DataGEMS

Data Discovery Platform with Generalized Exploratory, Management, and Search Capabilities



Report

D1.8
D1.5a Ethics Compliance

PU

24/7/2026

Disclaimer

This document contains description of the DataGEMS project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) protection rules so, prior to using its content please consult carefully the licensing information of this document and/or contact the Project and the Editor for any further information.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify the Project and the Editor immediately.

The editor and authors of this document have taken any available measure in order to ensure that its content is accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content. The information in this document is provided “as is” without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

This project has received funding from the European Union’s Horizon Europe framework programme under grant agreement No. 101188416.



Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency, neither of which can be held responsible for those views and opinions and are not liable for any use that may be made of the information contained therein.

Document Information

Type	Report
Identifier	D1.8
Title	D1.5a Ethics Compliance
Dissemination level	PU
Due Date	30/6/2026
Actual Delivery Date	24/7/2026
Lead Beneficiary	ARC
Other beneficiaries	
Editor	Cristina Contero Almagro
Authors	
Contributors	
Reviewer	Ana I. Pereira
Workpackage(s)	WP1
Keywords	Ethics, GDPR, AI Act, data governance, Trustworthy AI, fundamental rights
Version	1.0
Version information	Number of pages: 46 Last file save: 24/07/2026 2:07 PM File revision: 251 File name: D1.8_M18_D1.5a Ethics Compliance

Copyright Information

License: Creative Commons Attribution (CC-BY)
This document is released under the Creative Commons Attribution (CC-BY) license, which allows for the free use, distribution, and adaptation of the work, provided proper attribution is given to the original author(s). By accessing or using this report, you acknowledge and agree to comply with the terms and conditions of the CC-BY license. For the full text of the license, please visit: https://creativecommons.org/licenses/by/4.0/legalcode

Document Version

Version	Date	Change Description	Editor
1.0	24/7/2026	Submitted to EU	Cristina Contero Almagro

Document Tables

Table of Tables

Table 1 Provisional Inventory....28
Table 2 Recommendations and action plan....36

Table of Figures

Figure 1 **DataGEMS ethics and compliance governance cycle.** The figure illustrates a continuous process for identifying and mapping processing operations and AI systems; allocating legal roles and classifying risks; designing and documenting safeguards; monitoring implementation and escalating incidents or ethical concerns. The cycle should be coordinated through a multidisciplinary project ethics and data governance function and applied across the shared platform and all four use cases....35

Table of Contents

Disclaimer 2
Document Information 3
Copyright Information 3
Document Version 4
Document Tables 5
Table of Contents 6
Abstract 9
1 Introduction 11
1.1 Background 11
1.2 Nature of the assessment 11
2 Project background 12
2.1 Consortium and work package context 12
2.2 M18 implementation context 12
3 Purpose and scope 13
4 Methodology and sources reviewed 14
4.1 Source hierarchy 14
4.2 Assessment method 14
4.3 Questionnaire coverage 14
4.4 Documentary limitations 14
5 Legal and ethical framework 16
5.1 Grant Agreement and fundamental rights 16
5.2 GDPR 16
5.3 Pseudonymisation and anonymisation 16
5.4 Non-personal and mixed data 16
5.5 AI Act 17
5.6 Trustworthy AI and ethics-by-design 17
5.7 Copyright, database rights and text and data mining 18
5.8 Accessibility and national law 18
6 M18 data processing and AI landscape 19
6.1 Platform identity, access and logs 19
6.2 Discovery, profiling and recommendation 19
6.3 MathE 19
6.4 Lifelong learning 20
6.5 Weather 20
6.6 Language data 20
7 Fundamental rights and ethical assessment 21
7.1 Privacy and informational self-determination 21
7.2 Equality, fairness and non-discrimination 21

