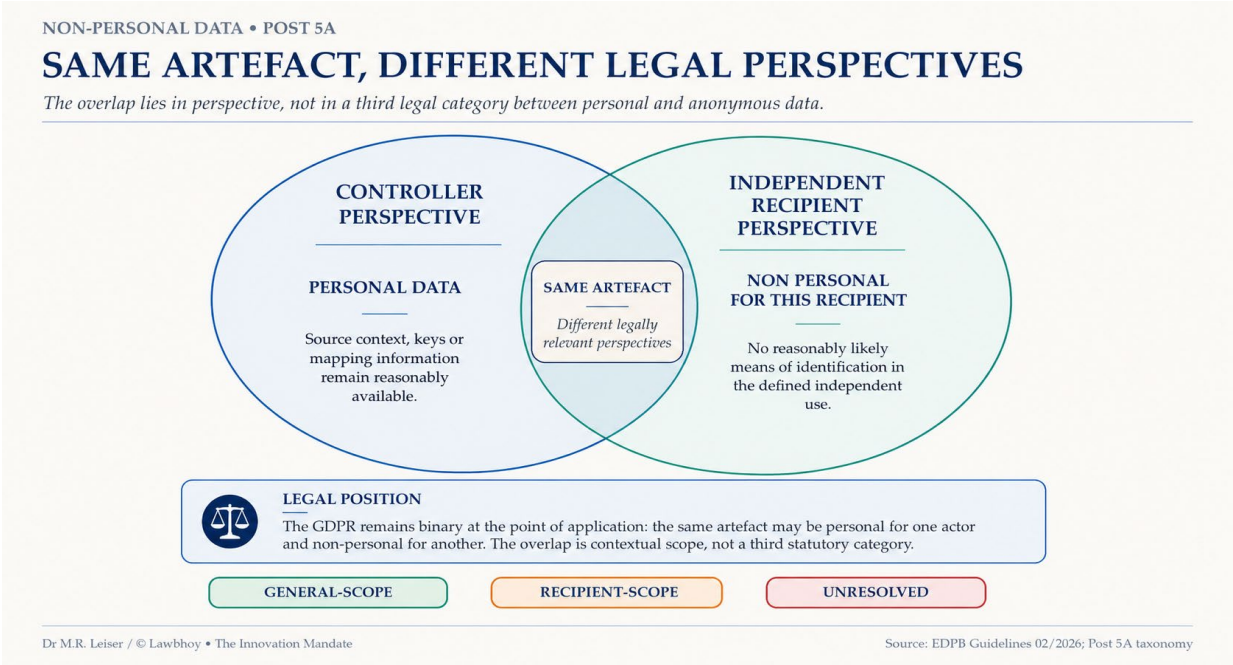
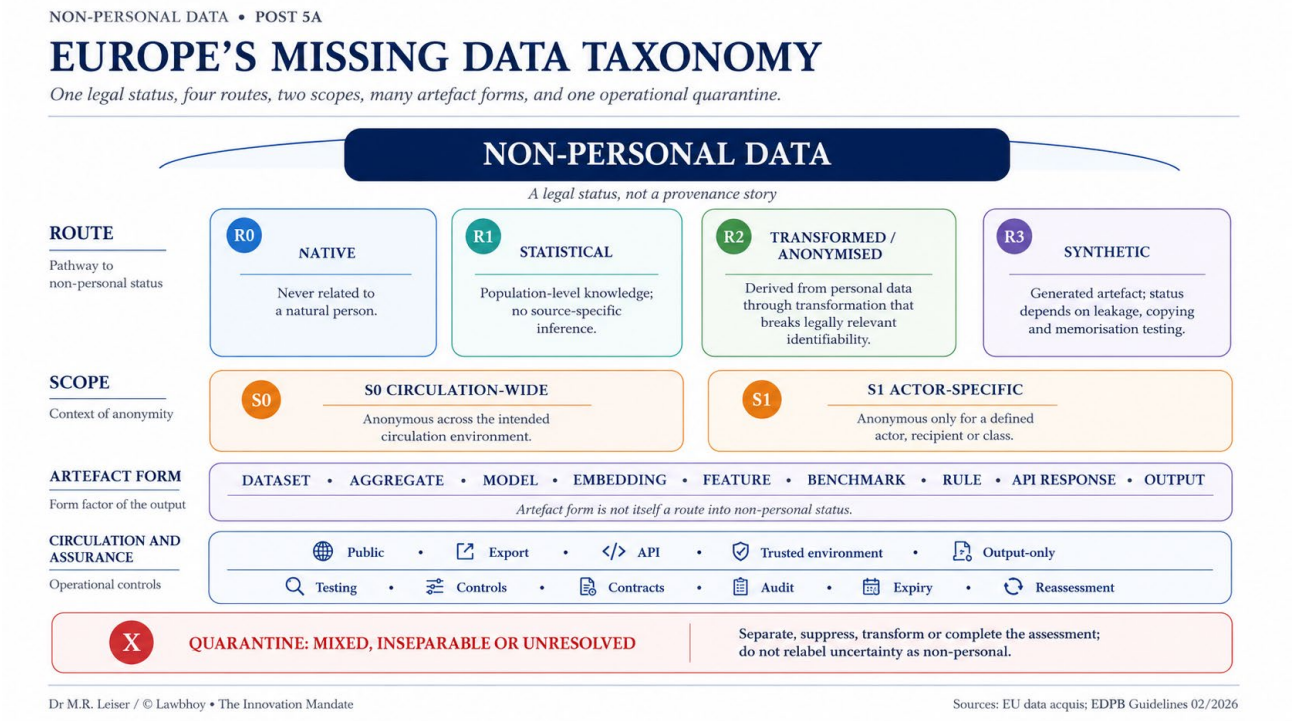
ANONYMISATION AFTER SRB • SUBSTACK POST 5A

Non-Personal Data Is Not One Thing

Europe's digital acquis needs a taxonomy that separates legal status from route, scope, artefact form, circulation and assurance.

