NOVA ID
RESEARCH DATA MANAGEMENT
COMPETENCE CENTRE

NOVA.ID
ASSOCIAÇÃO PARA A INOVAÇÃO E DESENVOLVIMENTO
NOVA SCHOOL OF SCIENCE AND TECHNOLOGY

Risk and Directrices to Sensible Data Management	
Work Package	**5. Policy on Data Curation and Sharing**
Task	Task 5.1
Delivery Date	December
Dissemination Level	Public
Authorship	Kevin Gallagher João Leitão
Reviewers	Juliana Monteiro
Conclusion	January 2025

Acknowledgements

The NOVA.ID-RDM-CC Competence Centre is funded under the PNCADAI - National Programme for Open Science and Open Research Data, as part of Measure RE-C05-i08 - More Digital Science of the PRR - Recovery and Resilience Programme.

This document aggregates the deliverables D15 and D16 in the original plan. This decision was made since the contents made sense to be presented together.

Executive Summary

This consolidated deliverable merges two planned WP5 outputs: a “Report on current risks for the management of sensitive research data” and a “Report with guidelines for the management of sensitive research data.” It is intentionally written as a cross-disciplinary institutional document because, as stated by the project team, the NOVA University ecosystem covers virtually all research fields and therefore manipulates very different types of data, ranging from interviews and audiovisual materials to scientific measurements, clinical and biological data, genetic data, and computer science artefacts. In such an environment, one of the central operational challenges is that “sensitive data” is not a single homogeneous category. Sensitivity arises from multiple overlapping dimensions, including personal data protection, research ethics, confidentiality obligations, intellectual property and trade secrets, and safety and security constraints.

Within the European legal context, the General Data Protection Regulation (GDPR) provides binding definitions and requirements for personal data, special categories of personal data (often referred to as sensitive personal data), pseudonymization, and security of processing. GDPR establishes that safeguards must be “appropriate to the risk,” placing risk assessment and proportional mitigation at the centre of compliant practice.

This document therefore proceeds in two parts. The first part systematizes the principal risk categories that typically arise when managing sensitive research data at scale in a heterogeneous university environment. The second part consolidates widely accepted technical and organizational practices, grounded exclusively in authoritative sources, that can be used as practical guidance for researchers, data stewards, and institutional stakeholders. It also clarifies how sensitive-data constraints can remain compatible with FAIR and Open Science, in line with the European Commission’s principle of being “as open as possible, as closed as necessary.”

Purpose, Scope, and Method

The purpose of this deliverable is to provide a concrete, verifiable, and cross-disciplinary reference on sensitive research data management for a university-wide context. It is not intended to replace local procedures, ethics committee requirements, or legal advice. Instead, it consolidates and explains common risk patterns and mitigation approaches that can be applied consistently across disciplines, while allowing for domain-specific tailoring.

The scope is research data across the full lifecycle, including collection, creation, processing, storage, access, sharing, publication, long-term preservation, and disposal. The document treats "data" in a broad research sense that includes not only conventional datasets but also associated digital objects (such as code, workflows, and tools), reflecting the scope of the FAIR principles as originally defined.

Methodologically, the document is based exclusively on publicly verifiable online sources, prioritizing binding EU legal text and guidance from recognized public authorities and standards bodies, including EUR-Lex (GDPR), the European Data Protection Board (EDPB), the former Article 29 Working Party (WP29) opinion on anonymization techniques, the European Data Protection Supervisor (EDPS), ENISA, NIST, and ISO.

The Core Challenge: What Constitutes "Sensitive Data" in a University Research Ecosystem

In practice, identifying what constitutes sensitive data is itself one of the most persistent challenges in research data management. The reason is that sensitivity is not determined solely by the file format or the scientific domain; it depends on the relationship between the data and identifiable persons, the presence of confidential or protected content, the plausibility of reidentification through linkage, the ethical and social consequences of misuse, and the contractual and regulatory frameworks governing access.

From a European legal perspective, the baseline category is personal data, defined as information relating to an identified or identifiable natural person. GDPR also defines "special categories of personal data," including data revealing racial or ethnic origin, political opinions, religious beliefs, trade union membership, as well as genetic data, biometric data for unique identification, health data, and data concerning sex life or sexual orientation. These categories are singled out because of their heightened risks and the strong restrictions imposed on their processing.[1]

However, a university-wide view must extend beyond "special categories" because many research activities produce data that are sensitive without necessarily falling into those legal categories. Interviews, focus groups, ethnographic notes, recordings, and video reports may contain personal data, contextual identifiers, and sensitive narratives even when the research topic is not biomedical. Similarly, scientific datasets may be sensitive due to their association with critical

