digital data sources accessed via API or as the results from simulation runs of existing agent-based models which will be through manual file transfer. Some secondary data will be manually extracted from websites and from meetings with stakeholders.

1.2 Specific Sources

Primary data collection through API requests to centralised internet of things (IoT) data repositories, such as the Urban Observatory. Additional API sources include the Open Geography Portal (Office for National Statistics) and the Nomanatim API (Open Street Map). Table 1 shows sample IoT pedestrian data from the urban observatory's database. All of the data from the urban observatory uses the same fields regardless of sensor type.

dt	value	veh_class	dir	location	category
2022-08-18 02:45:00	4	person	southwest_to_northwest	NclNorthumberlandStSavilleRowWest	flow
2022-06-11 20:15:00	9	person	northwest_to_southwest	NclNorthumberlandStSavilleRowWest	flow
2022-03-02 23:15:00	9	person	northeast_to_southeast	NclNorthumberlandStSavilleRowEast	flow

Table 1: Randomly sampled records from the urban observatory database

1.3 Licensing

- Urban Observatory data is available under the CCA4 license.
- Open Street Map data is available under the ODbL license.
- Open Geography Portal data is available under the OGv3 license.

2 Created Data

2.1 Sources

Data will be created through transformation of collected data. These transformations will take place within notebooks and scripts. The majority of the transformed data will not be

stored but will be fully reproducible through rerunning notebooks and scripts. The scripts will be accessible in a repository on my GitHub page.

2.2 Licensing

The data created by this project (including any scripts and notebooks) will be available through the MIT license.

3 Metadata Standards

3.1 Description

Adherence to established metadata standards for geospatial and IoT data to ensure consistency, clarity, and compatibility for data processing and analysis. The Newcastle University Research Data standards will be used and can be found here. Metadata standards will apply to the codebase, written documentation, and created data. A variety of standards will be used for these different data forms.

3.2 Codebase

The NERC guidelines will be applied to the following: data that were analysed; data processing methods; version of programming language used; list of libraries and their versions; scripts used; analysis techniques; additional datasets; methods of interpretation (visualisations); starting conditions (seeds) for any stochastic modelling. A README file in each folder as well as accompanying project data dictionary and metadata documents will be provided. Additionally, the SOLID principles will be followed for creating modules or libraries of code. These apply to object-oriented designs and help to ensure they are understandable, flexible, and maintainable:

- S - Single-responsibility principle
- O - Open-closed principle
- L - Liskov substitution principle
- I - Interface segregation principle
- D - Dependency inversion principle

All code will be written with the intent of it being published as a python package. Any code that needs sharing will be dockerised for maximum compatibility. The codebase will follow the paradigms of functional modularity, object oriented programming, config-driven development.

3.3 Written Documentation

For written documentation a logical file structure will be followed, and files will use this naming convention $YYYYMMDD\_AUTHORSINITIALS\_FILENAME$.

3.4 Created Data

Created data will follow various metadata guidelines depending on the context of the data. For streaming data, the Urban Observatories metadata standards will be adopted. These apply to the three areas shown in figure 1. The first is static data which relates to information such as the manufacturer of the IoT sensor, and the owner and licensing of the data. The second area is generated metadata, this relates to information about the data stream such as clear definitions for the different fields and how often the stream is refreshed. The last area is professional judgement, this will result from exploratory analysis and modelling, to find the limits of the data's suitability and utility for a range of tasks.