A PDF of Part A of my final post in my series on the EDPB's anonymisation guidelines and the EU's Digital Omnibus. The paper argues that "non-personal data" is a legal umbrella rather than a single coherent object, and that Europe now needs a clearer taxonomy capable of distinguishing native, statistical, anonymised, synthetic and derived artefacts without collapsing uncertainty into a convenient label.





Dr M.R. Leiser (Mark)

Substack (LawBhoy) • Riga Graduate School of Law • DigiData Consulting Ltd • Innovation Mandate

Non-Personal Data Is Not One Thing

Europe has built an entire data economy on a negative definition; the EDPB has supplied the missing hinge, but the Omnibus still needs a taxonomy.

OK, I know I said I was going to limit myself to writing only five posts, but this one was so big I had to split it into two. This is Post 5A, the first of two parts, to my final post in my series on the EDPB's Guidelines 02/2026 on Anonymisation and their consequences for the EU's Digital Omnibus. The first post set out what the guidelines get right and where they remain operationally thin; the second defended contextual anonymisation against the legal panic that followed EDPS v SRB; the third examined the unequal consequences for research, SMEs and AI development; and the fourth develops the institutional settlement that the Omnibus should now pursue. This final post addresses the problem beneath all four: Europe repeatedly uses the expression "non-personal data" as though it described a coherent legal object, when in fact it is a residual umbrella covering data with very different origins, risk profiles, access conditions and evidential foundations. The argument developed here is that the Union now needs a cross-acquis taxonomy, accompanied by a practical Data Status Passport, that makes every claim of non-personal status answerable to the same structured questions.

Guidelines 02/2026 provide the immediate doctrinal reason to stop treating non-personal data as an undifferentiated remainder. By asking for whom data is intended to be anonymous, requiring the assessor to identify the applicable perspective and the means reasonably likely to be used. Distinguishing population-level learning from specific and meaningful inference, the EDPB has made it clear that a conclusion of anonymity is neither an intrinsic property of a file nor a label that survives every change in actor, role, or circulation environment. The same framework also shows why the taxonomy cannot end with a binary status marker: a defensible claim must record how non-personal status was reached, whether it applies across a circulation environment or only to a defined independent recipient, which artefact carries the claim, and which tests, controls and reassessment triggers sustain it. Post 5A develops that classificatory language; Post 5B then operationalises it through a TRACE Data Status Passport, supported by an Anonymisation and Utility Impact Assessment, so that "anonymous" becomes a bounded, evidenced and reviewable claim rather than a sticker detached from the conditions that made it true.

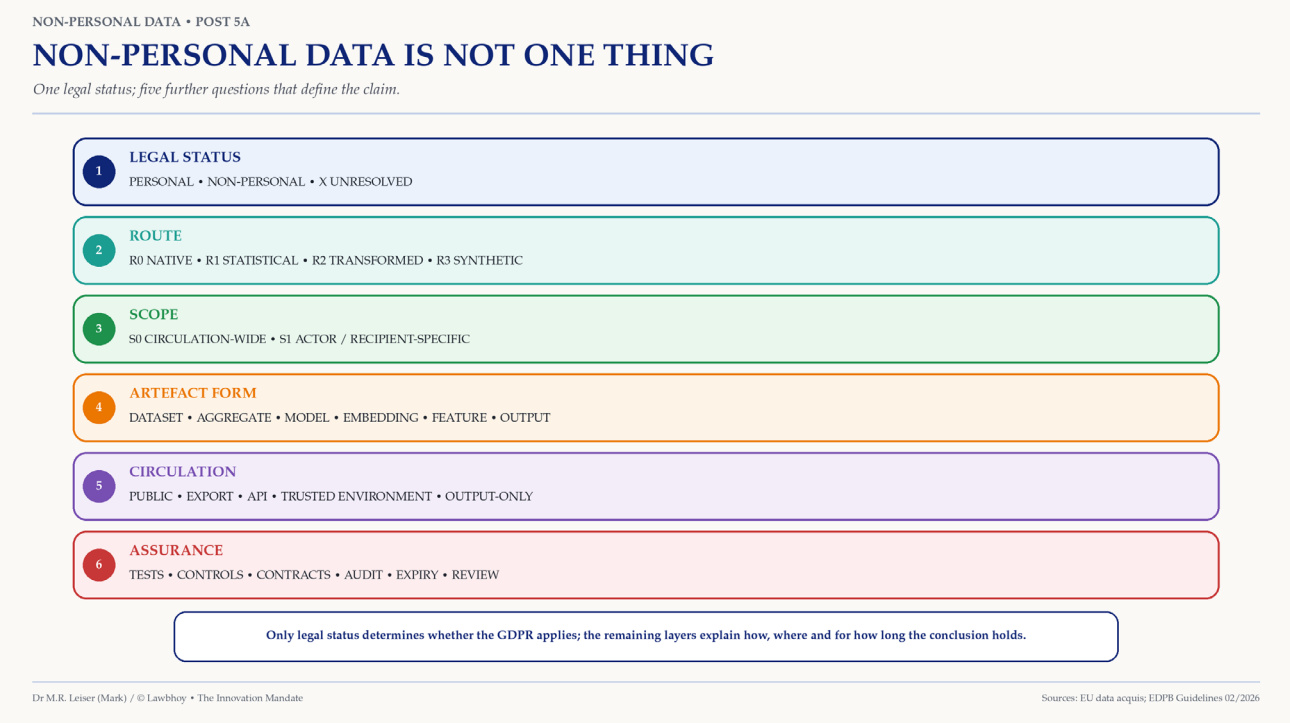
European data law has built an economic programme around a negative definition. The Free Flow of Non-Personal Data Regulation, the Data Governance Act, the Data Act and the AI Act all rely on the distinction between personal and non-personal data. In contrast, the European strategy for data assumes that information can circulate through common European data spaces and support research, competition and artificial intelligence. Yet the positive content of the second category remains remarkably thin because legislation generally defines non-personal data as data other than personal data. That formulation serves as a jurisdictional switch, but it does not explain what lies on the other side of the switch, how the artefact arrived there, whether the conclusion holds for every recipient or only for one, or what evidence must sustain the conclusion over time.[123]