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A32016R0679

infrastructure, environmental risk, dual-use implications, or confidentiality constraints. Furthermore, computer science and engineering projects may produce logs, telemetry, usage traces, network captures, or application-level datasets that can become personal data if they relate to identifiable individuals or can become sensitive if they enable inference about behavior. GDPR's approach makes the identifiability criterion central, and modern data environments make identifiability increasingly contextual and dynamic.[2]

This challenge is explicitly recognized in major regulatory and technical guidance. The WP29 Opinion on Anonymization Techniques emphasizes that effective anonymization is difficult because a dataset considered anonymous may later be combined with another dataset in a way that enables identification. This directly matches the institutional reality of a large university where data are heterogeneous, reused, and linked.[3]

Report on Current Risks for the Management of Sensitive Research Data

Risk as a structural property of the research data lifecycle

A central concept across EU law and security standards is that risk management is continuous across the lifecycle rather than a one-time compliance step. GDPR explicitly frames security of processing as a matter of appropriate measures relative to risk and includes both technical and organizational measures. It also defines "personal data breach" in a way that captures accidental or unlawful destruction, loss, alteration, unauthorized disclosure, or unauthorized access. These definitions matter operationally because they set the boundary conditions for what institutions must be prepared to prevent and respond to.[4]

In a multi-disciplinary institution, lifecycle risk is amplified by diversity of practices. Some research fields rely on long-term longitudinal data that require controlled linkability; others rely on qualitative artefacts where anonymization can reduce scientific value; others rely on high-throughput instruments producing large volumes requiring distributed storage and collaborative access. These differences change the threat model, the feasible safeguards, and the residual risk.

[2] https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng
[3] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
[4] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A32016R0679

Misclassification and under-classification risk

Reidentification risk is particularly acute in environments where datasets are shared internally across teams, reused across projects, or published with partial suppression. WP29's Opinion highlights that linking datasets can defeat naive anonymization and stresses that anonymization aims to prevent identification while retaining usefulness, but that effectiveness is difficult to guarantee. ENISA's work on pseudonymization further analyses adversarial models and attacking techniques such as brute force, dictionary attacks, and guesswork against pseudonymization, illustrating that pseudonymization must be designed with an explicit threat model.[5]

For a university with many disciplines, this risk has multiple concrete forms. In qualitative studies, indirect identifiers embedded in narrative content can enable identification even when names are removed. In biomedical and genetic contexts, genetic data are explicitly included among special categories of personal data in GDPR, reflecting the unique identifiability and sensitivity of such information. In digital trace datasets, device identifiers, IP-related information, or behavioral patterns can function as quasi-identifiers in practice.[6]

Unauthorized access and privilege accumulation

A second major risk category is unauthorized access, including both external compromise and internal overexposure. NIST SP 800-122 frames the problem of protecting personally identifiable information by emphasizing context-based assessment of what constitutes PII and what safeguards are appropriate. Although NIST is not EU law, its value for this document is that it provides a detailed and widely adopted security engineering perspective on confidentiality threats, safeguards, and incident response that complements the GDPR risk-based approach.[7]

In university contexts, unauthorized access risk is increased by the multiplicity of systems used for collaboration, the existence of shared drives, diverse authentication practices, and turnover of project members. Without systematic least-privilege enforcement and periodic access review, access rights tend to accumulate over time, increasing exposure.

[5] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
[6] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A02016R0679-20160504
[7] https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-122.pdf

Loss of confidentiality through insecure storage, transport, and backup

A third risk category concerns confidentiality failures in storage, transit, and backup. GDPR explicitly references encryption as an example measure for security of processing, and encryption is similarly treated as a core technical safeguard in major security frameworks. ENISA's security and resilience guidance for data-intensive services highlights that large-scale data systems face recurring security challenges and require structured good practices for resilient and secure operation.

In concrete terms, confidentiality failures frequently result from unencrypted portable devices, misconfigured cloud storage buckets, poorly protected backup media, or ad hoc transfer of files through channels not designed for sensitive content. These risks are magnified by the scale and distribution of research activities across many units.

Integrity and availability risks, including ransomware and accidental loss

Sensitive data management is not only about secrecy; it also involves ensuring integrity and availability for legitimate research purposes. GDPR's definition of personal data breach includes destruction and loss, which means that resilience is part of the compliance and governance picture, not merely an operational concern. ENISA guidance on security and resilience in data services similarly emphasizes that data infrastructures must anticipate and mitigate integrity and availability threats.

For research projects, integrity failures can undermine scientific validity, while availability failures can disrupt ongoing studies and damage trust with participants. Ransomware and destructive attacks have made backup strategy and restoration testing central concerns, and accidental loss remains common when storage is decentralized and informal.

