Minimal documentation standard for synthetic data collections

Synthetic data plays a crucial role in enabling innovation, wider data sharing and privacy preservation. This document provides a clear and concise framework for the documentation of synthetic data collections. Its aim is to ensure transparency, reproducibility, and usability, making synthetic datasets accessible and valuable for their intended applications, and avoiding misuse.

By following these guidelines, data creators can provide sufficient context and detail about the dataset's origins, purpose, methodology, and structure. This document outlines the essential elements to include when preparing documentation for a synthetic data collection, covering everything from data sources and generation techniques to quality assurance and structure.

We advise for the documentation accompanying synthetic data to consist of at least these five key sections, each summarising a critical aspect of synthetic data documentation:

1. **Introduction** – essential details about the dataset.
2. **Sources overview** – background information and references.
3. **Original purpose, applications, and limitations** – the rationale, use cases, and known constraints of the dataset.
4. **Methodology** – the techniques and tools used to generate and validate the data.
5. **Data structure and format** – a description of the dataset's structure, organisation, and format.

This structure ensures comprehensive and consistent documentation, providing users with the necessary information to understand, evaluate, and use synthetic data collections effectively.

1. Introduction

This section introduces the synthetic data collection and provides essential identifying information. It ensures users can quickly understand what the dataset is and who to contact for further information.

- A name for the synthetic data collection, ensuring synthetic is included in the title.
- A brief description of the dataset(s) being made available.

- Date when the synthetic data was generated.
- Name and contact information of the person or team responsible for the synthetic data (aiming for shared mailbox).

2. Sources overview

This section should provide a brief overview of the real dataset(s), metadata, research, and/or documentation used as references or inputs in generating the synthetic data. If real data was used, it should be cited appropriately. Any existing models or prior research that informed the generation process should also be cited. This should include the full citation including DOIs or other persistent identifiers wherever possible and access and use information.

Where synthetic data has been created from existing raw data, please include:

- Data types, number of observation units and variables in original data.
- Brief overview of original data, including main aims and topics.

3. Original purpose, applications and limitations

This section should aim to explain the intended purpose of the synthetic data and outlines its potential use cases. It provides transparency regarding any known limitations or inaccuracies and avoids misuse. The following questions should be addressed:

- What is the purpose of generating this synthetic data?
- Who can use the synthetic data, and for what purpose?
- What are the known limitations or inaccuracies in the synthetic data that may affect its usage or performance?

4. Methodology

This section describes the processes used to generate, validate, and assess the synthetic data. It is divided into two parts for clarity.

4.1 Generation

Details the techniques and tools used for generating the synthetic data, as well as steps to reproduce the dataset:

- Description of the technique(s) used to generate the synthetic data, including where applicable, any customisation made to standard techniques.
- Name and version of any software tools used for the synthetic data generation.
- Description of any methods used to confirm no matches with real data present in existing data (where applicable).
- Steps or instructions to reproduce the synthetic dataset, i.e. where the code is available, including the DOI/persistent identifier. Where code cannot be shared, an explanation should be provided.

NB: "random seeds" should be recorded and included in the documentation for full reproducibility.

4.2 Data quality and validation

Outlines the processes used to ensure data quality and validate its accuracy and usability:

- Description of any metrics used to assess the quality of the synthetic data.
- Distribution comparison, correlation metrics and statistical similarity.
- Explanation of the process used to validate the synthetic data.
- Any known issues in the synthetic data (e.g. outliers).

NB: this section might not apply to all data, i.e. very low fidelity where only manual checks are conducted; if this is the case, please explain why standardised quality control and validation were not necessary.

5. Data structure and format

This section defines the organisation and structure of the synthetic dataset, ensuring users can easily understand and integrate the data into their work.

NB: Variable names should be prefixed with synth_ to reduce potential misuse.

This section should include information about:

- Data types.
- Data format.
- Data volume.

Acknowledgements

This guidance document was made possible and inspired by the Evaluating the Benefits, Costs, and Utility of Synthetic Data project, hosted by the UK Data Service (UKDS). A key part of the project is advancing the understanding and application of synthetic data to promote innovation and data sharing. We acknowledge the funding support provided by the Economic and Social Research Council (ESRC) facilitated via ADR UK (Administrative Data Research UK).

We also extend our gratitude to key UKDS colleagues, specifically Cristina Magder, Dr. J. Kasmire, and Dr. Sharon Bolton, whose expertise and insights shaped this guidance. Their collective effort has ensured a robust and practical framework for documenting synthetic data collections, supporting transparency, reproducibility, and usability.