Feedback as part of the public consultation on the Guidelines 02/2026 v1 on Anonymisation issued by the European Data Protection Board

This feedback was drafted by five experts from three German research institutes (in alphabetical order): Daniel Buck (Head of the RDC DZHW), Dr Andreas Daniel (Deputy Head of the RDC DZHW), Dr Aleksander Kocaj (Head of the RDC IQB Berlin), Anne Rucker (Researcher at the RDC IQB Berlin) and Oliver Watteler (Senior Expert for RDM and Data Protection at Gesis). This feedback is submitted in our personal capacities and does not necessarily reflect the views of the institutions we work for or organisations we are associated with.

In Germany, there are more than forty research data centres (RDC) in the social, economic, and educational sciences

The RDC make microdata available for scientific reuse in the form of Scientific Use Files (SUF) and Campus Use Files (CUF) for research, university teaching and practical exercises. These data are derived from the institutes' own surveys, interviews and observational studies, and from archived studies conducted by third parties. The purposes of use, including the provision to third parties for research and teaching, are explained to participants as part of the informed consent process. Consent thus forms the basis for the provision of the data. Furthermore, official statistical data is also made available to researchers, educators and the public in accordance with the statutory frameworks.

Following processing and anonymisation, the data is made available once a data usage agreement has been concluded and the purpose of use has been verified. In individual cases, acceptance of the terms of use is sufficient (e.g., Campus Use File). There are four access methods, with increasing levels of control:

- **Download:** sometimes requiring registration, usually with a contract for data use and deletion deadline. This is the most common access method.
- **Remote Access:** Access after registration and conclusion of a contract for data use. Analysis on RDC servers on a virtual desktop via remote access. All imports and exports are checked by RDC staff.
- **On-site access:** like remote access, but in a secure room with additional physical security measures.
- **Remote Execution:** Users submit analysis scripts which are then run by the RDC staff and receive checked outputs without ever having access to the original data.

The degree of statistical anonymisation depends on the method of access. When determining whether a person is identifiable, account is taken of all means that, according to generally accepted judgement, are likely to be used. This also includes the possibilities of drawing on additional knowledge. These possibilities are, in some cases, severely restricted by access regimes, contractual obligations and technical and organisational measures.

We suggest that the following comments and amendments be considered in the final version of the guidelines

Para. 31 – Motivation, value and risk for the re-identifying entity. The relationship between excluded motivational factors and permissible factors should be clarified. Under the current wording, the value of re-identified data counts only 'to the extent that it offsets the costs', meaning that it can increase, but not reduce, the likelihood of identification. From an action-theoretical perspective, motivation is largely derived from value, costs and risks. The Guideline should either recognise a low achievable value to be considered symmetrically or explain why incentives count only in one direction. Furthermore, a low motivation for re-identification is a reasonable assumption for certain user groups, such as researchers affiliated to a research institution who are additionally bound by recognize a low achievable value to be considered symmetrically, or explain why incentives count only action-theoretical perspective, motivation is largely derived from values, costs, data-use agreements, pursue documented research purposes, and face institutional and professional sanctions in the event of misuse. For such groups, motivation can be assessed based on objective, verifiable characteristics of the group rather than individual intentions. The guidelines should take this into account and recognise it as a relevant factor in assessing the means whose use can reasonably be expected.

Furthermore, it is stated that an individual's motivation is difficult to assess and may change over time. It is therefore argued that motivation should not be used as a factor in assessing the data. However, many other factors that may change over time are nevertheless used in the document to support the assessments. For example, the technical capabilities available for deanonymization may change over time, as may the contextual data available to potential data users. We would therefore also wish to take motivation at the time of assessing the anonymity of the data into account. It should then be noted that motivation may change over time, just as the technical capabilities available for deanonymization may change.

A worked example of controlled access through an RDC. The Guidelines should consider the combined effect of access restricted to named and institutionally affiliated users, documented research purposes, technical access controls, binding restrictions on use and disclosure, and enforceable sanctions. The example should explain how these safeguards affect access to additional information and the means reasonably likely to be used for identification.