7.3 Education and human agency 21
7.4 Access to information and academic freedom 21
7.5 Misuse and over-reliance 22
7.6 Societal and environmental well-being 22
8 GDPR and data compliance assessment 23
8.1 Roles and accountability 23
8.2 Categories, sources and special-category data 23
8.3 Purpose limitation and legal bases 23
8.4 Transparency and platform terms 24
8.5 Data subject rights 24
8.6 Data minimisation and privacy by design 24
8.7 Accuracy 24
8.8 Storage limitation and retention 25
8.9 Security 25
8.10 DPIA and risk assessment 25
8.11 International transfers and external services 25
8.12 Anonymisation, pseudonymisation and re-identification 26
8.13 Data sharing and governance 26
9 Free flow of non-personal data and wider data governance 27
10 Artificial intelligence assessment 28
10.1 Provisional AI inventory 28
10.2 Prohibited practices 29
10.3 High-risk classification 29
10.4 Research and development context 30
10.5 AI literacy 30
10.6 Transparency and explainability 30
10.7 Human agency and oversight 30
10.8 Technical robustness, accuracy and safety 31
10.9 Bias, fairness and accessibility 31
10.10 Accountability and documentation 31
11 Use-case compliance assessment 32
11.1 MathE 32
11.2 Lifelong learning 32
11.3 Weather 32
11.4 Language data infrastructure 32
12 Governance, security and accountability 34
12.1 Current positive practices 34
12.2 Need for project-wide coordination 34
12.3 Incident and ethical escalation 34

13 Recommendations and action plan 36
14 Matters to revisit at M36 39
15 Conclusions 40
Appendix I Provisional processing and role map 41
Appendix II AI Act classification worksheet 43
Appendix III List of acronyms and abbreviations 44
Appendix IV References 45
I.1 Project sources 45
I.2 EU legislation and official guidance 45

Abstract

This deliverable provides an interim ethical and data compliance assessment of DataGEMS at Month 18 (M18). It is based on the Grant Agreement, the M18 ethics questionnaire responses, the M10 technical architecture reports, the pilot requirements report and official legal and ethical sources. The evidence indicates positive design work and several safeguards, but also material matters that require confirmation, consolidation or further implementation before the final assessment in D1.9 at M36.

DataGEMS processes or may process several distinct categories of data. These include dataset content and metadata, user account and identity-federation data, natural-language queries and conversation histories, access and service logs, user profiles and inferred interests, pilot-participant data and, in the MathE use case, student identity and learning-performance data. The lifelong learning design may involve personal or organisational skill profiles. Weather data are principally environmental and geospatial, although real-time services, accounts, logs or media-derived content may introduce personal data. The current language pilot uses historical and openly licensed corpora reported not to contain personal data relating to living individuals, while the platform is designed to support restricted user uploads.

The architecture contains useful compliance-enabling features. These include role-based access control, identity federation, token-based service authentication, protected or local servers, separation of external and internal identifiers, user-managed conversation history, audit and monitoring functions, human-in-the-loop query disambiguation, teacher validation of generated educational content, provenance explanations and a draft legal and ethical dataset metadata schema. Several respondents report DPO consultation, consent materials, pseudonymisation, DPIA activity and incident-reporting routes. These practices should be retained and converted into consistent, documented project-wide controls.

The main M18 data-protection issue is not evidence of a specific breach, but fragmentation and uncertainty. The materials do not provide a consolidated record of processing operations, controller/processor role allocation, privacy notices, platform terms of use, Article 26 arrangements, processor agreements, retention schedule, transfer register, completed DPIAs or vendor due-diligence records. Legal bases are sometimes listed cumulatively without mapping each purpose to one basis. Retention periods differ by data type and respondent and are not always confirmed as implemented. ARC reports a six-month retention period for component-level query logs and seven days for logs sent to the DataGEMS Monitoring component, with longer retention for critical security events. Its updated response nevertheless identifies final validation of the retention approach as an open point. Another respondent reports seven-day purging for transient service logs but longer, unspecified retention for authentication, accounting and audit data. These matters should be resolved through a data-inventory and retention exercise rather than by selecting a single period for all logs.

The treatment of anonymisation requires clarification. IPB describes replacement of names with codes and retention of the mapping separately. CNRS describes dictionaries, hashing and entropy-based methods. ZHAW reports that data used by it are fully anonymised. Where additional information exists or re-identification remains reasonably possible, the data remain personal data in pseudonymised form. The consortium should therefore avoid using "anonymous" as a general label unless a documented, context-specific re-identification assessment supports that conclusion. Pseudonymisation is a valuable safeguard but does not remove GDPR obligations.

