Metadata/README elements for synthetic structured data made with GenAI: Recommendations to data repositories to encourage transparent, reproducible, and responsible data sharing.

AUTHORS:
Ericka Johnson, Dept of Thematic Studies, Linköping University
David Rayner, Swedish National Data Service, University of Gothenburg
Jools Kasmire, Research Fellow, University of Manchester
Victor Hennetier, Dept of Thematic Studies, Linköping University
Saghi Hajisharif, Dept of Science & Technology, Linköping University
Helene Ström, Fair AI Data

Introduction

Publication of AI-generated synthetic structural data in data repositories is beginning to reveal the specific documentation elements that need to accompany synthetic datasets so as to ensure reproducibility and enable data reuse.

This document identifies actions that research repositories can take to encourage users to provide AI-generated synthetic datasets with appropriate structure and documentation. The recommendations are specifically for AI generated data, not (for example) data produced using pre-configured models or missing data created by statistical inference. Additionally, this document discusses metadata/README elements for synthetic *structured* datasets (tabular and multi-modal) and not textual data from LLMs or images for computer vision.

The document is the result of a workshop held on 23rd January 2025, with participants from the Swedish National Data Service, Linköping University and Manchester University. It also draws on survey responses about current practice from 17 data repositories and a review of existing metadata and README requirements.

Background

AI-generated synthetic structured datasets are generated using machine learning techniques with the aim of reproducing the essential elements of an existing dataset (Guépin et al., 2024 Jacobsen, 2023; Li et al., 2023; Offenhuber, 2024; Savage 2023). Synthetic data generation may be driven by the need to ensure privacy or to expand, enhance or substitute for real-world datasets which may be insufficient or non-existent. Sometimes synthetic data is produced to

create a portable or shareable dataset that is considered safe for open access, for example to share via a data repository.

While synthetic structured data may reproduce the essential elements of an original dataset, it will also inevitably introduce "intersectional hallucinations", which refer to anomalous inter-attribute relations within a dataset (Lee, Hajisharif & Johnson 2025). AI generated synthetic data also have a known tendency to minimize minority elements and amplify majority elements (Chen et al., 2024; Johnson & Hajisharif 2024). Thus, knowing in what ways a synthetic dataset demonstrates fidelities and in what ways it is 'different' from the original data is essential for successful and responsible re-use of synthetic data. Given that the goal of many data repositories is to provide access to data that is replicable and/or reusable, there is a clear need to establish protocols for documenting synthetic data.

Primary recommendations

Our **primary recommendations** are:

a) that data repositories establish a standardized way to label data as synthetic data, and that this information is prompted-for or required in the metadata or READMEs associated with synthetic datasets.
b) that data repositories provide users with a guide that explains how to properly document synthetic data. The extent to which documentation should be provided with the dataset or provided in associated articles or publications linked to the data will depend on the policies of the repository. An example is the guide provided by the Swedish National Data Service (2025).
c) that domain experts be prompted to document the context and motivation for generating synthetic data.

Documenting synthetic data – process and product

Reusability often refers to the data as a product. In the case of synthetic structured data, however, it may be the method of data generation (the data as a process) that is reusable, not the data itself. We therefore suggest that data repositories require information about both the process of data generation *and* details about the actual synthetic data.

Data as a process

The following elements should be included to describe the technical details of the synthetic structured data generation process:

- A description of the workflow.
- The generative model used (i.e. GAN, Diffusion, etc.). As techniques are constantly evolving, these requirements should be formulated in such a way to allow for and capture new techniques. The structure and hyperparameters (learning rate, number of epochs, etc.) of the generative model are also important factors for reproducibility and should be included.

- What raw data or inputs, if any, were used, including its mode of collection. A link to the source of the raw data should be provided where appropriate.
- Which (random) seeds were used.
- If a subset of raw data were reserved for testing, how was this subset selected?
- Versions of the software and packages used.
- Operating system information, values for relevant environment variables.
- A link to the source code (we suggest keeping code in a separate repository so it can be reviewed, improved, and re-released) and if appropriate a link to the weights of the trained model.
- Citation details (including DOIs) for related documents or the release versions of code.

Additionally, some cases of synthetic data are not based on raw data (e.g. agent based modeling/multi agent systems, digital twins). In such cases, this should also be clearly stated in the description of the data generation process.

If a repository considers that publishing the data generation model is out-of-scope, we suggest providing information on how models can be deposited in either a more generic repository or in a specific repository for models[1]. Links to the model can then be provided in the dataset metadata and/or README.