The emptiness becomes visible as soon as ordinary examples are placed beside one another. A turbine's wind-speed reading may never have related to a natural person; a national unemployment rate may have been generated from individual records but no longer reveal anything source-specific about a worker; a transformed clinical table may remain personal for a hospital that retains identifying material while being anonymous for an independent research institute; a synthetic dataset may contain no copied record yet still memorise a rare patient; and a model, embedding or benchmark may preserve enough source-specific signal to support membership inference or extraction. All these artefacts may qualify as non-personal in the right

circumstances, although they reach that status through different routes and carry different technical, legal and institutional risks.

Guidelines 02/2026 supply the missing legal hinge by accepting that anonymity may vary from one entity to another. Their core test asks whether information relates to a natural person and, if so, whether that person is identified or identifiable from the applicable perspective through means reasonably likely to be used. If either answer is negative, the information is anonymous from that perspective. The Board therefore treats contextuality as part of the legal standard rather than as a discretionary relaxation, while preserving the objective character of the inquiry through factors such as access, cost, time, auxiliary information, legal powers, technology and foreseeable developments.[4]

The Status Stack



Legal status answers whether the GDPR applies, while route, scope, artefact form, circulation and assurance define the claim.

A residual category carrying too much weight

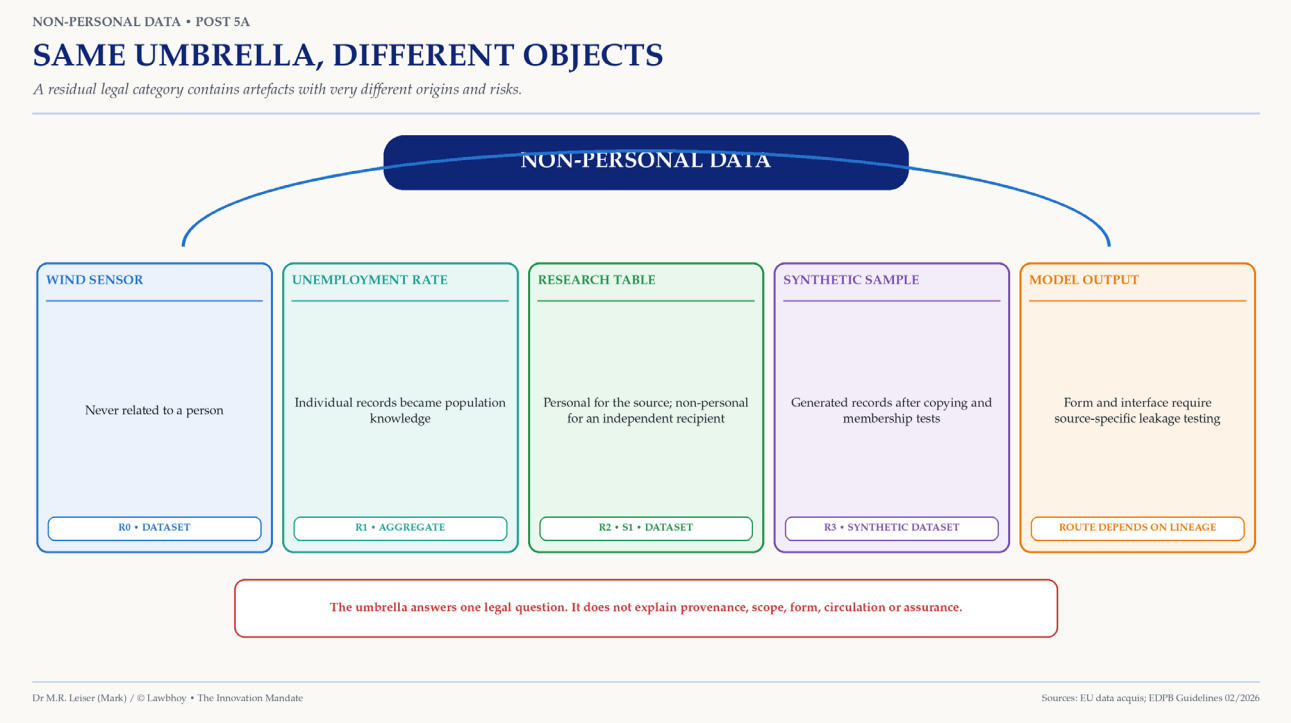
The historical starting point was a boundary around data protection rather than a positive law of non-personal data. Article 4(1) GDPR defines personal data broadly. At the same time, Recital 26 excludes information that does not relate to an identified or identifiable natural person and directs attention to the means reasonably likely to be used by the controller or another person. The binary legal question remains essential because it determines whether the GDPR applies, but the legislation says little about the internal structure of the category that remains outside its scope. That omission was tolerable when anonymous information appeared mainly as an exception at the edge of data-protection law; it became much harder to defend once the Union began treating non-personal data as an economic resource in its own right.[5]