DPIA information may refer to different scopes. CNRS, ZHAW and ARC report that a DPIA has been conducted, while IPB reports that a MathE DPIA is planned. No DPIA documents were supplied. The appropriate conclusion is not that the project has or has not completed a single

project-wide DPIA. Instead, the consortium should identify each high-risk processing operation, verify which DPIA covers it, record the controller and decision and conduct or update assessments where necessary. Particular attention is warranted for student profiling and adaptive learning, rich user profiling, systematic logging and possible future use of skills information in employment contexts.

DataGEMS uses or plans to use AI systems for dataset profiling, semantic retrieval and ranking, recommendations, query disambiguation, text-to-SQL, user profiling, quiz recommendation and MCQ generation. Under the EU AI Act, prohibited-practice and AI-literacy provisions have applied since 2 February 2025 and general-purpose AI governance obligations since 2 August 2025. As at the cut-off date, most remaining provisions are scheduled to apply from 2 August 2026 under Regulation (EU) 2024/1689. The Council gave final approval to the AI Omnibus on 29 June 2026. The revised timetable sets 2 December 2027 for stand-alone Annex III high-risk systems and 2 August 2028 for high-risk systems embedded in regulated products. Article 50 transparency obligations remain applicable from 2 August 2026.

AI classification must be based on the intended purpose of each system, not on the sector label alone. A tool used in education is not automatically high-risk. MathE functionality could fall outside Annex III where it merely recommends optional learning material or drafts content for teacher approval. It could become high-risk if its intended purpose is to evaluate learning outcomes, determine the appropriate level of education or materially influence access or progression in the way described in Annex III. Lifelong learning functions may be outside Annex III when used voluntarily by individuals, but employment-related ranking, task allocation, promotion, termination or performance monitoring would require separate high-risk analysis. Dataset search and profiling are generally lower-risk, although transparency, data protection, accuracy and human oversight remain relevant. Generative and conversational functions should also prepare for Article 50 transparency duties, including clear disclosure of AI interaction and appropriate marking or labelling of generated content where applicable.

The seven requirements for Trustworthy AI provide a proportionate framework even where an AI system is not legally high-risk: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. DataGEMS already has positive foundations in human validation, source grounding, explanations and modular testing. The M18 evidence does not yet show a consistent ALTAI-style assessment, AI inventory, model/vendor register, fairness metrics, environmental metrics or formalised override and contestability procedures across components.

The highest-priority actions are to: establish a project-wide processing and AI inventory; map controller, joint-controller and processor roles by operation; publish platform privacy information and terms before broader use; resolve purpose, legal basis and retention for query and profile data; verify DPIA scope and status; formalise rights handling across distributed services; document international transfers and third-party model services; define an AI classification and transparency procedure; implement role-based AI literacy; and create a consolidated ethics and data compliance action register for monitoring to M36. These actions are proportionate to the project's stage and will help convert existing good practice into demonstrable accountability.

1 Introduction

1.1 Background

DataGEMS is a Horizon Europe project developing a platform for profiling, discovering, linking, querying, analysing and reusing heterogeneous data. Its formal design covers structured, unstructured, historical and real-time sources and includes natural-language interfaces, recommendations, machine-learning models and reusable analytical patterns. The Grant Agreement requires the action to be carried out in line with the highest ethical standards and applicable law and requires beneficiaries to process personal data in compliance with the GDPR and other applicable rules.
D1.8 is the first ethics compliance deliverable and establishes an M18 baseline. D1.9 at M36 should assess the implemented platform, completed pilots and evidence of operation. This report therefore distinguishes binding duties, non-binding ethical principles, good practice, matters under development and matters for which further information is required.

1.2 Nature of the assessment

The report evaluates the information reviewed and does not replace the responsibilities of controllers, processors, deployers, providers, data stewards, DPOs, ethics bodies or competent authorities. Data protection and AI Act roles are functional and may differ between a project experiment, a partner-operated platform component, a pilot institution and a future external adopter.
No material security incident or established unlawful processing was reported in the questionnaire responses. The appropriate M18 approach is risk-based: recognise controls already in place, resolve material ambiguity and create a traceable compliance record before scale-up.

