Comments on EDPB Guidelines 02/2026 on Anonymisation

Submitted to the EDPB public consultation on Guidelines 02/2026 on Anonymisation (version 1.0, adopted 7 July 2026).

Submitter and scope

I am a data scientist based in Greece. I have built, and am bringing to market, a service that generates synthetic replicas of relational business databases for software development, testing and analytics, under the name Stable State Labs. These comments come from that practice: they concern how the draft Guidelines apply to synthetic data generated from personal data by statistical and machine-learning models.

I welcome the draft. In particular, paragraph 38 (anonymisation is itself processing), paragraph 41 (documentation of the process and its testing) and the three criteria in paragraphs 55 to 67 give practitioners a clear structure to design against. The comments below ask for six clarifications that would make the Guidelines easier to apply to synthetic data without lowering the standard.

Comment 1: synthetic data deserves its own short section

Synthetic data appears once in the draft, in paragraph 82, as an example of data that captures correlations rather than original records. That single mention understates how common the technique has become. Synthetic data is now a standard way for controllers to populate development, testing and analytics environments, and several national authorities have published material on it.

Synthetic data also differs from the techniques the draft discusses in one way that matters for the three criteria. Generalisation, suppression and noise addition transform existing records, so each output record still descends from one input record. A generative model instead learns the joint distribution of the source data and samples new records from it. There is no one-to-one correspondence between an output record and an input record. The risk is therefore not residual identifiers in a transformed record but memorisation: a model with too much capacity, or trained on too little data, can reproduce rare source records almost exactly.

I suggest a short dedicated section, or an annex, that (a) names synthetic data as a distinct technique, (b) states that the absence of one-to-one correspondence does not by itself satisfy any of the three criteria, and (c) identifies memorisation of rare records as the principal failure mode to test for.

Comment 2: the three criteria can be measured for synthetic data, and the draft could say how

Paragraphs 43 and 84 call for a technical analysis that reflects the state of the art in re-identification. For synthetic data, each of the three criteria has an established empirical test. Naming them, even as non-exhaustive examples, would help controllers and supervisory authorities compare assessments.

Criterion	What can be measured	What the result shows
No record isolation (para 55)	The share of synthetic records that match a source record on all quasi-identifiers, and the share whose quasi-identifier combination belongs to exactly one real individual	Whether any synthetic record singles out a real person
No linkage (paras 64 to 65)	The distance from each synthetic record to its nearest source record, compared with the distance from real records that were held out of training; membership-inference attacks, reported as the attacker's accuracy against the chance level	Whether synthetic records sit closer to the training individuals than an unseen person would, and whether an attacker can tell who was in the training data
No inference (para 67)	Attribute-inference attacks: given some attributes of a real individual, whether the synthetic data lets an attacker predict a sensitive attribute better than population statistics would	Whether the data supports specific inferences about particular individuals

One methodological point would be worth stating explicitly. The linkage and inference tests only have meaning against a baseline. The natural baseline is a holdout set: real records excluded from training, so that the test can ask whether synthetic records are closer to training individuals than to unseen ones. A controller that trains on every record cannot run this comparison. I suggest the Guidelines note that withholding part of the source data from the generation step is a precondition for a credible assessment, not a deviation from it.

Comment 3: clarify "specific and meaningful" inference for generative models

Paragraph 67 defines a meaningful inference as one that, among other conditions, "could not be obtained from general knowledge or from data about the population at large." Paragraph 82 then warns that inferences can be elicited by querying AI models or synthetic data. Read together, these two paragraphs leave a gap that will cause inconsistent assessments.

Synthetic data is useful precisely because it preserves population-level patterns: that customers in one region buy more of one product, or that account balance correlates with age. Those are inferences "about the population at large" and paragraph 67 excludes them. But a reader who starts from paragraph 82 may conclude that any inference drawn from synthetic data is suspect, because the data was built to support inference.

I suggest paragraph 82, or the section proposed in Comment 1, state the distinction directly: a synthetic dataset fails the No Inference criterion when it lets a party learn something about a particular individual that the party could not learn from the population statistics alone. Preserving population-level correlations is the purpose of the technique and is not in itself a failure. The test is whether the dataset reveals an individual-level signal that the aggregate does not, which is what attribute-inference attacks measure.

Two further details would help. First, whether the assessment should assume an attacker who already holds some true attributes of the target individual, since attribute inference is only possible on that assumption. Second, how the criterion applies to outliers: an individual who is unique in the population may be inferable from any faithful statistical summary, synthetic or not, and the Guidelines could say whether that situation is addressed by suppressing outliers before generation or by accepting a residual risk.

Comment 4: minimum content of the documentation record

Paragraph 41 asks controllers to keep records of the anonymisation process, including testing, that show the process is effective. This is the most useful operational requirement in the draft, and it would be stronger with a minimum list of contents. Without one, two controllers using the same technique will document it in incomparable ways, and a supervisory authority will have no fixed point to check against.

For synthetic data, I suggest the record include at least: the source schema and its fingerprint; the generation method and its version; the share of records withheld as a holdout and how they were chosen; each test run, with the threshold applied and the result; the date of the assessment; and the identity of the person who performed it. A record of this kind can be produced automatically by the generation tooling at the time of the run, which makes the requirement cheap to meet and hard to backdate.

Comment 5: the trained generator is itself an artefact to reassess or destroy

Paragraph 35 notes that re-identification likelihood rises over time and recommends periodic reassessment of anonymised data. For synthetic data there is a second artefact that the draft does not mention: the trained generative model. The model is what was fitted to the personal data, and a model that has memorised rare records can reproduce them on request even when the published dataset passed every test.

I suggest the Guidelines state that where a generative model is used, the model falls within the scope of the assessment and of any retention decision: either it is destroyed once the synthetic dataset is produced, or it is protected and reassessed in the same way as the dataset. Paragraph 82 already recognises that querying a model can elicit inferences; the retention consequence follows from that.

Comment 6: the simplified approach and smaller controllers

Paragraphs 45 to 49 offer a contextual approach, which assesses each recipient's capabilities, and a simplified approach, which assumes a uniform level of protection and errs toward treating data as personal. Paragraph 48 rightly says the simplified approach is not a different legal standard.

Most controllers that use synthetic data for development and testing are small or mid-sized. They share the data with internal teams and with a changing set of external developers, and they do not have the resources to assess each recipient. For them the simplified approach is the only realistic one. It would help if the Guidelines said so, and added that fixed, documented thresholds applied uniformly across recipients are an acceptable way to implement it, provided the thresholds are set conservatively and recorded under paragraph 41. That would give smaller controllers a proportionate route to compliance that supervisory authorities can still verify.

Closing

The draft gives synthetic data practitioners a workable framework for the first time. The six requests above are all for clarity rather than for a lower bar: a dedicated section on synthetic data, example tests for each criterion, a precise reading of inference for generative models, a minimum documentation list, treatment of the trained model as an artefact, and a proportionate route for smaller controllers. I would be glad to provide further technical detail on any of them.

Dimitris Vakkas, Data scientist, Stable State Labs, Greece, dimitris@stablestatelabs.com