Plan Overview

A Data Management Plan created using DMPonline

Title: Copy of Modelling International Airline Passengers to Generate Synthetic Passenger Name Record Data for Security Analysis

Creator: Muhammad Fathi Fadlian

Affiliation: The University of Sheffield

Template: Postgraduate Research DMP (The University of Sheffield)

Project abstract:

Passenger Name Record (PNR) data is a valuable source of information for aviation security, passenger mobility analysis, and the study of transnational travel behaviour. It contains structured details about passenger identity, bookings, itineraries, payment information, and travel companions, making it useful for analytical tasks such as watchlist matching, retrospective investigation, association detection, and the identification of Subject-of-Interest-indicative travel patterns. However, real PNR data is highly sensitive and is generally inaccessible for academic research due to legal, privacy, and security restrictions. This creates a significant data gap for researchers seeking to develop, test, and evaluate passenger screening, anomaly detection, and aviation security analytics methods.

This Confirmation Review presents a research framework for generating synthetic PNR data. Rather than only reproducing aggregate statistical distributions, the proposed framework aims to simulate international airline passengers as individual agents with demographic attributes, household structures, social networks, travel behaviours, and longitudinal travel histories. The framework integrates synthetic population generation, social network modelling, trip planning, flight booking, and structured XML PNR generation to produce individual-level records covering passengers, bookings, and flights. The resulting data is designed to preserve statistical, spatial, temporal, chronological, behavioural, and relational consistency while avoiding the use of real personal data.

The current proof-of-concept demonstrates the feasibility of generating structured synthetic PNR records from publicly available data and operational flight data. Preliminary evaluation shows that the framework can reproduce several high-level population and travel characteristics, while also revealing limitations that require further refinement, particularly in relation to behavioural calibration, group travel dynamics, seasonality, scalability, and Subject-of-Interest behaviour modelling. The remaining PhD work will focus on improving the framework, embedding Subject of Interest-indicative behavioural patterns, and developing a multi-dimensional validation strategy to assess realism, analytical utility, and fitness for aviation security research.

The expected contribution of this research is a privacy-preserving, reproducible, and extensible synthetic PNR generation framework that supports methodological experimentation in aviation security and passenger mobility analysis where access to real PNR data is not possible.

ID: 204309

Start date: 01-06-2024

End date: 31-05-2029

Last modified: 13-05-2026

Copyright information:

Copy of Modelling International Airline Passengers to Generate Synthetic Passenger Name Record Data for Security Analysis

Defining your data

- What digital data (and physical data if applicable) will you collect or create during the project?
- How will the data be collected or created, and over what time period?
- What formats will your digital data be in? (E.g. .docx, .txt, .jpeg)
- Approximately how much digital data (in GB, MB, etc) will be generated during the project?
- Are you using pre-existing datasets? Give details if possible, including conditions of use.

What data will you collect or create? Synthetic Passenger Name Record (PNR) datasets in XML format, generated entirely by the computational framework. No real passenger data is collected. Supporting data includes agent-based model configuration files, simulation logs, and statistical evaluation outputs. No physical data is involved.

How will the data be collected or created, and over what time period? Data is created programmatically using an agent-based modelling framework that simulates passenger populations, social networks, travel itineraries, and flight bookings. Generation runs on local machines and University of Sheffield HPC (Stanage). The proof-of-concept dataset was produced in 2024–2025; full-scale datasets will be generated iteratively from June 2026 to mid-2028.

What formats will your digital data be in? XML (PNR records conforming to PNRGOV EDIFACT-derived schema), Python scripts (.py), JSON and YAML (configuration/parameters), CSV (evaluation metrics and statistical outputs), .docx and .pdf (thesis and publications), .xlsx (analysis summaries).

Approximately how much digital data will be generated? The proof-of-concept (1,000 agents, 3-month window) produces approximately 50–100 MB of XML. Full-scale generation (target: 100,000+ agents, 12-month window) is estimated at 5–20 GB total across all experimental runs. Supporting code, logs, and evaluation outputs add roughly 1–2 GB. Total estimate: **10–25 GB**.

Are you using pre-existing datasets? Yes — publicly available secondary datasets only: OpenFlights (airport/route data, ODbL licence), GeoNames (geographic data, Creative Commons), and ONS/Eurostat demographic statistics (UK Open Government Licence / open access). No real PNR or passenger data is used at any stage. The TENACITy project outputs (analytical tools) also inform the SOI modelling approach but do not constitute input datasets.

Looking after data during your research

- Where will you store digital data during the project to ensure it is secure and backed up regularly? (University research storage)
- How will you name and organise your data files? (An example filename can help to illustrate this)
- If you collect or create physical data, where will you store these securely?
- How will you make data easier to understand and use? (E.g. include file structure and methodology in a README file)
- Will you use extra security precautions for any of your digital or physical data? (E.g. for sensitive and/or personal data)

Where will you store digital data during the project? Primary storage is a private GitHub repository for all code, configuration files, and smaller datasets, with version control providing a full change history. Large-scale generated datasets (XML outputs from HPC runs) will be stored on University of Sheffield research storage (managed via IT Services). Local working copies are maintained on a university-issued machine and backed up to OneDrive. HPC outputs on Stanage are retained under the university's standard research storage policies.

How will you name and organise your data files? The repository follows a structured directory layout: src/ (framework code), configs/ (simulation parameters), outputs/ (generated PNR data), evaluation/ (metrics and results), docs/ (thesis chapters and papers). Files are named descriptively with version or date suffixes, e.g. pnr_output_pop10k_3m_v2.xml (10,000 agents, 3-month window, version 2), eval_demographic_similarity_20270115.csv. Git branches and tags track experimental iterations.

