Comment on Guidelines 02/2026 on Anonymisation

Controlled research environments and textual, multimodal and historical data

Submitted by: Dr. Jan Krasni, Center for Research on Antisemitism (Zentrum für Antisemitismusforschung), Technische Universität Berlin.

Introduction

This comment complements the joint response to the consultation that originated in the Ethical, Legal and Social Aspects (ELSA) section of the German National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur, NFDI) and was consolidated by Patrick Brunner and Shlomi Hod. I am one of its co-authors. The following comment brings the perspective of a research infrastructure working with sensitive collections of historical and contemporary hate, and makes two points about the Guidelines of the European Data Protection Board (EDPB): how they apply to controlled research environments, from which only reviewed outputs are released, and how they apply to textual, multimodal and historical material in the social sciences and humanities (SSH). Section 5 adds a short observation on pending case law.

The points rest on an ongoing series of interviews with holders of sensitive collections. Roughly 35 of the 100 planned interviews have so far been conducted, covering institutional archives, historical libraries, civil society organisations, media observatories and research institutes. Preliminary findings suggest that continued control over their collections is an important condition for holders to open them to research. Two worked examples appear in the Annex.

Requests

1. **Outputs of controlled environments:** Confirm that, in controlled environments, the anonymity test applies to outputs intended for release without personal-data restrictions. These outputs are assessed from the perspective of their recipients or, if published, of the public, while the source material and its processing remain subject to the General Data Protection Regulation (GDPR) (paragraphs 10, 12, 37 and 89; section 1).
2. **Conditions under which originals are held:** Consider the legal, jurisdictional and technical conditions under which the originals are held as factors in the anonymity of released outputs, where they materially affect the availability of additional identifying information (paras. 29(b), 30, 32–33; Example 8; section 2).
3. **Measures under para. 34:** Consider naming, by way of example, measures that count as reliable, verifiable and enforceable in research settings. These should distinguish technical controls, disclosure review of outputs and accountability measures, with data use agreements as a complement (section 3).
4. **Roles:** Clarify how controller and processor roles in controlled environments relate to the anonymity of their outputs, given that these roles depend on who decides on purposes and essential means rather than on access to the data (para. 15; footnote 7; section 1). An example involving a research consortium in which the data holder, the operator of the environment and the research institution are established in different Member States would be particularly useful.

5. **Worked examples:** Add worked examples on unstructured text and archival material, such as those in the Annex. These would provide a textual counterpart to Example 24, showing how No Record Isolation, linkage and inference apply where a document is searchable or relates to several people. Example 1 could also serve as the stress test of data use agreements and secure processing environments proposed in the joint response.
6. **Layered records:** Clarify that, where the same document exists as a digitised image, a transcript and a catalogue record, each representation must be assessed in its own right and all of them together. This includes the risk that shared reference numbers re-link representations released separately (section 4; Annex, Example 2).
7. **Reassessment:** Clarify what counts as a material change in externally available information that should prompt the reassessment of anonymised material, such as the online publication of a large archival collection in another country (para. 35; section 4).
8. **Research beyond anonymisation:** Clarify that where anonymisation would remove the information a research question depends on, graded access with Article 89 GDPR safeguards remains available on an applicable legal basis, and that a failed anonymity assessment does not bar research. Cross-reference Guidelines 1/2026 on processing of personal data for scientific research purposes (section 4).
9. **Pending case law:** Take the pending reference in Case C-654/25 into account when finalising or later revising the last sentence of para. 29 (section 5).

1. In controlled environments, anonymity is assessed for released outputs

In secure processing environments (SPEs), trusted research environments and federated compute-to-data infrastructures, identifiable research data can be processed under controlled access arrangements and the applicable GDPR safeguards, without being transferred to external researchers for unrestricted use. Depending on the architecture, authorised researchers submit code for execution or work with the data through controlled interfaces; historians and qualitative researchers cannot always work through code alone. Outputs intended for release without personal-data restrictions undergo disclosure review and an assessment of anonymity from the perspective of their recipients (the “Five Safes” model; Desai, Ritchie and Welpton 2016). The source material and its processing remain subject to the GDPR.