Statistical disclosure control had already shown why the residual category required more structure. Suppression, generalisation, sampling, perturbation, and aggregation were used to preserve useful information while reducing the chance that a person represented in the source data could be identified. Latanya Sweeney's work on k-anonymity formalised part of that problem by requiring each released record to share specified quasi-identifiers with at least k-1 others. At the same time, later linkage research demonstrated that apparently innocuous combinations could become identifying when matched against

external information.[67] Those contributions did not prove that anonymisation was impossible; they showed that a claim of anonymity must account for the structure of the artefact, the availability of auxiliary data and the environment in which access occurs.

The legal literature developed the same point from several directions. Paul Ohm warned against treating anonymisation as a promise of irreversible safety, while Ira Rubinstein and Woodrow Hartzog argued for a disciplined risk-reduction process rather than an impossible guarantee. Nadezhda Purtova showed how an ever-expanding conception of personal data could turn European data protection into a law of everything. In contrast, Michèle Finck and Frank Pallas, followed by Inge Graef, Raphaël Gellert and Martin Husovec, exposed the instability of a binary vocabulary when technical capacity, data combination and institutional relationships change over time.[8][9][10] The sensible response does not abolish the legal boundary, because the GDPR still requires a threshold, but it does demand a more disciplined account of how non-personal status is reached, scoped and maintained.

Same Umbrella, Different Objects



A residual legal category contains native data, statistics, transformed research data, synthetic samples, and model outputs, each with different origins and risks.

Contextuality supplies the missing hinge

The Court of Justice gradually supplied the relational analysis that the legislation left implicit. *Breyer* treated a dynamic IP address as personal for an online media service where legal channels made identifying information from an internet provider reasonably available. *Scania* and *IAB Europe* confirmed that identifiability may depend on information or methods distributed across several actors. At the same time, *OC v Commission* required the relevant audience to include investigative journalists because the press release at issue was designed for public dissemination.[11][12][13][14] *EDPS v SRB* then held together two propositions that are often separated too quickly: pseudonymised information is not necessarily personal for every person in every circumstance, yet the particular legal obligation under examination may fix the applicable perspective.[15]

The EDPB converts that jurisprudence into a structured framework by asking for whom the data is intended to be anonymous, identifying the relevant entities and applicable perspectives, and then testing record

isolation, linkage and inference against the means reasonably likely to be used in that context. The analysis remains objective because the recipient cannot create non-personal status by declaration, contract or indifference. The outcome depends on evidenced relationships between actors, access to the artefact, access to auxiliary information, realistic chains of cooperation and the capacity to distinguish and treat a person differently. Contextuality therefore does not mean subjectivity; it means that identifiability is relational rather than intrinsic.[4]

The same framework also limits the practical reach of actor-specific conclusions. A processor does not acquire an independent anonymity perspective when it acts on the controller's instructions, and the controller can identify the individuals, because the controller's perspective follows the processing. A public release requires a much broader assessment than a transfer to a named independent research institute. At the same time, a recipient-specific conclusion cannot survive an onward transfer that introduces new identifying means. Contractual restrictions may reduce the likelihood that particular means will be used. Still, the EDPB treats them as evidence within the assessment rather than as a legal alchemy that turns personal data into anonymous information.[4]

The No Inference analysis adds an equally important distinction between population-level knowledge and source-specific connection. A historical dataset may reveal a general relationship between attributes and default risk. That relationship may later support an inference about a new applicant who never appeared in the source data. The new inference can become personal data about the applicant without converting the historical dataset back into personal data. By contrast, an inference that depends on the inclusion, behaviour or characteristics of someone represented in the source material preserves a connection to that person and may defeat the claim of anonymity.[4]

One status, five descriptive layers

A workable taxonomy should preserve the binary legal conclusion while separating five further questions that European policy repeatedly collapses into one label. Legal status asks whether the artefact is personal or non-personal for the actor and operation under examination. Route identifies the legal and technical path through which non-personal status was reached. Scope identifies the actors and circulation environment for which the conclusion holds. Artefact form records whether the object is a dataset, aggregate, model, embedding, feature, benchmark, rule or output. Circulation describes how the artefact may be moved or queried. Assurance records the testing, controls, contracts, audit, expiry and reassessment that make the claim defensible.

Only legal status determines whether the GDPR applies to the artefact for the actor in question, although the other layers supply the facts and evidence that support the conclusion. Pseudonymisation remains a safeguard applied to personal data; encryption controls intelligibility and access but remains reversible; aggregation describes form but may still expose outliers or support reconstruction; synthetic describes generation rather than privacy; and a clean room describes an access environment rather than the legal status of the data inside it. These concepts remain useful when they serve their proper functions, but each becomes a false friend when reduced to a status label.

