

International Journal of Population Data Science

Journal Website: www.ijpds.org



Synthetic data and clarifying the quagmire of terminology that surrounds it

Emily Oliver¹, Lora Frayling², Shah Suraj Bharat¹, Elizabeth Pattinson¹, Joshua Stock³, Fiona Lugg-Widger³, and Emma Gordon¹

¹ADR UK, London, United Kingdom

²HDI CIC, Cambridge, United Kingdom

³Cardiff University, Cardiff, United Kingdom

Objectives

Building on existing literature, our objective is to recommend definitions for key synthetic data terminology regarding its use cases in privacy preservation. By unpacking the words we use, we expose unsettled terminology and propose definitions which offer clarity and consistency, thus affording consensus to support increased adoption of synthetic data.

Methods

Taking into account the differing interests of creators, guardians and users of synthetic data, we applied a principled approach to selecting literature that could reveal insight on:

- The scope and purpose of synthetic data when used for privacy preservation
- Methods for characterisation and evaluation of it
- Clarity on quality control and regulations
- Guidelines for practitioners generating, distributing and using synthetic data.

We examined a range of academic and grey literature and, on finding a lack of definitions and consistency, dug deeper to explore the differences in terminology depending on type and usefulness for different use cases.

Results

Understanding the use case is key to creating accurate definitions. The terms 'utility' and 'fidelity' are separate but related concepts and feature commonly in literature, often interchangeably. Yet whilst 'utility' is dependent on context (is it right for this use case?), 'fidelity' is an inherent quality of the data. Different levels of fidelity can be important for some use cases, but increasing fidelity will not always increase utility. Furthermore, structural fidelity is different to statistical fidelity, yet this differentiation is rarely acknowledged. As such, terms related to fidelity remain ambiguous, while others,

such as 'microdata' are simply assumed by the author to be understood by the reader with no definitions provided.

Conclusion

Inconsistent and undefined use of terminology leads to miscommunication, and consequently a lack of reliability and efficiency. It is important to solve this through wide stakeholder engagement for purposes of good governance and public trust. We recommend a non-static glossary where contested terms are recognised and clarity is formed.



Temporary page!

$\LaTeX$ was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because $\LaTeX$ now knows how many pages to expect for this document.