**Response to the public consultation**
**on the EDPB Guidelines 01/2025 on Pseudonymisation.**
**Date: 13th March 2025.**
**Author: Fabio Ricciato.**

**Declaration: I am submitting this feedback in my personal capacity. The following feedback should not be construed as representing the official position of my employer.**

---

I very much welcome these guidelines and I acknowledge that the content of the document will be very helpful to promote meaningful use of pseudonymization. I have appreciated several points made in the document. In my feedback below I will not repeat the many positive points, but rather focus on a few points for improvement.

I have two kinds of feedback:

- Part A. Relatively small suggestions for improving presentation of certain key concepts and formulation of specific sentences.
- Part B. A suggestion for content extension in the direction of considering secret sharing as an advanced form of pseudonymization, and data processing automatization in a secure computation environment as an important measure to complement (not replace) pseudonymization.

**Part A.**

| | |
|---|---|
| (i) | I welcome the concept of "pseudonymization domain". You may include at point 10. a more explicit and compact definition by spelling out the two qualifying conditions of the pseudonymization domain (PD):<br>• The pseudonymized data must not get out of PD<br>• The auxiliary information must not get into the PD.<br>You may highlight that it is always up to the data controllers to adopt measures to ensure (enforce) that both such conditions are met also vis-à-vis potential attackers that, by definition, do not limit themselves to legal means.<br><br>You could also highlight that two kinds of attackers are to be counteracted:<br>• Attackers that try to exfiltrate pseudonymized data out of PD.<br>• Attackers that try to infiltrate auxiliary information into the PD. |
| (ii) | At point 10. the final sentence reads: "*to determine who should be precluded from attributing the pseudonymised data to individuals*". This sentence is used to explain the role of PD. I find this sentence not entirely correct. The set of actors that should be "precluded from attributing the pseudonymised data to individuals" includes the PD but is not limited to the PD. Actually, we want to "preclude" everybody in the whole world (except the legitimate controller(s) of non-pseudonymised data), not only the PD. What distinguishes the PD from the rest of the world is not the "preclusion" condition, but rather that PD entails a higher risk compared to the rest of the world because it contains |

| | |
|---|---|
| | pseudonymised data. I think that you should find a clearer way to introduce the concept of PD, and perhaps my remark (i) above will be somewhat inspiring. |
| (iii) | Referring to "*the means they are reasonably likely to use for attribution*" (e.g. in point 42) it would be useful to highlight (e.g., in a separate numbered item 42bis?) that what is to be considered "reasonably likely" must be assessed also in relation to the nature of the data, their scale and information content. This relates to the proportionality principle. To illustrate, consider one small dataset that contains few summary variables for a small sample of the total population (e.g., the annual income for a 0.1% sample of all residents for a single month), and a large dataset containing very granular and detailed information for the whole population (e.g. every single purchase and financial transaction made by every resident, along with the details of the purchased item, for 10 years etc.). The second dataset is much more valuable than the first one for rogue actors. Assume that breaching into the data set would involve advanced attack means which would cost to a potential attacker 10 Mio EUR ("attack budget"). It may appear "reasonably likely" that a potential attacker (including, say, a foreign rogue governmental actor) would be ready to invest such a large "attack budget" to target the second dataset but not the first one. In other words, what is "reasonably likely" in terms of "means of attribution" (or "attack budget") must be assessed proportionally to the value of the target data.<br>I think it would be useful to spell out clearly this principle, e.g., in a dedicated point. |
| (iv) | At point 59 .in the sentence "*No one in the pseudonymisation domain [...] should be able to easily use the data to the disadvantage of the data subject* [...]"<br>I would consider droping the word "easily". |
| (v) | At point 112 the sentence "*Copies of data should be deleted as soon as they are no longer needed*" I suggest writing "securely deleted" instead of simply "deleted". There are different ways to erase the data from a disk, and some of them are as "weak" to allow recovering of the erased data by an attacker (with some efforts). Mentioning the need to "securely delete" the data reminds the reader of the importance of choosing a strong erasure method. |
| (vi) | Point 125 reads: "*The pseudonymisation secrets are stored in multiple locations which increases the chance of unauthorised access or use*". I propose to change the sentence into: "***Copies of** the pseudonymisation secrets are stored in multiple locations which increases the chance of unauthorised access or use*", i.e., to prepend "copies of" in front of the sentence. This is to avoid confusion with the case where the same (single) copy of the pseudonymisation secret is **divided** among multiple locations by **secret sharing**, which actually reduces the chance of unauthorized access or use. In fact, denoting by $p$ the probability of data breach at a single location and by $n$ the number of locations, it is trivial to show that (under the assumption of data breaches being i.i.d. across the different locations) **multiplying the number of secret copies increases the total risk linearly by n** (the total risk scales as $P=1-(1-p)^n \cong np$ for small $p$) while **dividing the number of secret shares decreases the total risk exponentially** (the total risk scales as $P=p^n$). You may find useful to add this explanation in a separate point (see Part B below). |