The status stack also prevents the opposite error, in which personal data lineage permanently contaminates every downstream artefact. National statistics do not remain personal merely because tax records contributed to it. In contrast, a model does not necessarily remain personal merely because personal data appeared in the training process. Lineage creates a reason to test, document and monitor; it does not answer the legal question by itself. Treating lineage as an irreversible taint would reproduce the law-of-everything problem and render the AI Act's preference for anonymised, synthetic, or other non-personal alternatives largely meaningless.[310]

Four routes into non-personal status

Four routes are relevant to a claim of non-personal status: R0 (native non-personal data), R1 (statistical or population-level non-personal data), R2 (transformed or anonymised data derived from personal data) and R3 (synthetic non-personal data). Each route records how the conclusion was reached; it does not determine the scope for which that conclusion holds.

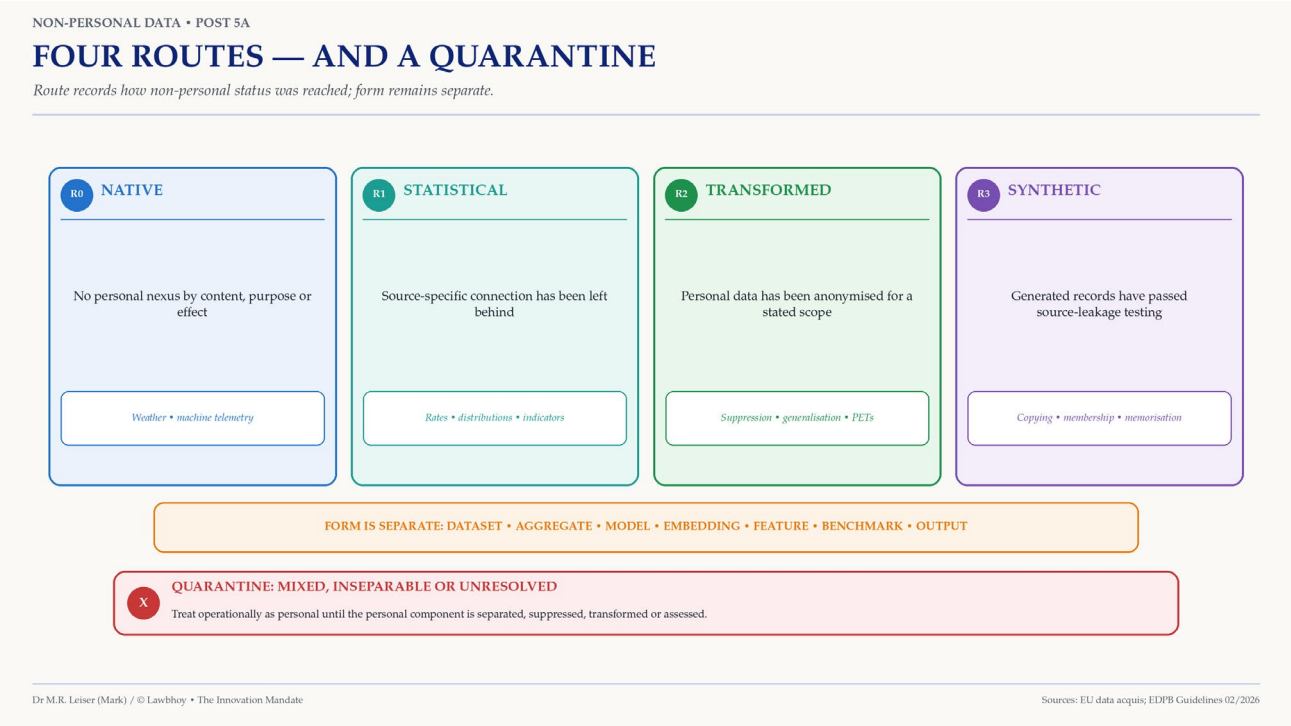
R0 covers native non-personal data: information that never related to an identified or identifiable natural person by content, purpose or effect. Weather readings from remote sensors, machine vibration measurements, and many forms of industrial telemetry may fall under this route. However, a machine source does not guarantee the absence of a personal nexus. Technical data can become personal when an organisation uses it to evaluate a worker, driver, household or device user, so R0 records the absence of that relationship in the operation under examination rather than the mere fact that a machine generated the information.

R1 covers statistical or population-level non-personal data produced through generalisation or aggregation. Individual records may have contributed to an average, rate, distribution, coefficient, indicator, or generalised finding. Still, the source-specific connection has been lost, and no specific and meaningful inference can be drawn about a person represented in the original data. The route therefore requires attention to small cells, outliers, overlapping tables, repeated queries and reconstruction attacks, because an aggregate becomes personal where the analytical form still reveals or reconstructs information about an identifiable individual. The EDPB's distinction between general learning and source-specific inference provides the central legal test.[4]

R2 covers personal data that has been transformed or anonymised for a stated scope. Personal data has undergone suppression, generalisation, perturbation, aggregation, deletion, privacy-preserving computation, or other transformations until the likelihood of identification becomes insignificant in practice within the stated scope. The route deliberately does not specify whether the conclusion holds throughout a circulation environment or only for a named independent recipient, because scope belongs on a separate axis. The same transformed table may therefore be personal for the source controller, non-personal for an independent research institute and unsuitable for public release.

R3 covers synthetic non-personal data and treats generation as provenance rather than proof. The artefact contains generated rather than directly observed records, but a high-fidelity generator may reproduce rare sequences, expose membership, memorise unusual examples or permit source-specific inference. R3 therefore requires method transparency and tests for copying, membership, memorisation, extraction and meaningful inference, with the result attached to the artefact version and access interface rather than to a product family. A synthetic label should open the privacy analysis rather than conclude it.