2 Project background

2.1 Consortium and work package context

DataGEMS runs from 1 January 2025 to 31 December 2027 and is coordinated by ARC. The consortium includes academic institutions, research centres, public infrastructure organisations and companies across the EU and Switzerland. The technical architecture is distributed across dataset profiling, dataset discovery, user assistance, data and model management, platform integration and pilots.

The architecture creates distributed responsibility. A use-case provider may determine why pilot data are processed, a technical partner may determine important implementation means, CITE provides identity and access infrastructure and external providers may host models or compute. The actual purposes, essential means, instructions, contracts and user relationships must be examined operation by operation.

2.2 M18 implementation context

D2.1 describes profilers that may inspect dataset content and generate semantic metadata using statistical methods, LLMs and OCR. D3.1 describes in-dataset and cross-dataset discovery, ranking, text-to-SQL and use of external services such as Wikidata, OpenStreetMap and model endpoints. D4.1 describes query recommendation, user profiles, logs, disambiguation, dataset recommendation and provenance explanations. D5.1 describes the MoMa catalogue, user profiles, query and access logs, orchestration, role-based access and storage. D7.1 translates these capabilities into four use cases involving education, skills, weather and language data.

At M18, some components are prototypes and some functions are planned. The data protection and AI assessment must therefore cover both existing processing and reasonably concrete planned processing, while clearly marking future functionality such as personalised dataset recommendations or broader external pilots.

3 Purpose and scope

D1.8 assesses the ethical aspects and data implications of DataGEMS, including fundamental rights, Trustworthy AI, the GDPR, non-personal data, data sharing and the AI Act. It covers platform-level and use-case-level activities where supported by evidence. It examines processing purposes, data categories, roles, legal bases, transparency, rights, minimisation, retention, security, DPIAs, pseudonymisation, transfers, AI classification, human oversight, fairness, explainability, accountability and environmental considerations.
The report does not attempt to determine every applicable national-law derogation for research or every intellectual-property issue in a dataset. National law and dataset-specific licences remain relevant and should be assessed by the responsible institution. Copyright and text-and-data-mining issues are included where they are material to language-data and profiling activities.

4 Methodology and sources reviewed

4.1 Source hierarchy

The Grant Agreement was treated as the primary source for project obligations and scope. The responses to the D1.8 questionnaire, designed by the External Ethics Advisor and distributed among the partners, were used as evidence of the M18 position reported by ARC, CITE, CNRS, IDS, IPB/MathE, UNIVR and ZHAW. Architecture and pilot reports were used to understand data flows and intended functions. The D1.6 questionnaire was reviewed for cross-report consistency, particularly where respondents described AI, consent, user interaction or expected use. Official EU legislation, Commission materials, EDPB guidance and official national sources were reviewed as well.

4.2 Assessment method

For each material activity, the analysis seeks to connect: the processing or AI function; the applicable legal or ethical requirement; the evidence available; the current M18 position; the risk or uncertainty and a practical action. Findings are expressed in qualified terms where the underlying documentation was not supplied.
The report distinguishes:

- **Binding legal obligation**, where an applicable law imposes a duty on the relevant actor.
- **Ethical principle**, where the requirement derives from fundamental-rights or Trustworthy AI frameworks and may also inform legal interpretation.
- **Recommended good practice**, where the action supports accountability or reduces risk but is not necessarily mandated in every case.
- **Under development**, where the project documents describe planned functionality or controls.
- **Further information required**, where questionnaire evidence is missing, inconsistent or not accompanied by documentation.

4.3 Questionnaire coverage

Seven identifiable responses were reviewed: ARC, CITE, CNRS, IDS, IPB/MathE, UNIVR and ZHAW. The responses provide useful coverage of platform coordination, discovery, identity and access, user assistance, data management, MathE and language data. No direct D1.8 response was identified from NOA, DHBW, SCAYLE, IG or UPCité. Non-response is not treated as a compliance gap; it limits partner-specific evidence, particularly for weather, lifelong learning, model hosting and some operational services.