The Guidelines already supply the elements for this distinction. Para. 10 accepts anonymisation without deletion of the originals, which then remain personal data for the controller (para. 37). Footnote 43, although written for inference, models the given data as the output A(D) of a process applied to the original data D. Union legislation follows the same pattern. The Data Governance Act (Regulation (EU) 2022/868, Article 2(20)) defines an SPE as an environment whose provider determines and supervises all processing, including display, download, export and the calculation of derivative data through computational algorithms. Under the European Health Data Space Regulation (Regulation (EU) 2025/327, Article 73(2)), whose provisions on secondary use apply from 2029, users may download only non-personal data, including data in anonymised statistical format, after review by the health data access body.

Two consequences follow. First, published outputs must be assessed from the perspective of the public (para. 12, Example 1), while inside the environment only authorised users count as having access (para. 89). Second, SPEs limit the chains of entities through which

identification could proceed (paras. 9 and 28) and make them assessable. Recipients never hold record-level data, so they cannot pass it on or sell it. Residual risks remain, among them cumulative outputs, collusion, model extraction and security failures; these are assessed together with the released outputs and the exposure of the environment holding the originals.

Roles also need clarification. Physical possession of data, access to them and legal responsibility for their processing are separate questions. The operator of an SPE may process data on behalf of a controller (para. 15). An external researcher who decides on the purpose and essential means of an analysis may be a controller, or a joint controller, without access to the data (footnote 7; Case C-210/16, Wirtschaftsakademie); writing code does not by itself settle the question. The same Regulation sets out two models for health data. The health data user is controller for processing under a data permit inside the SPE, with the health data access body acting as its processor (Article 74(1), third subparagraph). National contact points or authorised participants that place data in the secure processing environment managed by the Commission are joint controllers, with the Commission as processor (Article 75(9)).

2. The conditions under which originals are held are part of the assessment

If outputs are anonymous only because the originals are protected, the protection of the originals becomes part of the anonymity assessment. Example 8 illustrates this for poorly secured originals. The Guidelines also count foreign intelligence agencies among the relevant entities (para. 30) and accept that legal prohibitions do not restrain actors who are not bound by them (paras. 32–33).

The legal and jurisdictional conditions under which the originals are stored and processed should therefore be taken into account where they materially affect the availability of additional identifying information. This includes the provider's ability to access the data, the applicability of third-country disclosure obligations (for example under the United States Clarifying Lawful Overseas Use of Data Act, the CLOUD Act), the effectiveness of technical protections, and the likelihood that such access could enable re-identification of released outputs. The relevant question is not data location alone, but *whether the governance, legal and technical arrangements effectively restrict the means reasonably likely to be used for identification.*

3. Paragraph 34: measures that are reliable, verifiable and enforceable

Para. 34 rightly treats contractual prohibitions as a complement to technical measures, which must be reliable, verifiable and enforceable. The EDPB could usefully name, by way of example, measures that meet this standard in research settings, distinguishing three kinds:

- **Technical controls that prevent exposure:** code-only or remote access, and interfaces that do not allow record-level data to be exported.
- **Disclosure review of each output before release:** This includes established rules of thumb (for example, at least ten unweighted units per table cell and no single contributor above 50 % of a cell), applied by trained staff as conventions requiring judgement rather than as proof of anonymity (Bond, Brandt and de Wolf, n.d.). It also includes review of trained models (GRAIMATTER 2022).
- **Accountability measures that document what was approved:** audit logs of who ran which analysis on which data, and persistent identifiers that trace each released output to its analysis. These support verification but do not themselves prevent re-

identification. The identifiers belong in a protected provenance register, since publicly linking separate analyses could itself aid linkage.

