Dienst Uitvoering Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap

Contact
synthetische.data@duo.nl

memo Privacy and Legal Aspects of Synthetic tabular data

Executive summary

Since 2022, DUO has been creating synthetic tabular data products. The Information Products and Compliance departments have collaborated to clarify and document the legal and privacy aspects of these products. In essence: 1) Synthetic tabular data enables the sharing of data that closely resembles the original data without disclosing personal information. 2) The privacy risks associated with synthetic tabular data are comparable to those of aggregated data.

Introduction

DUO (Dienst Uitvoering Onderwijs) is the Dutch education executive agency. As part of the Ministry of Education, Culture and Science, DUO is responsible for the execution of various education related policies and services.

This memo outlines DUO's objective regarding its privacy policy for synthetic tabular data. It has been drawn up after coordination between Compliance and the developers of synthetic tabular data within DUO. This memo discusses the definition of synthetic tabular data and describes how it can be assessed from a legal and privacy perspective.

What are synthetic tabular data?

Synthetic tabular data are artificially generated tabular data.[1] Tabular data are data represented in table format, often as a file or a single table within a database. The goal of synthetic tabular data is to enable providing data that closely resembles the original data without disclosing personal information. The potential applications of synthetic tabular data are broad, but the current focus areas within DUO are facilitating research and software testing. Synthetic tabular data can be generated to reflect the original data with varying level of representativity, using global statistics of the original data, machine learning techniques based on the underlying statistical relationships between variables, or a combination of both. It is important to distinguish synthetic tabular data from other techniques. Synthetic tabular data (as defined in this memo) are not pseudonymized data—regardless of the method of encryption—nor are they masked data, rule-based generated data, simulation-generated data, or data produced through random sampling.

Synthetic tabular data and the GDPR

Creating synthetic tabular data requires datasets which contain personal information. Therefore generating synthetic datasets means processing personal data and consequently falls within the scope of the General Data Protection Regulation (GDPR). This means that the production of a synthetic dataset must adhere to the principles of the GDPR. Processing must be necessary for a legitimate purpose and must have a valid lawful basis as outlined in Article 6(1) of the GDPR.

[1] Alvaro Figuera & Bruno Vas, Survey on Synthetic Data Generation, Evaluation Methods and GANs, 2022; César Augusto Fontanillo López en Abdullah Elbi, 'On the Legal Nature of Synthetic Data', 2022.

DUO collects and processes personal data about individuals to execute its public tasks ranging from financing primary public education to providing grants to students in higher education. This information is also used for scientific, historic or policy research. The use of personal data for these purposes constitutes what is known as further processing, meaning that the data is used for a purpose different from the original purpose of collection. DUO specifically has the public task to *provide and manage educational information to support policy.*[2] The production of synthetic tabular datasets falls within the scope of this public task and helps DUO comply with the principles of data minimisation and subsidiarity as not all (policy) research requires genuine personal data.

In certain cases external parties receive datasets from DUO to do their own research. Creation of synthetic tabular data in such cases conforms to further processing for scientific or historical research purposes or statistical purposes and is considered to be a compatible lawful processing operation.[3]

As the resulting fully synthesized dataset no longer contains personal information, it therefore falls outside the scope of the GDPR. *"Synthetic data is not real data about a person. (…), a single record in a synthetic dataset does not correspond to an individual or record in the original (real) dataset. (…) It does not include an identifier that corresponds to an actual natural person. It does not reference the physical, physiological, genetic, mental, economic, cultural, or social identity of an actual natural person. In short, a fully synthetic dataset does not meet the GDPR definition of "personal data.""*[4]

The European Commission Joint Research Centre also states that synthetic tabular data does not contain personal information: "*As opposed to real data, synthetic data are de-personalized (not personal) data, and therefore can be used in cases in which the target is not to identify a certain person."*[5]

If data is not fully or correctly synthetized synthetic tabular data may not be completely anonymous, resulting in a pseudonymous synthetic dataset[6]. For example, if multiple rows in the original dataset are linked to the same individual, the synthetic tabular dataset might replicate that individual's information. In such cases, the dataset would still contain personal data and must therefore be processed in accordance with the GDPR. To prevent this, DUO has developed a framework which specifies the definition and conditions regarding the use of synthetic data. The datasets synthesized by DUO under these conditions are ensured to contain only anonymous data. As a result, no 1-to-1 relationship can ever be derived.