4.4 Documentary limitations

The following supporting documents were not included in the reviewed corpus: a consolidated record of processing activities; privacy notices and platform terms; consent forms; completed DPIAs; legitimate-interest assessments; controller-to-controller or joint-controller arrangements; processor agreements; transfer impact assessments; a vendor and model register; security test or incident records; an approved retention schedule; and a consolidated AI inventory and classification record. Respondents refer to some of these materials, but this report could not verify their scope, approval, date or implementation.
CNRS, ZHAW and ARC report DPIA activity, while IPB reports that a MathE DPIA is planned; no supporting DPIA documents were supplied, so their precise scope and status could not be

independently verified. IDS reports no AI/ML in its own work, while technical documents show AI-supported language profiling and retrieval implemented by other partners. ZHAW reports fully anonymised data, while other respondents describe coded or hashed student data. These differences are analysed in this report D1.8.

5 Legal and ethical framework

5.1 Grant Agreement and fundamental rights

Article 14 of the Grant Agreement requires the action to follow the highest ethical standards, applicable EU, international and national law and basic EU values. Article 15 requires personal data to be processed lawfully, fairly, transparently, for specified purposes, with minimisation, accuracy, storage limitation and appropriate security.
The Charter of Fundamental Rights is relevant to privacy and data protection, non-discrimination, education, freedom of expression and information, academic freedom, workers' rights, environmental protection, the rights of the child and effective remedy. Not every DataGEMS function engages each right to the same degree. The rights most directly engaged at M18 are privacy and data protection, equality and non-discrimination, access to information, education, academic freedom and effective contestability of automated assistance.

5.2 GDPR

The GDPR applies where personal data are processed in the context of EU establishments or where its territorial scope is otherwise met. Core requirements include lawfulness, fairness and transparency; purpose limitation; minimisation; accuracy; storage limitation; integrity and confidentiality and accountability. Controllers must identify a legal basis for each purpose, provide information to data subjects, facilitate rights, implement data protection by design and by default, maintain records where required, use appropriate processors and conduct DPIAs where processing is likely to result in high risk.
Research status does not create a general exemption. Article 89 provides for appropriate safeguards, particularly minimisation and measures such as pseudonymisation, and Member State law may contain specific provisions or derogations. The applicable controller must assess the concrete processing, legal basis and national framework.
Controller, joint-controller and processor roles are functional. The EDPB's final Guidelines 07/2020 emphasise that roles depend on who determines purposes and essential means. Consortium membership or task allocation does not alone decide the role. Where partners jointly determine why and how pilot data are used, Article 26 analysis may be necessary.

5.3 Pseudonymisation and anonymisation

Pseudonymised data remain personal data where the information can be attributed to a person using additional information. Hashing, coded IDs and separate mappings are common safeguards but do not automatically anonymise data. Anonymisation requires that identification is not reasonably likely, taking account of all means reasonably likely to be used by the controller or another person. The EDPB's Opinion 28/2024 stresses that anonymity of AI models and datasets requires a case-by-case assessment and that data protection by design is essential.

5.4 Non-personal and mixed data

Regulation (EU) 2018/1807 seeks to reduce unjustified localisation restrictions for non-personal data and supports availability and portability. Where a dataset contains both personal and non-personal data and the elements are inextricably linked, the GDPR applies to the personal-data part and prevails in the event of conflict. DataGEMS should therefore avoid categorising a whole

dataset as non-personal merely because most fields are environmental, organisational or technical.
The Data Governance Act applies to specified data-intermediation and reuse contexts and promotes trusted sharing. The Data Act has applied since 12 September 2025 and includes rules relevant to access, contractual fairness and switching between data-processing services. DataGEMS may not fall within every provision, but open interfaces, exportability and avoidance of vendor lock-in are relevant good practice for its data-management and cloud dependencies.