Data use agreements *complement these measures but do not replace them*. The model discussed here should also be distinguished from arrangements in which the recipient obtains a copy of the underlying data and processes it in its own confidential-computing environment. Confidential-computing technologies may support an SPE, but they do not determine its governance or access model.

4. Textual, multimodal and historical material

For text, images and audio, the criteria behave differently from tabular data:

- **Uniqueness:** Unstructured records frequently contain distinctive combinations of expressions, events, relationships and contextual attributes, so they will often fail No Record Isolation. Para. 99 and Example 24 show that this is not decisive: a unique record can still be anonymous where its attributes cannot be matched to other available data. That route rarely applies to textual collections, whose content is intelligible by design and often recorded elsewhere or searchable by its wording. The uniqueness of a document is also not the same as the isolation of a record relating to one individual (paras. 54–55). Whether No Record Isolation is retained or revised (the joint response questions its value), the Guidelines should explain how it applies where a document relates to several people. They should also explain how it interacts with linkage and inference, which are the decisive risks for textual collections.
- **Content as search key:** Verbatim search, web archives and reverse image search link material back to its source. Removing names with named-entity recognition (NER) does not satisfy No Linkage. Quotations in publications are linkage vectors: in a review of Twitter studies, 72 per cent quoted tweets, and searching the quotes led to the authors in 84 per cent of cases (Ayers et al. 2018).
- **Writing style:** Writing style enables authorship attribution at internet scale (Narayanan et al. 2012).
- **Several people per record:** One item can relate to its author, to the person addressed, targeted or depicted, and to the person who reported or collected it (Glossary, "Record"). Some are deceased and outside the GDPR (Recital 27 GDPR); many are not, and archives often cannot tell which.
- **Catalogue metadata:** Metadata of restricted collections are needed for findability. For small sensitive collections, however, they can themselves enable membership inference ("12 incident reports, town X, 2024"). They should be assessed together with the material they describe, as should the transcript and digitised image of the same document (Annex, Example 2).

Changing availability of auxiliary information: Identifiability changes when previously restricted collections become accessible, including in other countries. In April 2025, for example, Argentina's national archive (Archivo General de la Nación) placed around 1,850 documents on Nazi activities in Argentina online. Declassified in 1992, they had until then been consultable only in a designated reading room; they can now be downloaded worldwide. Such openings are valuable for historical research, and nothing in this comment argues against them. They do, however, *change the auxiliary information against which records, transcripts and metadata held in European archives must be assessed*.

Most persons named in such files are deceased and outside the GDPR, but witnesses, informants and relatives may not be. Where linkage concerns living persons, earlier

assessments may no longer hold (paras. 29(c), 35). Access regimes also differ between Member States. Spain's Democratic Memory Act, for example, entitles anyone to consult in full the documents that *establish or may establish their status as a victim, including third parties' personal data appearing in them, regardless of the documents' date* (Ley 20/2022, Article 27(2)). Material released in one country may thus become linkable to records that an archive elsewhere makes available under its own rules. Para. 35 treats periodic reassessment as good practice, and the joint response proposes making it an obligation. Either way, archives need to know what counts as a material change in available information, since no archive can monitor every collection worldwide.

In hate and extremism research, the relevant entities are *not hypothetical*. Organised networks and their opponents have strong incentives and open-source intelligence (OSINT) capabilities, so the value of re-identification is high (para. 31). German criminal law now penalises the dangerous dissemination of personal data, including so-called enemy lists (§ 126a Strafgesetzbuch, StGB). The provision shows both the threat and the limits of legal prohibitions (para. 33). Severity must include third parties and small exposed groups. Realism cuts both ways, however: in many other domains, the full list in para. 30 overstates the threat.

Anonymisation and the value of research data: For much SSH material, the information that makes identification possible is the information the research needs: a place, an occupation, a family history, the wording of a post. In historical and qualitative research, removing information is not always a neutral privacy intervention. It may change the meaning of a source, prevent contextual interpretation, undermine the verification of findings, or erase the very social relationships and discriminatory practices under investigation.