Four Routes and the Quarantine



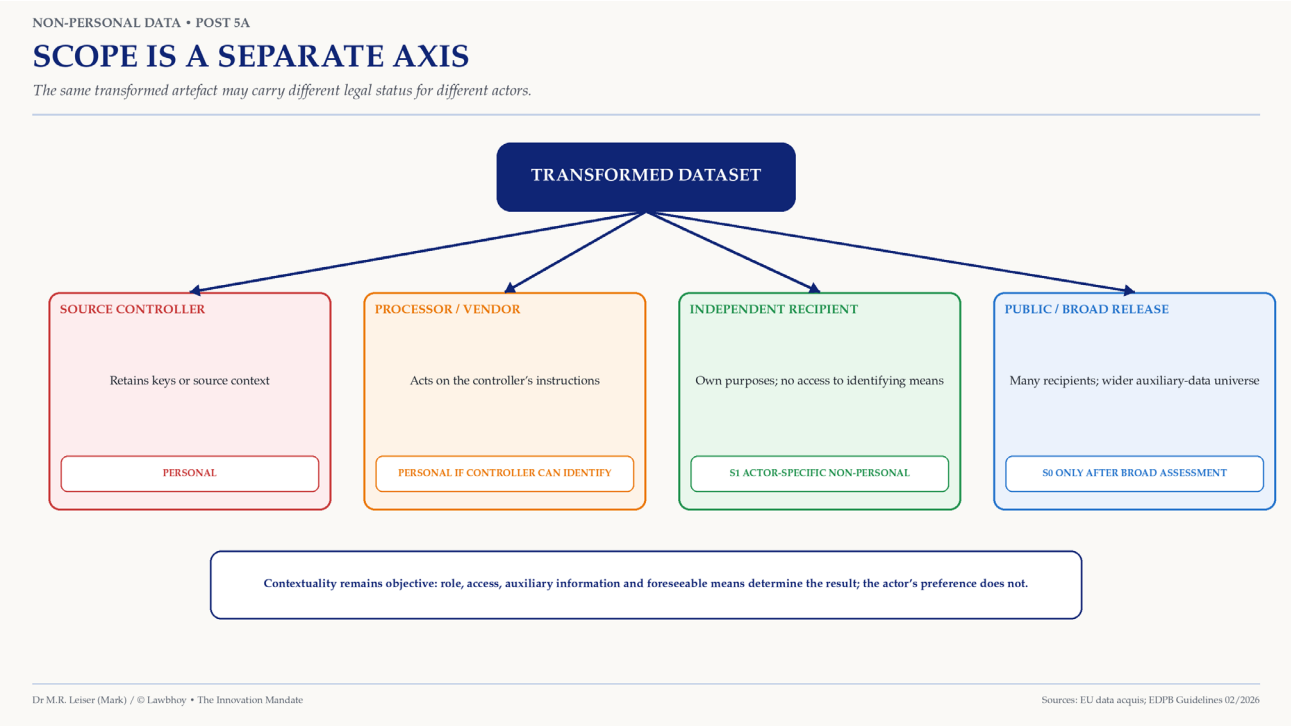
R0 native, R1 statistical, R2 transformed, and R3 synthetic describe routes into non-personal status, while X quarantines unresolved material.

Scope must travel with the claim.

Separately from route, scope operates at two levels: S0 (a circulation-wide finding within a defined release environment) and S1 (an actor-specific or recipient-class finding). Scope records for whom, and under which access and onward-transfer conditions, the conclusion holds. S0 does not mean anonymous to every conceivable person for all time; it means that the assessment has covered the actors, access conditions and auxiliary information reasonably associated with the stated circulation environment. S1 identifies the named recipient or class, the role analysis, the access modality and the onward-transfer boundary that make the conclusion true. The difference prevents a narrow bilateral conclusion from being advertised as an unrestricted property of the artefact.

The distinction also makes the processor rule visible within the status claim itself. An independent research institute that determines its own purposes and means may receive an R2 artefact with S1 scope, provided that identification remains insignificant from its perspective. A processor acting on behalf of a hospital or retailer does not receive the same conclusion merely because it lacks the key. After all, the controlling entity's perspective governs the relationship between the processor and the controller. A public release requires broader S0 evidence because the recipient universe and auxiliary data expand. At the same time, a trusted environment may support an S1 conclusion without, by itself, transforming identifiable source material into non-personal data.[4]