5.5 AI Act

Regulation (EU) 2024/1689 entered into force on 1 August 2024. Prohibited-practice rules, the AI-system definition and AI-literacy obligations have applied since 2 February 2025. Governance and general-purpose AI model obligations have applied since 2 August 2025. Under the Regulation as enacted, most other provisions are scheduled to apply from 2 August 2026, with product-related high-risk rules later.
At the time of preparation of this report D1.8, the Digital Omnibus on AI had reached an advanced but incomplete legislative stage. The European Parliament gave final approval on 16 June 2026. The Council gave final approval to the AI Omnibus on 29 June 2026. The revised timetable sets 2 December 2027 for stand-alone Annex III high-risk systems and 2 August 2028 for high-risk systems embedded in regulated products. Article 50 transparency obligations remain applicable from 2 August 2026.
The AI Act classification is based on intended purpose. Annex III includes certain systems used in education and vocational training, notably for access or admission, assignment, evaluation of learning outcomes, assessment of the appropriate level of education and monitoring prohibited behaviour during tests. It also includes specified employment and worker-management uses. A recommendation system does not become high-risk solely because it is used by a university or HR department; its intended function and effect must match the listed use.
Article 4 requires providers and deployers to take measures to ensure sufficient AI literacy of staff and others dealing with AI systems on their behalf, considering technical knowledge, experience, context and affected persons. This obligation already applies. Article 50 transparency obligations for certain interactive, generative and manipulated-content systems are scheduled to apply from 2 August 2026. The Commission published a final Code of Practice on marking and labelling AI-generated content on 10 June 2026, but it was still undergoing adequacy assessment. While the code is voluntary, the legal duties are not.
DataGEMS partners that merely use third-party general-purpose models are generally deployers or downstream providers rather than providers of the underlying GPAI model, unless they develop or place such a model on the market under their name or make substantial modifications that change legal role. They should nevertheless perform supply-chain due diligence, respect instructions, assess data flows and understand whether a component placed into service becomes an AI system for which they are provider.

5.6 Trustworthy AI and ethics-by-design

The High-Level Expert Group's Ethics Guidelines for Trustworthy AI and ALTAI are non-binding. They organise assessment around seven requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being and accountability. These requirements remain useful for lower-risk AI systems and for project research before all AI Act duties apply.
The Horizon Europe ethics-by-design approach supports continuous assessment from research design through deployment. For DataGEMS, this means recording intended purpose, affected

groups, data provenance, model dependencies, foreseeable misuse, human review, uncertainty, incident response and changes between pilot and operational use.

5.7 Copyright, database rights and text and data mining

Language corpora, educational resources, PDFs, media content and database contents may be protected by copyright, related rights or database rights. Directive (EU) 2019/790 provides text-and-data-mining exceptions subject to conditions, including lawful access and, for the general exception, rights reservation. National implementation differs. Open licences and public-domain status should be verified rather than inferred from availability.

5.8 Accessibility and national law

The European Accessibility Act has applied to covered products and services from 28 June 2025, and the Web Accessibility Directive applies to public-sector websites and mobile applications. Whether a particular DataGEMS deployment falls within scope depends on the operator and service. Accessibility is in any event an ethical and adoption requirement for educational, public-sector and EOSC-facing services.

National laws supplement the GDPR and implement copyright, research and employment rules. Relevant jurisdictions include Greece, Italy, Portugal, France, Germany, Spain and Poland, with Swiss law relevant to ZHAW. The responsible partner should identify the national rules applicable to its controller activities, research institution, employment setting and pilot population.

6 M18 data processing and AI landscape

6.1 Platform identity, access and logs

CITE reports that the authentication and authorisation infrastructure stores identity-federation information including name, external identifiers and email. It processes logs, accounting records, session characteristics, operational metadata, request forwarding and search-conversation history for authentication, authorisation and usage accounting. External identifiers are retained within the AAI, while internal components receive a new internal user identifier. CITE reports seven-day purging of transient service logs, longer retention for authentication, audit and accounting information and user control over conversation history.
D5.1 describes a broader architecture for access and query logs containing user or session IDs, login times, natural-language query text, translated queries, execution status, datasets, errors and, where local rules allow, device and IP information. It also describes user profiles containing role, affiliation, ORCID-derived interests, explicit feedback and inferred preferences. The difference between architectural possibility and active collection should be documented. Data protection by default requires unnecessary fields to remain disabled rather than collected merely because the schema supports them.