Governance and accountability failures

A further risk category is governance failure, meaning that even technically sound measures are undermined by unclear responsibilities, inconsistent decision-making, or absence of documentation. GDPR's accountability model implies that organizations must be able to demonstrate compliance with principles and safeguards. ISO/IEC 27001 provides a recognized framework for establishing, maintaining, and continually improving an information security

management system (ISMS), which is widely used to structure governance, risk assessment, control selection, and continuous improvement.[8]

In large academic ecosystems, governance risk often arises when there is no agreed process for classifying data, approving sharing, defining access roles, or deciding when and how anonymization should be attempted. The result is a patchwork of local solutions that is difficult to audit and difficult to scale.

Publication and sharing risks in Open Science contexts

Open Science policy encourages broad sharing of research outputs, including data, but European guidance is explicit that openness is conditioned by legitimate constraints. The European Commission describes Open Science compliance in terms of open access to publications and FAIR data, "according to the principle as open as possible, as closed as necessary." This principle becomes operationally important because it establishes that restricting access to sensitive data is not a failure of Open Science; it is often a requirement for responsible practice.[9]

The risk in practice is that researchers may interpret openness as a default obligation to publish raw data, including sensitive information, or may publish insufficiently controlled derivatives. Conversely, teams may decide not to share anything due to uncertainty, resulting in avoidable loss of scientific value. The risk report therefore leads directly to the need for guidelines that clarify feasible sharing models, including controlled access and open metadata.

Report with Guidelines for the Management of Sensitive Research Data

Guiding principles and baseline requirements

The first guideline is that sensitive data management must begin with a principled and documented approach to identifying sensitivity. For personal data, GDPR provides definitional anchors and highlights special categories as particularly restricted. For anonymization and pseudonymization, authoritative guidance clarifies that these are not merely technical labels but processes whose effectiveness depends on the threat model, context, and separation of

[8] https://www.iso.org/standard/27001
[9] https://rea.ec.europa.eu/open-science_en

additional information. EDPB guidelines characterize pseudonymization as a method to control attribution of personal data to individuals by denying that ability to some actors, and GDPR defines pseudonymization as processing where personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and subject to technical and organizational measures.

The second guideline is that safeguards must follow a risk-based logic. GDPR's security framework explicitly requires measures appropriate to risk, and this is consistent with structured security management approaches such as those embodied in ISO/IEC 27001. The institutional objective in a multi-disciplinary environment is therefore to define a repeatable process that makes risk assessment and control selection standard practice rather than an ad hoc judgement.

Identification and classification as a repeatable process

Because identifying sensitive data is itself difficult, the guideline is to treat classification as a process that is revisited at key lifecycle points, rather than a one-time decision. This is especially important in a university-wide environment where research fields and data types differ significantly, as described by the project team. Data that begin as non-sensitive may become sensitive when combined with other data or when enriched. WP29's analysis explicitly warns that datasets can become identifying when combined, reinforcing the need for iterative classification that considers plausible linkability.

A practical implication is that classification must explicitly consider direct identifiers, indirect identifiers, and contextual identifiability. It must also include non-personal dimensions of sensitivity such as confidentiality obligations and security constraints. Although this document does not prescribe an institutional classification taxonomy, it establishes that any robust approach must be able to represent these dimensions explicitly and document the reasoning.

Data minimization and purpose limitation in research design

A key guideline is that sensitive data management should begin at research design time through minimization and purpose limitation, rather than being treated as an after-the-fact storage problem. GDPR places these principles at the core of lawful processing, and they also reduce operational risk by limiting exposure and reducing the amount of information that must be protected. In many research contexts, this translates into designing instruments and protocols to

collect only what is required, structuring consent and participant information appropriately where relevant, and planning retention and deletion criteria early.

Pseudonymization for controlled linkability and reduced exposure

Where research requires linking records over time, pseudonymization is a common and recommended approach, but only when implemented with clear separation of roles and information. EDPB guidelines and ICO guidance converge on the requirement that the additional information enabling reattribution must be kept separate and protected, and that pseudonymized data remain within data protection scope. This is particularly relevant across biomedical, behavioral, and social science contexts, but it also applies to engineering and computer science datasets where linkability to individuals must be restricted for most actors while preserved for a limited set of authorized roles.[10]

ENISA provides a technical view of pseudonymization techniques and explicitly analyses attacker models and attack types. The practical consequence is that pseudonymization must be chosen and evaluated relative to plausible adversaries, rather than assumed to be safe by default. This is especially important when data will be shared across teams or institutions, or when high-dimensional data may enable inference about identity.[11]

Anonymization as a risk-managed process rather than a binary state

Anonymization is often treated in research as the preferred route to enable open sharing, but authoritative guidance consistently cautions against simplistic assumptions. WP29's Opinion provides an EU-level analysis of the effectiveness and limits of anonymization techniques and explicitly warns that anonymization is not straightforward, particularly because linkability can reintroduce identifiability. The EDPS provides further clarification by framing anonymization as the process of rendering personal data anonymous and addressing common misunderstandings. ICO anonymization guidance similarly positions anonymization as requiring understanding of techniques, strengths and weaknesses, and suitability for specific situations.