Scope Is a Separate Axis

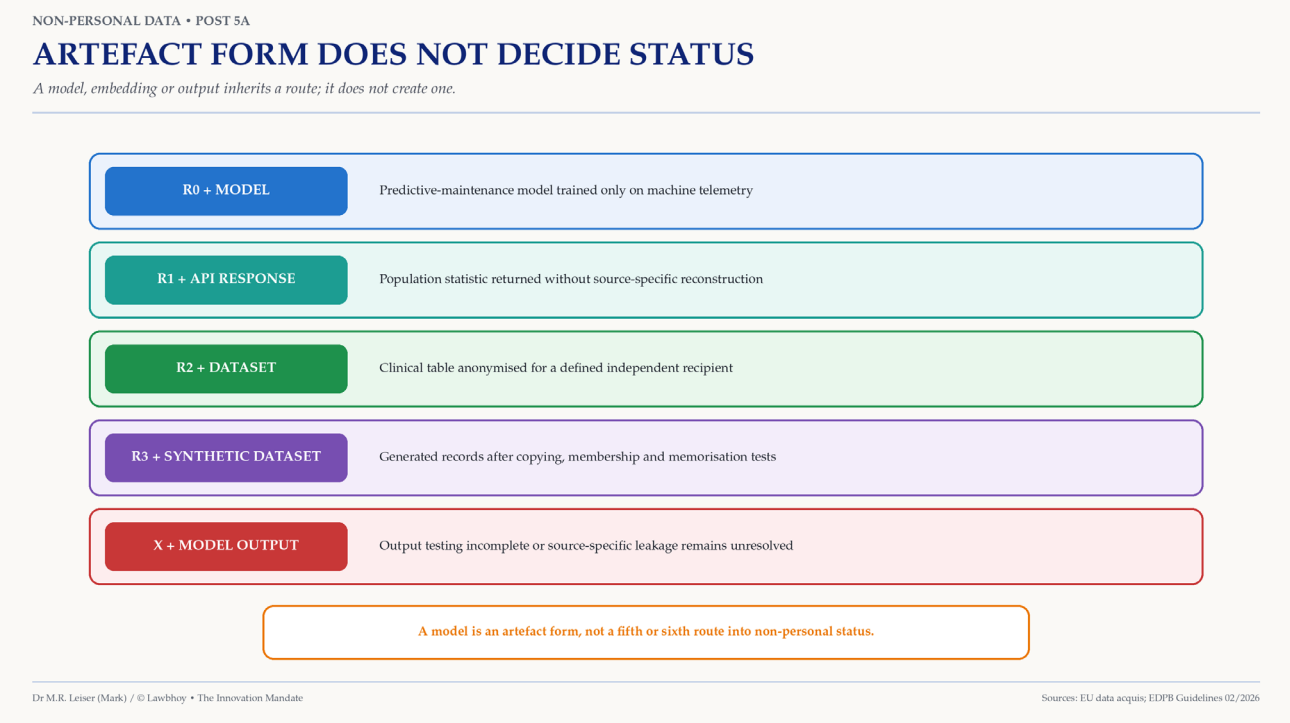


The same transformed dataset may remain personal for a source controller, personal for its processor, non-personal for an independent recipient and unsuitable for public release.

Artefact form is not a route.

Modern data economies increasingly circulate models, embeddings, features, retrieval indices, benchmarks, and outputs rather than row-level datasets. Still, these artefacts do not constitute an additional route to non-personal status. A model trained entirely on machine telemetry may follow R0; an API response containing a population statistic may follow R1; a model output derived from transformed personal data may follow R2; and a synthetic dataset follows R3 only after the required privacy tests. Artefact form determines which risks and tests become salient, while route explains how the non-personal conclusion was reached.

This separation becomes especially important for artificial intelligence because the same model family may contain versions with very different privacy behaviour. Training data, retrieval layers, prompt interfaces, output controls, and model updates can alter the likelihood of extraction, memorisation, or source-specific inference, even when the product name remains unchanged. A status claim must therefore be attached to a versioned artefact and a defined interface. At the same time, the form layer indicates to the assessor whether record isolation, linkage, membership, extraction, inversion, or prompt-based disclosure requires testing.[16]



Models, embeddings and outputs inherit a route and require form-specific testing; they do not create an additional legal route.

Quarantine is not a third legal category

X marks mixed, inseparable or unresolved material that cannot yet support a defensible claim of non-personal status. A bundle containing personal and non-personal components should not be described as wholly non-personal merely because most fields are technical or aggregate. At the same time, a model output with incomplete leakage testing should not be placed outside the GDPR because the organisation expects the missing evidence to be favourable. X is an operational instruction to separate, suppress, transform, narrow circulation or complete the assessment, rather than a permanent category sitting between personal and non-personal data.