Product

Synthetic structured datasets inevitably contain stochastic variability, meaning that different datasets can be obtained by running the same code multiple times with different random seeds. We therefore suggest that metadata/READMEs also contain information about:

- whether the dataset is entirely synthetic or augmented. If it is augmented, what are the proportions of real and synthetic data?
- missing edge cases at the single-attribute level and inter-attribute level.
- inter-attribute combinations in the raw data that have diminished frequency in the synthetic data.
- inter-attribute hallucinations that have been observed in the synthetic data.
- details of the verification/validation process: how was the model tested, etc.
- how the synthetic data are structured at the file-level: are the input data in a folder marked "raw" or "input", and output in an "output" folder?

Privacy and specific circumstances

A common use-case for synthetic data is when privacy assurance is necessary for sensitive data. In such cases, we recommend the metadata/README contain information about disclosure risk, indication risk, reidentification risks, K-anonymity, etc. This type of synthetic data requires extra care and should only be made freely available if specific individuals cannot be re-identified by any reasonably likely means.

[1] For example, https://www.comses.net/

We also suggest that repositories include instructions on creating the metadata/README that will prompt domain experts to explain the specific circumstances of their synthetic data. Why was it generated? What is the fundamental hypothesis behind the synthetic dataset's use? What is its subject and purpose(s)? Data creators should be encouraged to disclose, for example, if the dataset was created for exploratory research, to represent sensitive data, to allow for work by a distributed team, to enable data portability, to create categories or support classification decisions, etc.. Encourage data submitters to consider sensitive areas and intersections within the data, as well as how many relational intersections are valuable to combine when using the dataset for new research purposes.

Summary and discussion

Synthetic structured data may be produced where scientific research requires data with no personal information, data that are portable and shareable, data which are not obtainable for practical or ethical reasons, or large datasets for machine learning. However, the details of the generation process and the variations inherent in synthetic data need to be documented, either in a dataset's metadata/README or in the articles accompanying the dataset.

Many aspects of synthetic data are still emerging, and in some cases, we lack established routines or even vocabularies for them. We hope the recommendations in this policy document will serve as a starting point for further discussions. In particular, we aim to encourage those working with data repositories to collectively establish best practices for managing synthetic data and developing vocabularies to describe them. For example, we might promote an accepted keyword or subheading, such as **SYNTHETIC_DATA**, or suggest appending **"_synth"** to filenames containing synthetic data. Additionally, controlled vocabularies should include subcategories to distinguish between fully synthetic and blended/augmented data.

With a well-defined vocabulary and clear metadata guidelines, repositories can help researchers to describe both their datasets and the processes used to create them in an open, transparent, and reproducible manner, ensuring responsible data sharing within the scientific community.

References

Chen W, Yang K, Yu Z, et al. (2024) A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review* 57(6): 137.

Guépin F, et al. (2024) Synthetic Is All You Need: Removing the Auxiliary Data Assumption for Membership Inference Attacks Against Synthetic Data. In: Katsikas, S., et al. Computer Security. ESORICS 2023 International Workshops. ESORICS 2023. *Lecture Notes in Computer Science, vol 14398.* Springer, Cham. https://doi.org/10.1007/978-3-031-54204-6_10.

Jacobsen BN (2023). Machine learning and the politics of synthetic data. *Big Data & Society*. 10(1).

Johnson E and Hajisharif S (2024) The intersectional hallucinations of synthetic data. *AI & Society*. https://doi.org/10.1007/s00146-024-02017-8.

Lee, Hajisharif & Johnson (2025) The ontological politics of synthetic data: normalities, outliers, and intersectional hallucinations. *Big Data & Society*.

Li X, Wang K, Gu X, Deng F, Wang FY (2023) Parallel eye pipeline: An effective method to synthesize images for improving the visual intelligence of intelligent vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. *53*(9), 5545-5556.

Offenhuber D (2024) Shapes and Frictions of Synthetic Data. *Big Data & Society.* 11 (2): 20539517241249390. https://doi.org/10.1177/20539517241249390.

Savage, N (2023) Synthetic data could be better than real data. *Nature Machine Intelligence. doi: https://doi.org/10.1038/d41586-023-01445-8.*

Swedish National Data Service. (2025). Managing and publishing synthetic research data (Version 1). *Zenodo. https://doi.org/10.5281/zenodo.14887525*