UNECE High-level Group for the
Modernisation of Official Statistics

Business Case for Practical Guide to Synthetic Data

This business case was prepared by Kate Burnett-Isaacs, and is submitted to the HLG-MOS for their approval.

Type of Activity			
☒	New project	☐	New activity
☐	Extension of existing project	☐	Extension of existing activity
Projects are undertaken by separate project teams. Projects are expected to produce a significant contribution to achieving the HLG-MOS vision		*Activities are undertaken by Modernisation Groups. These activities produce smaller, more detailed outputs to help achieve the HLG-MOS vision*	
See here for more details: https://statswiki.unece.org/x/nwEzCw			

Purpose

Data has become a valuable commodity, providing information for statisticians, economists, and data scientists to generate more timely and granular insights. National statistical offices (NSOs) are striving to provide greater transparency and openness and so are looking to expand safely sharing of data, expertise and best practices both internally as well as with external partners. In addition, different types of users are increasingly searching for quality data sets to support testing, evaluation, education and development purposes. These aspects provide more value to users and bring the need to uphold data integrity and confidentiality to the forefront.

The demands for timely, integrated data compiled from ever-growing sources of increased complexity, along with the unequivocal commitment to trusted data protection call for a modernized, interoperable approach to mobilizing these large and complex data sources. Synthetic data can be a solution to providing rich data while respecting integrity and confidentiality imperatives.

Synthetic data can find its roots in edit and imputation methods, however synthetic data uses are becoming broader and increasing in complexity. New methods are emerging for generating and evaluating confidentiality of synthetic data, and more guidance is needed to maximize utility while ensuring confidentiality. Utility is seen as the value that a data set brings to a particular usage, for example, systems testing or model testing. Increasingly, utility also encompasses the desire that distributions or results of the synthetic data closely approximate those found in the real data. For example, extracting more detailed insights using increasingly big data sets requires new methods such as machine learning and modeling techniques. The integrity of these new methods requires that, as much as possible, data sets (both those from survey or non-survey sources) preserve the structure, characteristics and often the distributions of the original data as much as possible. However, the more closely the synthetic data set emulates the real data, the higher the risk that confidential information in the original data set could be disclosed. As governments become more open with their data and work, the confidentiality aspect remains a top priority. Once properly explored and understood, synthetic data can play an important role in the way that NSOs share data while maintaining public trust.

Synthetic data is a relatively new topic, particularly its use by NSOs. A better understanding of not only the theoretical methods of how to create synthetic data are needed, but an international consensus on practical applications and best practices is required for consistency, transparency and comparability within statistical agencies, and with users in academia and the private sector. Additionally, in order for synthetic data to be a viable option for NSOs to distribute and disseminate microdata, clear methods of communication on the utility and risk of using this type of data must be well available to stakeholders and users. In order for synthetic data to be used to its potential, an international consensus on practical methods and uses must be achieved.

Description of the project

Building on the success and network of the Blue Sky Thinking Network’s Working Group on Synthetic Data, ‘the practical guide to synthetic data’ project sets out to develop a hands-on guide for creating and using synthetic data primarily geared towards data protection and disclosure control. The target audience of this guide includes NSOs as well as their clients such as academia, the private sector and the general public. The guide will focus on how to use synthetic data in practical applications, considerations for implementation, and important aspects to share with users. This guide can serve as the foundation for future standards as synthetic data is more broadly adopted within NSOs and by their users.

The project will be divided into four work packages, with the scoping work already completed through the Working Group on Synthetic Data.

WP1: Use cases for synthetic data: The methods and measures of synthetic data are highly dependent on the reason for using synthetic data. The Working Group on Synthetic Data has identified four broad categories of uses for synthetic data: public data release, testing analysis, training and testing technology. This work package will detail the use case categories and highlight their specific methodology and measure requirements.

WP2: Recommended methods for creating synthetic data: This work package will gather an inventory of methods for creating synthetic data both currently in use and in research. The core outcome of this work package will be the assessment of the inventory of methods and recommendations on the methods suitable for different use cases of synthetic data.

WP3: Measuring the analytical value and/or disclosure risk of synthetic data sets: This work package will gather an inventory of utility and disclosure risk measures. This work will outline the circumstances in which the measures apply, associate them with the use case categories and provide clear explanations of what the measures signify and how they should be interpreted by those that create the synthetic data sets and those that use them.

WP4: Experimenting with the recommendations: Since the focus of this guide is on practical applications of synthetic data, this final work package will test the recommendations of the guide with real life scenarios of synthetic data. Activities will include pilot projects from NSOs, presentations geared towards user needs and hands-on events such as hackathons or training workshops.

Alternatives considered

The work to build consensus and understanding of creating and using synthetic data could continue as a Blue Sky Thinking Network (BSTN) working group, however with membership over25 participants, spanning 9 national statistics offices, academia and the private sector, this initiative has outgrown the scope of the BSTN. In addition, the target audience for synthetic data and the planned products are beyond the current methods of BSTN communication. A formal project would provide the proper scope, oversight and communications for the intended deliverables.

How does it relate to the HLG-MOS vision and other activities under the HLG-MOS?

The Synthetic Data project relates to all HLG-MOS visions and values by creating a collaborative initiative to promote sound methods and practices of synthetic data while upholding statistical integrity of those methods and the confidentiality of the data in question.

Proposed start and end dates

Start: January 2021	**End:** December 2021