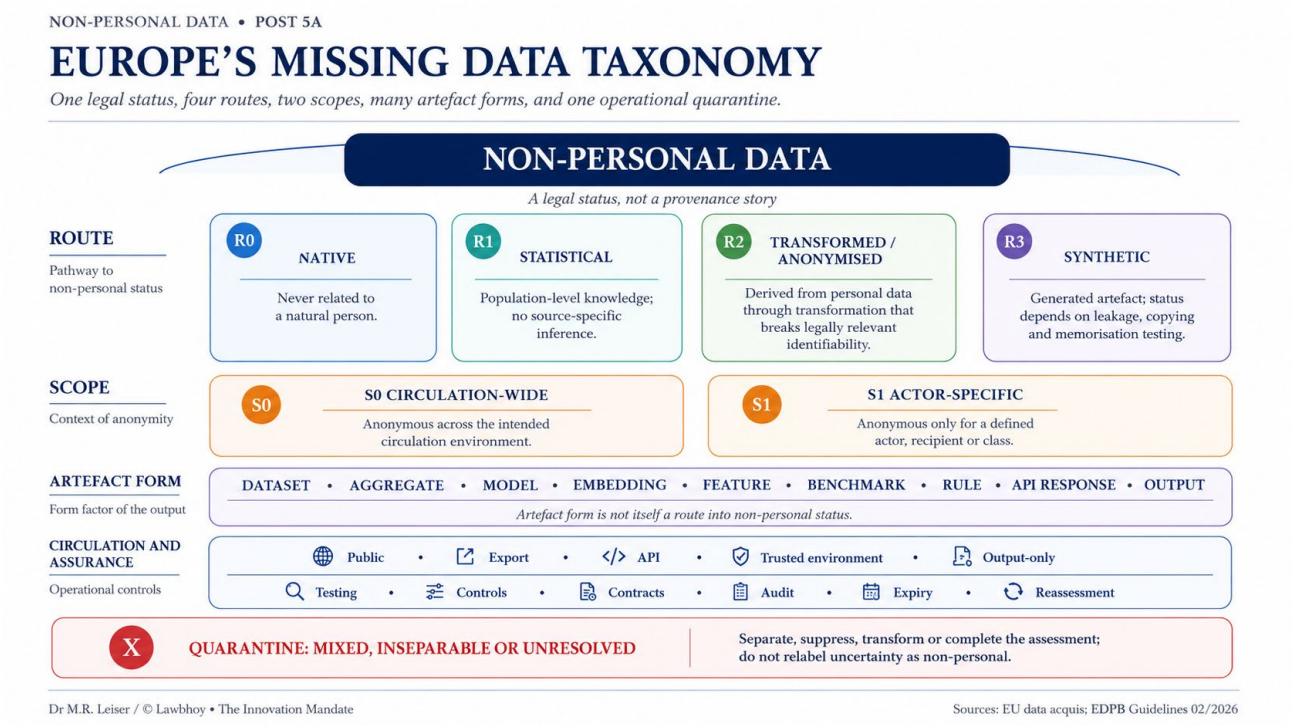
The treatment of mixed datasets should remain proportionate to the actual vulnerability. One identifiable record can require that the operational whole be treated as containing personal data when its components cannot be processed separately. Still, it should not create a legal presumption that every other record is identifiable through the same route. The assessment should ask whether the personal component can be separated, whether the vulnerability is shared, and whether the remaining material can support its own claim to status. Quarantine preserves caution without allowing a single anomaly to become an unexplained contagion across an entire database.[4]

What the Digital Omnibus Should Do

The Digital Omnibus already proposes consolidating parts of the data acquis into the Data Act, creating an opportunity to add conceptual consolidation alongside structural simplification. The legislation should retain the broad umbrella definition of non-personal data while recognising the separate layers of status, route, scope, form, circulation and assurance. It should require actor-specific claims to identify the recipient or class and the boundary for onward transfer. In contrast, general release claims should state the circulation environment that formed part of the assessment. Mixed or unresolved material should remain within the operational personal-data regime until separation or further assessment supports a narrower conclusion.[17]

A cross-acquis taxonomy would not weaken the GDPR's rights floor because the legal status test remains anchored in Article 4(1) and Recital 26. It would instead require the EDPB, Commission, AI Office, data-space authorities and national regulators to explain which route they recognise, what scope they accept, which tests they require and why the resulting burden is proportionate to the rights risk. That approach follows the Innovation Mandate's central discipline: rights remain fixed by law, while economically consequential supervisory discretion must become transparent, evidence-based and no more burdensome than the legal objective requires.[18]

Non-personal data is one legal status, not one kind of artefact.



Sources and references

1. Regulation (EU) 2018/1807 on a framework for the free flow of non-personal data, especially Articles 1–3; official text.

2. European Commission, COM(2019) 250 final, Guidance on the Regulation on a framework for the free flow of non-personal data.

3. Regulation (EU) 2022/868 (Data Governance Act), Regulation (EU) 2023/2854 (Data Act), and Regulation (EU) 2024/1689 (AI Act), especially Articles 3(51) and 59(1)(b) of the AI Act; AI Act official text.

4. European Data Protection Board, Guidelines 02/2026 on Anonymisation, especially pp. 2–18, 21–29 and Annex 1.

5. Regulation (EU) 2016/679 (GDPR), Article 4(1) and Recital 26; official text.

6. Latanya Sweeney, "k-Anonymity: A Model for Protecting Privacy" (2002) 10 International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 557.

7. Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymisation of Large Sparse Datasets" (2008) IEEE Symposium on Security and Privacy 111.

8. Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" (2010) 57 UCLA Law Review 1701; Ira Rubinstein and Woodrow Hartzog, "Anonymization and Risk" (2016) 91 Washington Law Review 703.

9. Nadezhda Purtova, "The Law of Everything: Broad Concept of Personal Data and Future of EU Data Protection Law" (2018) 10 Law, Innovation and Technology 40.

10. Michèle Finck and Frank Pallas, "They Who Must Not Be Identified: Distinguishing Personal from Non-Personal Data under the GDPR" (2020) 10 International Data Privacy Law 11; Inge Graef, Raphaël Gellert and Martin Husovec, "Towards a Holistic Regulatory Approach for the European Data Economy" (2019) 44 European Law Review 605.

11. Case C-582/14 *Breyer v Bundesrepublik Deutschland* EU:C:2016:779.

12. Case C-319/22 *Gesamtverband Autoteile-Handel v Scania* EU:C:2023:837.

13. Case C-604/22 *IAB Europe v Gegevensbeschermingsautoriteit* EU:C:2024:214.

14. Case C-479/22 P *OC v European Commission* EU:C:2024:215.

15. Case C-413/23 P *EDPS v SRB* EU:C:2025:645.

16. European Data Protection Board, Opinion 28/2024 on certain data-protection aspects related to the processing of personal data in the context of AI models.

17. European Commission, COM(2025) 837 final, Proposal for a Digital Omnibus Regulation.

18. The Innovation Mandate lecture series, especially the treatment of digital regulation as economic infrastructure, supervisory discretion, proportionality, assurance and evidence.