Article 89(1) GDPR reflects this: where the purposes can be fulfilled with data that do not permit identification, they must be; where they cannot, processing continues with safeguards. Such processing still requires a legal basis under Article 6 GDPR, a condition under Article 9(2) GDPR where special categories of data are involved, and compliance with applicable Union or Member State law (para. 38). Graded access, up to analysis in a controlled environment, can preserve analytically necessary information while protecting the persons concerned. *Anonymisation is therefore one route to research with sensitive data, not its precondition*.

5. Means reasonably likely to be used and pending case law

The last sentence of para. 29 treats any legal means of accessing additional information as objectively a means reasonably likely to be used. The pending Case C-654/25 (Undelam) is a request for a preliminary ruling from the Bundesgerichtshof, lodged on 6 October 2025 and concerning dynamic internet protocol (IP) addresses. In it, the Court has been asked whether it suffices that, under certain conditions, there may be legal possibilities to identify the data subject, or whether those conditions must be fulfilled in both fact and law in the specific case. The EDPB should take this reference into account when finalising or later revising para. 29, as the judgment may clarify the relevance of conditional legal means of identification.

Annex: two worked examples

Example 1: controlled analysis of hate material

A research infrastructure holds digitised antisemitic propaganda, witness interviews and social media posts reported to civil society organisations. Its catalogue metadata are public. Researchers with an approved project analyse the material inside an SPE.

- **Catalogue:** For a small sub-collection, public metadata can reveal that a particular person's report is included. Metadata are therefore assessed together with the material and published at a level of aggregation that prevents membership inference (para. 79).
- **Aggregate outputs:** Only reviewed outputs leave: aggregate findings and coding schemes. *Thresholds matter.* If a rare category occurred once in a given month, someone who knew of a particular post could infer how it had been classified, which is a specific and meaningful inference (para. 67).
- **Quotations:** Before release, quotations are checked for searchability. Where a passage can be found online, it is paraphrased, omitted, contextually redacted or released only under restricted access, depending on the material and the legal basis.
- **Models:** Trained models leave only after review (para. 82; GRAIMATTER 2022).
- **Result:** Where the disclosure assessment establishes that the released outputs do not permit identification through means reasonably likely to be used by their recipients, those outputs can be considered *anonymous from the recipients' perspective*. The underlying material remains personal data, subject to applicable GDPR obligations and Article 89 safeguards (para. 37).

Example 2: archival letters

A research archive holds letters written by members of the public to institutions in recent decades, now being prepared for digitisation. The archive is considering publishing some letters in redacted form, keeping others closed, and making its catalogue metadata only partly public.

Anonymity can only be assessed layer by layer, and for each context of use (public release, reading room, computational analysis):

- **Digitised image:** The image may show handwriting, signatures and letterheads, which can be recognisable to relatives, neighbours and colleagues, whom para. 30 counts among the relevant entities. *Must redaction cover the image as well as the transcript, and should handwriting itself be treated as identifying?*
- **Transcript:** A transcript makes every passage searchable at scale (paras. 28, 92–93). *If redaction is automated, must every letter be checked manually? Under para. 36, a single missed name would make the whole published set personal data unless its parts are treated separately.*
- **Catalogue metadata:** Records give names, places and dates and summarise the content. They are personal data in their own right, and shared reference numbers can re-link layers released separately. *Should metadata be assessed with the same rigour as the documents they describe?*
- **Content:** Some writers (and/or wittnesses) are still alive. The GDPR does not apply to the deceased (Recital 27 GDPR), but an archive usually cannot establish who has died, and letters may mention living relatives. Letters to institutions often reveal political opinions, a special category under Article 9 GDPR. Re-identification works through context: a combination of place of residence, occupation and a reference to a family member's past can single out a writer once every name has been removed. *Is it sufficient to remove names, addresses and signatures, or must such contextual details*

go too, even though this would destroy much of the letters' value as historical sources? On what criteria should an archive decide?

