CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Omar Ali Fdal
Statice GmbH
Eisenacher Str. 1,
10777 Berlin
Germany

Paris, le 13 février 2023

N/Réf. : AGE/BPS/CS231015

Monsieur Ali Fdal,

Vous avez sollicité un avis de la Commission nationale de l'informatique et des libertés (CNIL) concernant les travaux menés par la société Statice sur l'évaluation des risques résiduels de réidentification pour des jeux de données synthétiques. Les méthodes soumises à la CNIL doivent faire l'objet d'une publication sous le nom d'Anonymeter dans le cadre du symposium « Privacy Enhancing Technologies » (PETs), qui se tiendra à Lausanne en 2023.

Premièrement, le service de l'expertise technologique de la Commission accueille favorablement l'approche de Statice consistant en une évaluation pratique des risques de réidentification se basant sur les critères d'*individualisation*, de *corrélation* et d'*inférence* définis dans l'avis 05/2014 du G29 concernant les techniques d'anonymisation. L'analyse de la documentation fournie n'a pas révélé d'éléments suggérant que les méthodes proposées ne pourraient pas permettre de vérifier le respect des trois critères susmentionnés dans le contexte de la production et de l'utilisation de jeux de données synthétiques.

Deuxièmement, le service de l'expertise technologique tient à saluer l'approche adoptée par Statice consistant à se soumettre au processus de publication académique et de revue par les pairs de ses travaux, ainsi que la publication du code source correspondant. En effet, l'évaluation indépendante, la large diffusion et l'amélioration continue des outils présentés sont le meilleur moyen d'assurer leur robustesse et leur adoption.

Troisièmement, bien qu'Anonymeter semble être un outil prometteur pour l'évaluation des risques résiduels de réidentification pour les ensembles de données synthétiques, le service de l'expertise technologique souligne que d'autres indicateurs issus de la littérature scientifique pourraient utilement lui être associés. À titre d'exemple, l'évaluation du risque de réidentification réalisée sur des exemples choisis (tels que les valeurs aberrantes) pourrait s'avérer complémentaire aux méthodes d'évaluation globale proposées par Statice.

Par ailleurs, les services de la Commission souhaitent attirer votre attention sur les points suivants :

- La génération d'un jeu de données synthétiques basé sur des données personnelles constitue un traitement de données à caractère personnelles. Il est donc nécessaire de s'assurer de sa conformité avec la réglementation en vigueur en matière de protection des données, à savoir le règlement général sur la protection des données (RGPD) dans l'Union européenne.

- L'anonymat d'un jeu de données synthétiques ne peut être déterminé qu'au cas par cas, c'est-à-dire pour chaque jeu de données généré, et ne doit donc pas être présumé à partir d'analyses effectuées sur d'autres jeux de données provenant du même fournisseur ou du même outil de synthèse de données.

- Les résultats produits par l'outil Anonymeter ont vocation à être utilisés par le responsable du traitement des données pour décider si les risques résiduels de réidentification sont acceptables ou non, et si l'ensemble de données peut être considéré comme anonyme. Cependant, en tant que fournisseur de solutions, la société Statice devrait également fournir des éléments pratiques décrivant précisément comment utiliser l'outil et interpréter les résultats obtenus (tels que des exemples, des tutoriels, des seuils, etc.).

- Les domaines scientifiques de l'anonymisation et de la réidentification des données sont en perpétuelle évolution. Les méthodologies existantes doivent donc être régulièrement réévaluées pour tenir compte des derniers développements.

En conclusion, le service de l'expertise technologique estime qu'Anonymeter est un outil prometteur et pertinent dans le contexte de la protection des données personnelles. Il recommande donc aux chercheurs et institutions de le tester dans divers contextes, afin de confirmer son utilité et sa fiabilité, et d'affiner les critères d'évaluation et les seuils acceptables.

Je vous prie d'agréer, Monsieur, l'expression de mes salutations distinguées.

Bertrand Pailhès,
Directeur des technologies et de l'innovation

Omar Ali Fdal
Statice GmbH
Eisenacher Str. 1,
10777 Berlin
Germany

Paris, February 13th, 2023

N/Réf.: AGE/BPS/CS231015

Dear Mr. Ali Fdal,

You have requested an opinion from the Commission nationale de l'informatique et des libertés (CNIL) concerning the work carried out by the company Statice on the evaluation of the residual risks of re-identification for synthetic datasets. The set of methods evaluated is to be published under the name *Anonymeter* in the upcoming Privacy Enhancing Technologies Symposium (PETs), which will take place in Lausanne in 2023.

First, the technology experts department of the Commission welcomes the approach consisting of a practical privacy risk assessment based on the *singling out*, *linkability* and *inference* criteria defined in the opinion 05/2014 on Anonymisation Techniques of the WP29. After analyzing the provided pieces of documentation, we have not identified any reason suggesting that the proposed set of methods could not allow to effectively evaluate the extent to which the aforementioned three criteria are fulfilled or not in the context of production and use of synthetic datasets.

Second, the technology experts department would like to commend the approach taken by Statice of going through the peer review process and academic publication of its work, as well as the publication of the corresponding source code. Indeed, the independent evaluation, wide dissemination and continuous improvement of the presented tools are the best way to ensure their robustness and adoption.

Third, while *Anonymeter* appears to be an efficient tool for the evaluation of the residual risks of re-identification for synthetic datasets, the technology experts department would like to point out that other indicators found in the scientific literature could usefully be associated to it. As an example, privacy risk assessment performed on well-chosen examples (such as outliers) could turn out to be complementary to the global assessment methods proposed by Statice.

Furthermore, the services of the Commission would like to draw your attention to several key points:

- The generation of a synthetic dataset based on personal data constitutes a processing of personal data. It is therefore necessary to ensure that it complies with the regulations in force on data protection, namely the General Data Protection Regulation (GDPR) in the European Union.
- The anonymity of a synthetic dataset can only be determined on a case-by-case basis, i.e. for each generated dataset, and should therefore not be assumed from analyses performed on other datasets coming from the same provider or data synthesis tool.
- The results produced by the tool *Anonymeter* should be used by the data controller to decide whether the residual risks of re-identification are acceptable or not, and whether the dataset could be considered anonymous. However, as a solution provider, the company Statice should also

provide practical elements describing precisely how to use the tool and interpret the results obtained (such as examples, tutorials, thresholds, etc.).

- The scientific fields of data anonymization and re-identification are in perpetual evolution. Existing methodologies should therefore be regularly reassessed to take into account the latest developments.

In conclusion, the technology experts department believes *Anonymeter* is a valuable tool, relevant in the context of personal data protection. They encourage further testing in various contexts by researchers and institutions, in order to confirm its usefulness and reliability, and to refine the assessment metrics and acceptable thresholds.

Yours sincerely,

Bertrand Pailhès,
Director of technology and innovation