Data generated from rules has no source record: a category Guidelines 02/2026 does not yet cover

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

From Axolile Lungu, a qualified chartered accountant, dual citizen of South Africa and Ireland, founder of synthetic cfo · 23 October 2026

1 Who I am and why this matters to me

synthetic cfo generates synthetic ERP company ledgers for testing fraud detection, audit analytics and AI models. The ledgers follow the table structures of SAP ECC, SAP S/4HANA and Oracle Cloud; the ERP is a simulation, not affiliated with SAP or Oracle. Every figure is generated from double-entry accounting rules and published statutory rules. No real company's data is used at any stage, and no language model touches the numbers.

The output looks exactly like personal data. A generated company has employees with pay, tax withheld and a position; since engine 1.15.0 (20 October 2026) payroll withholds what the 2025 rules of each of ten jurisdictions call for. It has customers and vendors with invoices and payments. None of these people exist. I welcome the guidelines and support their framework. I write because the framework does not say how it applies to data like this, and the people who build and use such data need to know.

2 Summary

1. The draft is built around a transformation: an original dataset of personal data, a process applied to it, and the resulting "given data". Paragraphs 10, 70, 71(b) and 79, footnotes 1 and 43, and the treatment of synthetic data in paragraph 82 all assume that a source record exists.
2. Data generated from rules has no source record. No information relating to any individual is processed at any stage. It is neither anonymised data nor synthetic data in the sense of paragraph 82.
3. Applied mechanically, the three criteria give the wrong signal for such data. The record of an invented employee, with its own position, pay and tax, will usually be unique, so the No Record Isolation criterion (paragraph 55) reads as violated, although no natural person stands behind the record.
4. I propose four short edits and one example (section 5), and I set out where such data can still become personal data (section 4), so that the clarification does not become a loophole.

3 Where the draft assumes a source record

- **Paragraph 10** defines anonymisation as "any processing which successfully produces anonymous data under the GDPR, whether or not the original data is deleted". Footnote 1 recalls Recital 26: information "which starts as personal data" can later become anonymous.
- **Paragraph 70** defines the original dataset as "the information that was used to generate the given data", and makes a meaningful inference turn on a connection to it.
- **Paragraph 71(b) and footnote 43** test reliance by comparing A(D), produced from the original dataset D, with A(D'), produced from D without the individual's records.
- **Paragraph 79** describes membership inference as inferring that "an individual's data was included in the original dataset".
- **Paragraph 82** addresses synthetic data only as data "aggregated to represent correlations between elements of the original information".

These are the right tools when D contains personal data. When the information used to generate the data contains no record relating to any individual, D has no members. A(D) and A(D') are identical for every person, membership inference has nothing to find, and the paragraph 82 risk of eliciting information about a particular individual has no

source to draw on. The tests pass trivially, which shows they were not written for this case. The guidelines are silent on it, and silence invites two errors: treating rule-generated data as personal because a criterion is "violated", or describing it as "anonymised", a label that implies personal data was processed when none was (paragraph 40 asks for no ambiguous or misleading statements).

4 The category, and where it stops

What I mean. Data generated where every input is one of: (a) rules, such as double-entry bookkeeping and the posting logic of each document type; (b) published parameters, such as statutory withholding rates; (c) a configuration, such as industry, jurisdiction, ERP, fraud level and volume; and (d) a random seed. None of these is information relating to an individual. Any inference such data supports comes from general knowledge, which paragraphs 67 and 70 and footnote 42 already treat as not meaningful. The individuals in the data are invented, so they fall outside the glossary's "individual", which is a natural person.

Why it can be verified. Where generation is deterministic, the provenance claim can be checked rather than trusted. The synthetic cfo public sample is one fiscal year of an invented manufacturer in Germany on SAP ECC: 18 tables holding 12,028 rows, built from seed 20260801 on engine v1.15.0. Given the same seed, configuration and engine version, the engine rebuilds its database and every CSV byte for byte. A third party with access to the engine can therefore confirm that nothing but the stated inputs went in. That is stronger evidence than documentation alone (paragraph 41), and a transformation of personal data cannot offer it.

Where it stops. The category should be narrow, and the guidelines should say so:

- **Calibration on personal data.** If any parameter is fitted to record-level personal data, such as a real payroll or ledger, the generator learns from an original dataset. Paragraph 82 and the full framework then apply.
- **Coincidence and use.** An invented name or attribute combination may match a real person. A coincidental match is not information about that person by reason of content (paragraph 16(a)). But if the data is used to evaluate or treat a real person, it relates to that person by purpose or effect (paragraph 16(b) and (c)), and the guidelines apply from that perspective.
- **Mixing.** Where generated records are combined with real ones, paragraph 36 applies to the combined dataset as it stands.

5 Proposed changes

1. **New paragraph after paragraph 6:** "Data generated without processing any information relating to an individual, for example data produced only from rules, published parameters, a configuration and a random seed, does not relate to a natural person by reason of its content. Its generation is not anonymisation within the meaning of paragraph 10, and paragraphs 37 to 41 do not apply to it. It may nevertheless relate to a natural person by reason of its purpose or effect (paragraph 16(b) and (c)) where it is used to evaluate or treat a real individual."
2. **Paragraph 82:** specify that it concerns synthetic data produced from an original dataset containing personal data, and add: "Where the information used to generate the given data contains no information relating to individuals, the assessment should focus on the inputs to generation (see paragraph 51) rather than on the three criteria."
3. **Paragraph 51 (data mapping):** add the inputs to generation to the mapping step. Where any input is personal data, or a parameter fitted to personal data, the full framework applies.
4. **Paragraph 41 (documentation):** recognise that where a deterministic process regenerates the given data exactly from documented inputs, this regeneration is strong evidence of what was, and was not, processed.
5. **A new example in section 3.4.1:** "A software vendor generates the ledger of a fictitious company, including payroll for invented employees, from accounting rules, published withholding rules, a configuration and a seed. Each employee record is unique, so the No Record Isolation criterion appears to be violated. However, no record relates to a natural person: the data is not personal data, unless it is used to evaluate or treat a real individual."

I would be glad to share the public sample, its configuration and the regeneration steps with the EDPB Secretariat or any supervisory authority that wishes to test these points.

Axolile Lungu · founder of synthetic cfo · app.syntheticcfo.com