> **Important note:** Not all RDCs offer every method of access, as controlled access infrastructures are expensive to set up and operate and may impose restrictions on certain user groups. Highly controlled access should therefore not be treated as the main method of access. On-site access requires a considerable investment of time and money to travel to the RDC's headquarters (or a partner institution). Operating a remote access infrastructure is expensive because, in addition to the necessary hardware, software licences must usually be purchased as well. Remote computing systems with remote execution require extensive knowledge of the datasets, as these cannot be previewed in advance. Furthermore, analysis scripts must be submitted 'blind' during remote execution. Feedback on analysis results or error messages is delayed and only provided after intermediate verification steps. Consequently, access to research data, esp. for early-career researchers, could be restricted and present additional barriers (potentially constituting a restriction on non-discriminatory access to data).

Para. 35 – Increasing likelihood. The assumption that the likelihood of re-identification "typically" comes without any specific reference. As in other cases the text (i.e. "sufficient evidence" in para. 33) would greatly benefit from more precision. One could also argue that data can be outdated after twenty or thirty years. The paragraph clearly refers to the fact that data protection measures should be evaluated against the state of technology, and it would be good if the authors could base their arguments on such best practise guidelines.

Para. 85 – The case of scientific surveys should be added in more detail. We would suggest adding these aspects.
1. The uniqueness of cases and differentiation of sample and population data.
This refers to the distinction between uniqueness within a sample and uniqueness within a whole population. Uniqueness within a sample means that there may still be other instances of this specific type within the corresponding full population. This often constitutes the most important safeguard for survey data.

> Explanation: The distinction between uniqueness within the sample and uniqueness within the population should be added, which often constitutes the most important safeguard for survey data. A complete survey data set without unit non-response, or a complete data set from official statistics, guarantees that a unique case in the data set can be unambiguously matched to a single case in the population. In the case of an incomplete sample, a unique case in the data set with a specific combination of characteristics could still exist multiple times within the population.

2. Measurement and response errors. This refers to measurement and response errors in self-reported data, which should be recognised as factors that reduce the re-identification risk.

3. Time as a protective factor in re-identification. This refers to time, which can act as a protectional instance as characteristics become partly outdated over time and are no longer relevant attributes that can be used for re-identification.

Furthermore, the phrase 'the number of individuals whose data are contained in the record' should be clarified, as a record, by definition, refers to a single individual.

Para. 87 / Example 20 – Verifiability. The guidelines should explicitly state as a verification step, that a match whose accuracy cannot be verified by the body responsible for re-identification does not constitute effective reidentification. This would ensure that the reasoning regarding sampling uncertainty and the lack of knowledge about participation is explicitly established in the main text, rather than being left to an example.

Paras. 35, 41 and 83 – 88 – Scope and documentation of the assessment. The guidelines should clarify what constitutes an adequately supported anonymity assessment. This should include guidance on selecting relevant perspectives and identification scenarios, documenting assumptions about auxiliary information, evaluating the reliability of results, and determining when the assessment is sufficiently

comprehensive. Legal requirements should be distinguished from recommended good practice. Guidance should also identify material changes that warrant reassessment and explain when an existing assessment can be updated or reused. This would support consistent and proportionate assessments across research data centres.

The assessment of alternative protection measures should also consider their effects on scientific validity. Suppression, generalisation and other transformations may affect estimates, associations, longitudinal or multilevel structures, and the analysis of small population groups. Information loss therefore concerns the interpretability and validity of research findings as well as the amount of information retained. The guidelines should encourage the selection of effective safeguards that preserve the intended scientific purposes, including the possibility of studying rare or disadvantaged groups.

Paras. 63, 69, examples 18, 20 – unsubstantiated statements. The Guidelines should refrain from unsubstantiated statements like “data will often also satisfy”, "it may often be possible", “analysing certain attributes”, or “[inference] with a significant confidence” without any reference.

Proportionality, loss of information and freedom of research. The guidelines aim to ensure that data is anonymised ‘where possible and proportionate’ (p. 2) to enable ‘free and fruitful use of data’ (p. 2). However, they do not address the extent to which the required level certainty in the assessment (paras. 21, 36, 48 and 88), relates to the loss of analytical value and the freedom of scientific research as provided for in Article 13 of the Charter of Fundamental Rights of the European Union (CFREU). Freedom of research should not be used as an argument for the illegal and ethically questionable re-identification of identifiable data, but it should play a part in determining the proportionality of the assessment criteria and the expected margins of safety. The guidelines should also make explicit reference to Article 89(1) of the GDPR and Guidelines 1/2026 for cases in which (overly strict) anonymisation would negate the scientific value of the data.

Note: DeepL was used for language editing and proofreading for parts of this text. All content has been written, reviewed, and verified by the authors, who take full responsibility for the text.