- **Originals:** Archives retain originals in order to make them available, under conditions, to researchers. Para. 10 accepts anonymisation without deletion, and Example 8 shows that retained originals must then be protected. Every redacted version therefore remains linkable to its original, both by the archive itself, for which it remains personal data (para. 37), and by every authorised user of the originals. Access to the originals may enable linkage for authorised users without necessarily making the redacted versions identifiable to the public (para. 89 and, by analogy, Example 21). The public assessment must nevertheless consider other reasonably available information, including external collections and independently released representations of the same document. *Can the redacted versions therefore be anonymous for the public where access to the originals is effectively and permanently restricted (paras. 29(b), 89–90)? And what weight do reading-room conditions and user agreements carry, given para. 34?*

These questions *do not show that archival material cannot be anonymised*. They show that the assessment must cover all layers together and every context of use. Where it succeeds, redacted letters can be published. Where it fails, the letters can still be opened to research through *analysis in a controlled environment*, as in Example 1. A worked example along these lines would let archives apply the Guidelines consistently rather than each deciding alone.

About the author

Dr Jan Krasni is a researcher at the Center for Research on Antisemitism, Technische Universität Berlin, and project lead of a proof-of-concept infrastructure for analysing sensitive collections of historical and contemporary hate without transferring them (see Acknowledgement; Krasni 2025).

Acknowledgement

The contribution of Jan Krasni was supported by ERIHAD.OS, funded by the OSCARS project, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751.

References

Literature

Ayers, J. W., Caputi, T. L., Nebeker, C., et al. (2018). Don't quote me: Reverse identification of research participants in social media studies. npj Digital Medicine, 1(1). https://doi.org/10.1038/s41746-018-0036-2

Bond, S., Brandt, M., & de Wolf, P.-P. (n.d.). Guidelines for the checking of output based on microdata research. Data without Boundaries. https://cros.ec.europa.eu/system/files/2024-02/Output-checking-guidelines.pdf

Desai, T., Ritchie, F., & Welpton, R. (2016). Five Safes: Designing data access for research. UWE Bristol Economics Working Paper 1601.

GRAIMATTER (2022). Green Paper: Recommendations for disclosure control of trained machine learning models from Trusted Research Environments. Zenodo. https://doi.org/10.5281/zenodo.7089491

Krasni, J. (2025). Towards an ethical strategy for research data infrastructures: Digitalizing archives of historical hate. Shagi / Steps, 11(4), 205–221.

Narayanan, A., Paskov, H. S., Gong, N. Z., et al. (2012). On the feasibility of internet-scale author identification. IEEE Symposium on Security and Privacy, 300–314. https://doi.org/10.1109/sp.2012.46

Legal and EDPB sources

Regulation (EU) 2016/679 (General Data Protection Regulation), Official Journal L 119, 4 May 2016, p. 1.

Regulation (EU) 2022/868 (Data Governance Act), Official Journal L 152, 3 June 2022, p. 1.

Regulation (EU) 2025/327 (European Health Data Space Regulation), Official Journal L, 2025/327, 5 March 2025.

European Data Protection Board, Guidelines 02/2026 on Anonymisation, version 1.0, adopted 7 July 2026 (version for public consultation).

European Data Protection Board, Guidelines 1/2026 on processing of personal data for scientific research purposes, adopted 15 April 2026 (version for public consultation).

Court of Justice, Case C-210/16, Wirtschaftsakademie Schleswig-Holstein, judgment of 5 June 2018.

Court of Justice, Case C-654/25, Undelam, request for a preliminary ruling from the Bundesgerichtshof lodged on 6 October 2025 (pending).

Ley 20/2022, de 19 de octubre, de Memoria Democrática (Spain), Boletín Oficial del Estado no. 252, 20 October 2022, Article 27.