**Part B.**

- **Role of automation.** The document text seems to implicitly assume that humans have access to and process the data. It would be convenient to elaborate a bit more on the role of **automated processing**. The scenario where the processing of personal data is fully automated, i.e., is data are processed by machines, facilitates putting in place protection measures compared to scenarios where data are manipulated directly by humans. Eventually, people must configure the machines that process the data, so **who controls the machine accessing the data eventually controls the data**. But when machines are involved in the chain, one can adopt technologies (e.g., secure processing environments) to restrict who can access the machine, under what conditions and for what purposes, and such restrictions may be enforced technologically (this is what makes the processing environment "secure"). Plus, one can adopt multi-party technologies where the power to configure the machine (or equivalently, the secure environment) is divided up among multiple parties (acting as joint controllers or joint processors, depending on the case). This adds an important level of security that complements pseudonymization and should be encouraged (see point (vi) above). To simplify, I would say that a weaker form of pseudonymisation embedded in a "strong" secure computation environment may be acceptable, and in some cases even preferable over a stronger form of pseudonymization without surrounding security measures. In other words, the combined package of pseudonymization-cum-security may lower the risk of attribution down to acceptable levels that would not be achievable purely by pseudonymization. I think the document should encourage the adoption of security measures to complement (not replace) pseudonymization.

- **Secret-sharing and secure multi-party computation as strong pseudonymization**. The only reference to secret sharing is found, in brackets, in point 108: "*(For added security, they may also be divided up, e.g. by secret sharing, and stored by different entities.)".* I think that the document falls short of recognizing the value of of secret sharing as an advanced (maybe *the most advanced*) form of pseudonymization. Secret sharing was already considered as an effective supplementary measure in the context of international transfer (see the EDPB *Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data*[1] and specifically "Use Case 5: Split or multi-party processing" at page 33-34). The current document may explicitly state that secret sharing (and in general multi-party computation) represents an advanced form of pseudonymisation, in line with the content of the recent ENISA report[2] on "*Data pseudonymisation: Advanced Techniques and use cases (January 2021)"*. The document may highlight that all "classical" pseudonymisation techniques (i.e., all those

[1] *https://www.edpb.europa.eu/system/files/2021-06/edpb_recommendations_202001vo.2.0_supplementarymeasurestransferstools_en.pdf*

[2] https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Data%20Pseudonymisation%20-%20Advanced%20Techniques%20and%20Use%20Cases.pdf

considered in the current version of the document) work by “splitting” each data element into two “pieces”, namely (i) the pseudonymised data element and (ii) the auxiliary information. The former is non-attributable (to an identifiable person) but is still intelligible (i.e., the actors in the pseudonymised domain can read the value of the variable but cannot identify the invidual to which the variable is referred to). Secret sharing, and in general multi-party computation, takes one step further in multiple directions. First, it divides the data element into a number of “pieces” (i.e., “shares”) that may be larger than two, which clearly increases the level of protection. Second, each individual piece is not only non-attributable, but also non-intelligible by the individual actors within the pseudonymised domain (they can not even read the value of the variable). From the perspective of data protection, a well designed secret sharing scheme may offer a level of protection that is larger or much larger to that of “classical” pseudonymisation, at the cost of higher resource consumption (communication bandwidth, computation and organisational costs). Failing to mention secret sharing as a powerful and “strong” form of pseudonymisation in the document may not only represent a missed opportunity for *promoting* such techniques, but may even contribute to *demote* it among potential adopters. This would be a major loss for the cause of data protection, at a time where such techniques are becoming viable and affordable in real world deployments (see e.g. https://cros.ec.europa.eu/joconde). I suggest to include one or more dedicated points to explain the value of secret sharing as an advanced form of pseudonymisation and to include one or two examples of applications in the annex.