In practice, the guideline is to treat anonymization as a documented, context-aware process that includes a structured assessment of reidentification risk, including foreseeable linkage with external information. The objective is not to claim absolute impossibility of reidentification, but

[10] https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf
[11] https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices

to reduce risk to a level consistent with the applicable legal and ethical framework and the intended data sharing context, while preserving scientific utility.

Security of processing: encryption, access control, auditing, and resilience

Security controls must be designed to protect confidentiality, integrity, and availability across the lifecycle. GDPR explicitly references encryption and requires security measures appropriate to risk. ISO/IEC 27001 provides a governance structure for selecting and maintaining controls, and NIST SP 800-122 provides detailed guidance on protecting personally identifiable information and on developing incident response plans for incidents involving such information.

From a practical standpoint, the guideline is that sensitive research data should be protected by secure storage and secure transmission, robust authentication, least-privilege access allocation, logging and monitoring, and tested backup and recovery procedures. ENISA's guidance on security and resilience in data-intensive services provides further support for the proposition that data infrastructures require systematic practices to address security and resilience challenges, not only point solutions.

Controlled sharing and publication models compatible with FAIR and Open Science

A frequent operational tension in research data management is between openness and sensitivity constraints. This document adopts the European Commission's framing that open access to FAIR data operates under the principle "as open as possible, as closed as necessary," which legitimises restrictions grounded in privacy, confidentiality, and security. The FAIR guiding principles themselves emphasize that the principles apply broadly to digital assets and aim to maximise reuse and machine-actionability, but they do not require unconditional openness; rather, they support structured discoverability and access conditions.

In practice, this means that sensitive datasets can remain FAIR when metadata are sufficiently rich to support discovery, when access conditions and restrictions are clearly described, and when controlled access mechanisms are used where appropriate. The guideline is therefore not "publish everything," but "publish what can be published safely," and for what cannot, publish the maximum feasible information about existence, provenance, structure, and access pathways so that reuse can occur under legitimate controls.

Documentation, accountability, and continuous improvement

A final guideline is that effective sensitive data management requires documentation of decisions and continuous review. GDPR's accountability approach implies that organizations must be able to demonstrate that safeguards exist and that risk is managed. ISO/IEC 27001 similarly frames information security as a system of continual improvement. EDPB guidance on pseudonymization and WP29 guidance on anonymization reinforce that these measures depend on context and evolving threat models, meaning that periodic reassessment is integral rather than optional.[12]

For a large university ecosystem with diverse data types and practices, continuous improvement is particularly important because identifiability risks evolve as new external datasets become available and as analytic methods advance. Maintaining the effectiveness of safeguards therefore requires periodic review of reidentification risk assumptions, access patterns, storage configurations, and incident learnings.

Concluding Remarks

This consolidated deliverable reflects WP5's contribution by providing an evidence-based synthesis of current risk patterns and established mitigation techniques for sensitive research data management across a heterogeneous university environment. It explicitly recognizes that identifying sensitive data is a primary challenge and that disciplinary diversity amplifies the need for systematic and repeatable approaches to classification, risk assessment, and safeguard selection. It also clarifies that responsible restrictions on sensitive data sharing are compatible with European Open Science expectations under the principle of being as open as possible while closing data access when necessary for legitimate reasons.

References

All references were consulted in December 2025.

- Regulation (EU) 2016/679 (GDPR), EUR-Lex (HTML and consolidated PDF versions).
- EDPB, Guidelines 01/2025 on Pseudonymisation (PDF).

[12] https://www.iso.org/standard/27001

- Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques (WP216) (PDF).
- EDPS, “10 Misunderstandings Related to Anonymisation” (PDF).
- ENISA, “Pseudonymisation techniques and best practices” (publication page and supporting PDF references).
- ENISA, “Good Practices and Recommendations on the Security and Resilience of Big Data Services” (PDF).
- NIST, SP 800-122: Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) (PDF and NIST publication page).
- ISO, ISO/IEC 27001:2022 overview page (ISMS).
- European Commission (REA), Open Science guidance stating “as open as possible, as closed as necessary.”
- Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Scientific Data (2016)