6.2 Discovery, profiling and recommendation

ARC reports that natural-language queries, timestamps, session identifiers and query logs are personal data. Cross-dataset discovery can operate without a persistent user profile, but queries may themselves contain names, health information, political opinions or other personal or special-category content. Structured profiling is intended to profile datasets rather than people, but the profiler may inspect personal data contained in uploaded datasets. The legal position therefore depends on the source dataset and processing purpose, not only on the component label.
D4.1 describes future personalised query and dataset recommendation using profiles and prior queries. ARC states that its current dataset recommender is content-based and does not use personal data beyond authorisation, while future versions may use user history or profiles. This is a meaningful design distinction and should be reflected in versioned privacy documentation and DPIA screening.

6.3 MathE

CNRS and IPB report processing of student learning paths, performance and platform data for adaptive quizzes and learning-resource recommendations. IPB identifies name, email, school, birth year, gender and country as collected fields and reports use of coded IDs with a separate mapping. CNRS reports local protected storage, project-limited sharing and deletion at project end, with consent allowing withdrawal and deletion.
The technical functions include reinforcement-learning quiz composition, LLM-based MCQ generation and machine-learning recommendation using decision trees, random forests or clustering. Teachers or administrators review generated questions and can influence recommendations. These human controls are positive, but the intended purpose and actual role of the recommendation in assessment or progression require formal definition.

6.4 Lifelong learning

D7.1 anticipates individual and organisational skill models, role comparison and training recommendations. Specific datasets were not fixed at that stage. The D1.8 questionnaire corpus does not contain a direct lifelong-learning provider response. The processing may range from non-personal taxonomy matching to detailed worker profiles. If real employee data are used, purpose, power imbalance, accuracy, access, objection, employment-law rules and Annex III classification become significant. Pilot design should not proceed on the assumption that consent is freely given in an employment relationship, due to the imbalance of power.

6.5 Weather

The core weather datasets are environmental and geospatial, including ERA5-Land, historical stations, real-time stations, elevation, land cover and population density. Some real-time station services are restricted. User accounts, queries and logs remain personal data. Planned media crawling may process content about identifiable persons and should have source, purpose, retention and rights analysis. No direct NOA D1.8 response was supplied, so the current data flows and operational arrangements should be confirmed before pilots expand.

6.6 Language data

IDS reports that the current pilot uses open-domain historical textual datasets and no personal data relating to living individuals. It reports open licences or public-domain material and local or project-protected servers. It also notes that the platform may support restricted uploads under the data owner's responsibility.
The statement that IDS itself does not use AI/ML is role-specific. D2.1, D3.1 and the D1.6 response describe LLM-supported profiling, multilingual retrieval, embeddings and generated summaries for the language use case. The project AI inventory should therefore identify the technical partner responsible for each AI component even where the use-case provider does not implement it.

7 Fundamental rights and ethical assessment

7.1 Privacy and informational self-determination

The platform may observe what users search, which datasets they access, what interests are inferred and how students perform. These records can reveal professional interests, research topics, education level or sensitive personal circumstances. The ethical issue is not limited to confidentiality; persistent profiles can influence what users are shown and may create self-reinforcing assumptions. DataGEMS should collect only the information needed for a defined function and provide a non-personalised route where feasible.
Free-text queries deserve specific controls. Users may enter special-category data even if the system does not request them. The interface should warn against unnecessary personal data, apply filtering or redaction where proportionate, restrict log access and avoid reusing queries for benchmarks without a compatible purpose, legal basis and de-identification assessment.

7.2 Equality, fairness and non-discrimination

Data and model bias may affect dataset ranking, educational recommendations, skill matching and multilingual retrieval. Content-based systems can reproduce gaps in available metadata. Language tools may perform better for English or contemporary sources than lower-resource languages or historical variants. Student recommendations can be influenced by past performance and demographic proxies. Skills taxonomies may encode occupational assumptions and disadvantage non-standard career paths.
Fairness requires use-case-specific analysis. A single aggregate accuracy score is insufficient. The project should test performance by language, topic, data format and user group, document known limitations and involve domain experts and affected users. Where demographic data are not lawfully available, fairness can still be assessed through synthetic cases, qualitative review, error analysis and monitoring of proxy effects.