Physical data? Not applicable — no physical data is collected or created.

How will you make data easier to understand and use? Each major directory includes a README documenting file structure, naming conventions, and generation parameters. Simulation configurations are stored as self-documenting YAML files specifying all input parameters. The framework codebase includes inline documentation and a top-level README covering installation, usage, and reproduction steps. Evaluation scripts include comments linking outputs to specific research questions.

Will you use extra security precautions? The GitHub repository is set to **private** throughout the project. Since all data is entirely synthetic and contains no real personal information, no additional encryption or access restrictions are required beyond standard university IT security policies. The framework is designed to be inherently privacy-preserving — no real PNR or passenger data is used or stored at any stage.

Storing data after your research

- Which parts of your data will be stored on a long-term basis after the end of the project?
- Where will the data be stored after the project? (E.g. University of Sheffield repository ORDA, or a subject-specific repository)
- How long will the data be stored for? (E.g. standard TUoS retention period of minimum 10 years after the project)
- Who will place the data in a repository or other long-term storage? (E.g. you, or your supervisor)
- If you plan to use long-term data storage other than a repository, who will be responsible for the data?

Which parts of your data will be stored on a long-term basis? The framework source code (final release version), representative generated PNR datasets used in published evaluations, evaluation scripts and result files, simulation configuration files, and the doctoral thesis. Intermediate experimental outputs and working drafts will not be retained long-term.

Where will the data be stored after the project? The framework source code and representative datasets will be deposited in ORDA (Online Research Data Archive, University of Sheffield) alongside the thesis. The codebase will also be made available as a public GitHub repository, as referenced in the CR document [98].

How long will the data be stored for? Minimum 10 years after project completion, in line with the standard University of Sheffield retention period. The public GitHub repository will remain accessible indefinitely.

Who will place the data in a repository? I (Muhammad Fathi Fadlian) will deposit the data and code in ORDA and ensure the GitHub repository is made public, with oversight from my primary supervisor, Prof. Vitaveska Lanfranchi.

If you plan to use long-term data storage other than a repository, who will be responsible? The public GitHub repository will be maintained by me as the project author. Should the repository become unmaintained, the ORDA deposit serves as the persistent archival copy under the university's stewardship.

Sharing data after your research

- How will you make data available outside of the research group after the project? (E.g. shared in a repository, either openly or with controlled access)
- Will you make all of your data available, or are there reasons you can’t do this? (E.g. personal data, commercial or legal restrictions, very large datasets)
- If there are reasons you can’t share all of your data, how might you make as much of it available as possible? (E.g. anonymisation, participant consent, sharing analysed data only)
- How will you make your data as widely accessible as possible? (E.g. include a data availability statement in publications, ensure published data has a DOI)
- What licence will you apply to your data to say how it can be reused and shared? (E.g. one of the Creative Commons licences)

How will you make data available outside of the research group? The framework source code is already publicly available as an open-source repository. After project completion, representative datasets and evaluation scripts will be openly accessible via ORDA. No controlled access restrictions are anticipated, as all data is entirely synthetic.

Will you make all of your data available, or are there reasons you can't? The vast majority will be made available. The only exceptions are full-scale generated datasets that are prohibitively large for repository storage (potentially 10–20 GB of XML). In such cases, the generation code and configuration files will be shared instead, enabling full reproduction. Any outputs produced in collaboration with the TENACITy consortium will be shared subject to consortium agreement.

If there are reasons you can't share all data, how might you make as much available as possible? For large datasets, representative subsets will be deposited alongside the full configuration files and random seeds required to reproduce the complete output. This ensures any researcher with access to the open-source framework can regenerate the data in full. No anonymisation is required as no real personal data exists in the project.

How will you make your data as widely accessible as possible? All publications will include a data availability statement linking to the ORDA deposit and GitHub repository. Deposited datasets in ORDA will be assigned a DOI for persistent citation. The GitHub repository README will link to the ORDA record and published papers. Conference and journal papers (e.g. the ISCRAM 2025 paper) already reference the open-source repository.

What licence will you apply? The framework source code is released under the **MIT Licence**. Deposited datasets and evaluation outputs will be released under **Creative Commons Attribution 4.0 International (CC BY 4.0)**, permitting reuse and redistribution with appropriate credit.

Putting your plan into practice

- Who is responsible for making sure your data management plan is followed? (E.g. you with the support of your supervisor)
- How often will your data management plan be reviewed and updated? (E.g. yearly and if the project changes)
- Are there any actions you need to take in order to put your data management plan into practice? (E.g. requesting University research storage via your supervisor.)

Who is responsible for making sure your data management plan is followed? I (Muhammad Fathi Fadlian) hold primary responsibility for day-to-day implementation of the plan, with oversight and support from my primary supervisor, Prof. Vitaveska Lanfranchi. My secondary supervisor, Neil Ireson, provides additional guidance particularly regarding data outputs related to the TENACITy collaboration.

How often will your data management plan be reviewed and updated? The plan will be reviewed annually and updated whenever significant changes occur to the project scope, data generation approach, or storage arrangements. Key review points include after the Confirmation Review viva, at the start of the iterative framework development phase (Jun 2026), and during thesis preparation (2029).

Are there any actions you need to take to put the plan into practice? Request University of Sheffield research storage allocation via my supervisor for hosting large-scale HPC-generated datasets. Ensure the GitHub repository is correctly configured with the MIT Licence and appropriate documentation before the next publication. Confirm TENACITy consortium data-sharing agreements with Neil Ireson for any collaborative outputs. Register the project with ORDA ahead of the first dataset deposit to ensure a DOI is reserved for citation in upcoming papers.