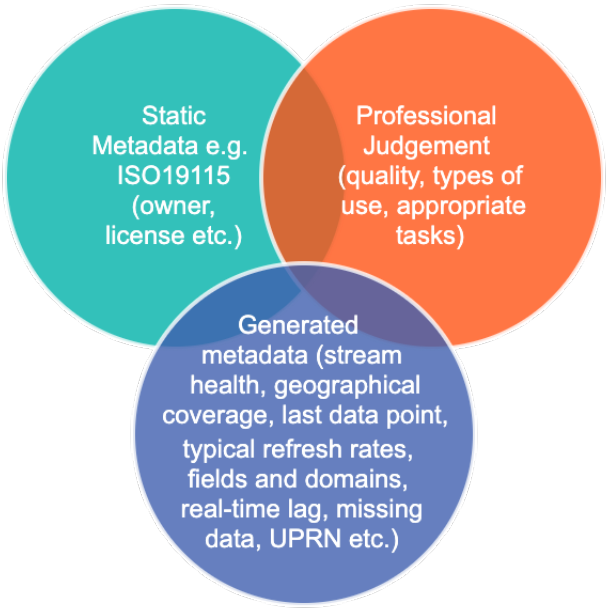

Figure 1: Urban observatory metadata engine for (near) real-time / streaming data

4 Data Structure and Storage

4.1 Structure

Data structure applied to the codebase and created data. The codebase will be structured following style guides of the relevant programming language. For example, PEP 8 and PEP 257 will be followed when using Python. Created data will be stored in the most suitable format for the AI and machine learning algorithms that are used. This will mostly depend on the language and machine learning framework used, for example if models are developed in PyTorch then it makes sense to store data in NumPy arrays or JSON.

4.2 Storage

Data storage will utilise secure, scalable cloud storage solutions for data preservation, with appropriate backup and recovery systems. This will involve using Microsoft OneDrive

for storing written documents and temporary small datasets, Git will be used for version control of the codebase, which will be periodically pushed to GitHub. Google Cloud Platform will be utilised for storage of large-scale streaming data and model outputs. Cloud service companies offer multiple redundancy so there is no need for further back-up. My personal computer will automatically perform a daily sync with iCloud to ensure that any temporary local files are recoverable.

5 Data Sharing

This project is committed to open data principles. This includes publishing in open-access journals where possible, making data free and open for use, and using open-source data ensure that proprietary licenses are not passed down in the process of data transformation. The FAIR principles, guidelines for good data management and stewardship, will be adhered to:

- Findable: Data should be easy to find for both humans and computers. Following the metadata guidelines in section 4 will ensure that data is findable.
- Accessible: Once found, data must be accessible. Using open-access platforms like GitHub discussed in section 5 ensures data is accessible.
- Interoperable: Data should be compatible with other datasets, tools, and workflows for analysis, storage, and processing. Following established frameworks like PyTorch for machine learning ensures that data outputs are interoperable.
- Reusable: Data should be well-described and richly documented to enable reuse and replication. Using design and style guides such as SOLID and PEP ensures that data is easily to interpret and manipulate for other users.

All significant work will be stored on GitHub and be made publicly available through my website.

6 Data Preservation

Long-term preservation plans will include clear guidelines on data retention periods and archival methods. Long-term storage will utilise cloud-based platforms (e.g., Google Cloud Platform) to ensure data safety and accessibility over time. Data on GitHub will be held for a period of 10 years post-project completion, with active maintenance for at least the first year. All data will be stored in accessible, non-proprietary formats (e.g., JSON) to facilitate long-term usability.

7 Ethical Considerations

All geospatial and IoT data used in the project will comply with data protection laws, such as GDPR. Personal identifiers will be removed or anonymised to protect individual privacy. Potential bias in data collection and algorithm development will be considered. Efforts will be made to ensure that AI models do not perpetuate or amplify existing societal biases.

Ethical considerations are covered in more depth in the responsible research and innovation plan.

8 Responsibility

8.1 Management

Carrow Morris-Wiltshire will be responsible for:

- Implementing and updating the DMP.
- Data collection, processing, and analysis.
- Ensuring data publication and public access at the project end (this includes managing any resource requirements).

8.2 Oversight

Reviews and audits will be periodically carried out to ensure adherence to the data management plan and compliance with legal and ethical standards. These will occur annually prior to progress review meetings and will involve external review.