[2] Organisatie- en mandaatbesluit OCW 2008, bijlage 1, hoofdstuk 9, https://wetten.overheid.nl/BWBR0023543/2022-10-15#Bijlage.

[3] Recital 50 GDPR.

[4] Khaled el Emam, Lucy Mosquera en Richard Hoptroff, Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data, First edition, Boston, 2020, ch. 6.

[5] European Commission Joint Research Centre, Multipurpose Synthetic Population for Policy Applications, LU: Publications Office, 2022, p. 17.

[6] "(…) according to the identifiability test as enshrined in the definition of personal data, synthetic data will be considered pseudonymous or anonymous data depending on the appropriateness of the data synthesis and the related ex-post control mechanisms." (López en Elbi 2022).

Finally, it is important to note that since synthetic tabular data no longer contains personal information, the GDPR does not restrict the sharing, publication, or distribution of synthetic datasets. This allows DUO's data to be used by a broader audience and for a wider range of purposes. Additionally, the European Commission highlights other possibilities, such as enhancing certain characteristics of synthetic datasets compared to their original counterparts.[7].

Comparison between synthetic tabular data and aggregated data
The residual risks of synthetic tabular data are comparable to those of aggregated data.

Aggregated data is generally classified as anonymous data under the GDPR, provided that the data cannot be traced back to individuals.[8] Aggregated data plays an important role in shaping government policies, and the GDPR recognizes that statistical analysis can serve the public interest. For example, statistical analysis conducted by National Statistical Offices - such as Statistics Netherlands (CBS) - is essential for evidence-based policymaking by the government. When statistical aggregated data is used, the rules for processing statistical data apply, which include, among others, standards for confidentiality and security.

Lopez & Elbi argue that aggregated data, where information cannot be traced back to an individual, should fall outside the GDPR.[9] The same principle applies to synthetic tabular data: while existing individuals contributed to the data in the synthetic dataset, they cannot be identified as such. Therefore, synthetic tabular data should be treated the same way as aggregated data.

A difference between synthetic tabular data and aggregated data is that synthetic data maintains a probabilistic relationship between individuals and the dataset. This means that neither outsiders nor the data owner can identify which specific individuals contributed to a given data point in the synthetic tabular data. For aggregated data, this only applies to outsiders.

When assessing privacy within synthetic tabular data, DUO applies its Statistical Disclosure Control framework.[10] Privacy is ensured by generating synthetic tabular data using state-of-the-art, scientifically validated techniques and by implementing control mechanisms on the generated dataset. This approach also acknowledges that as the analytical value increases (i.e., the more precisely the original data is replicated), the privacy impact also increases.

Treating synthetic tabular data similarly to aggregated data has several practical implications. The same confidentiality standards must be applied, particularly regarding group-level information, the irreversibility of the synthesis process, and the conditions for publishing algorithms. Additionally, any considerations made when publishing an aggregated dataset-such as its potential impact-should be applied in the same manner to the analog synthetic dataset.

[7] "(…) it enables the possibility to enhance certain characteristics that might not be cut by the real data, e.g. outliers, biases, etc." EC Joint research centre, 2022, p.17.

[8] Bart van der Sloot, Sascha van Schendel, en César Augusto Fontanillo López, 'The influence of (technical) developments on the concept of personal data in relation to the GDPR', Tilburg Institute for Law, Technology and Society (TILT), 2022, p.40.

[9] César Augusto Fontanillo López en Abdullah Elbi, 'On the Legal Nature of Synthetic Data', 2022.

[10] Kader Statistische Beveiliging.

Conclusion

When creating synthetic tabular data, personal data is processed as defined by the GDPR. This means that generating a synthetic dataset requires a purpose and a (legal) basis. However, synthetic tabular data itself (as defined in this memo) no longer falls under the GDPR once it is (correctly) generated. To ensure this, it is essential that there is no 1-to-1 relationship with real individuals and that the synthesis process must be irreversible. Even though synthetic tabular data are not subject to the GDPR, “confidentiality and security standards” still apply, like with aggregated data.