7.3 Education and human agency

Personalised learning can support education, but learners should not be reduced to an inferred level or automated pathway. Teachers should retain authority over pedagogical content and students should be able to request explanation or alternative material. The MathE evidence of teacher validation and physical support during pilots is positive. The project should define which decisions the system may make, which require teacher approval and which must never be automated.
If pilot participants include minors, child-specific transparency, consent or authorisation, safeguarding and data minimisation require additional attention. The reviewed material refers to higher education and does not establish participant ages. This should be confirmed rather than assuming that all participants are adults.

7.4 Access to information and academic freedom

DataGEMS can improve access to data and evidence, but ranking systems also shape visibility. The platform should avoid presenting the top-ranked dataset as the only valid source, explain relevance and support diverse or alternative results. Researchers should be able to inspect original data, metadata and transformations and to use the platform without surrendering control over research questions or unpublished interests.

7.5 Misuse and over-reliance

Weather outputs may be mistaken for operational warnings, generated MCQs for validated educational content and skill-gap results for objective employee assessments. The system should communicate intended use, limitations and the difference between exploratory assistance and authoritative decision support. High-consequence deployments require integration with institutional procedures, named human responsibility and documented override.

7.6 Societal and environmental well-being

DataGEMS may support public-interest research, climate literacy and education, but compute-intensive LLM and embedding workflows create cost and environmental impacts. The materials do not contain energy or carbon metrics. The consortium should monitor model choice, inference volume, hardware, caching and the necessity of generative processing. Environmental monitoring is recommended good practice and aligns with Trustworthy AI even where not a specific legal duty.

8 GDPR and data compliance assessment

8.1 Roles and accountability

No consolidated role map was supplied and the likely arrangements vary. A pilot institution that determines why student or worker data are used may be a controller. A technical partner operating a component on documented instructions may be a processor for that use, while acting as a separate controller for security logs or independent research. Partners that jointly design the purpose and essential means of a pilot may be joint controllers. CITE may be a processor for authentication functions or a controller for specific security and accounting purposes depending on the agreed arrangements.
The consortium should create an operation-level role table covering at least account management, query logging, user profiling, dataset onboarding, each pilot, model-service calls, dissemination and evaluation. For joint control, Article 26 arrangements should allocate transparency and rights duties while making the essence available to data subjects. For processing on behalf of another party, Article 28 terms should cover instructions, confidentiality, security, subprocessors, assistance, deletion or return and audits.

8.2 Categories, sources and special-category data

The reviewed evidence identifies identity, contact, account, session, log, query, educational, demographic, organisational skill and inferred-preference data. Sources include users, federated identity providers, MathE, pilot institutions, public profiles such as ORCID and system-generated observations. Dataset uploads may contain additional personal data not anticipated by platform designers.
The onboarding process should require uploaders to declare personal-data and special-category status, source, controller, purpose, lawful basis, permitted users, retention and restrictions. Automated scanning may support but should not replace uploader accountability. Free-text queries and user-uploaded datasets may contain special-category or criminal-offence data; the platform should restrict unnecessary collection and require additional Article 9 or Article 10 analysis where processing is intentional.

8.3 Purpose limitation and legal bases

Questionnaire responses refer to consent, contract, legitimate interests and educational service necessity, but these cannot be treated as interchangeable. Each processing purpose should have one clearly reasoned Article 6 basis, with an Article 9 condition where applicable. Operational authentication may be necessary to provide a requested service. Security logging may rely on legitimate interests or a legal obligation depending on the actor. Newsletter subscriptions may rely on consent. Research use of student performance may require consent, public task or legitimate interests depending on institution, national law and safeguards. Consent must be freely given, specific, informed, unambiguous and withdrawable without detriment. In education or employment, imbalance may make consent inappropriate for core processing. An "umbrella" form may support several related experiments only if purposes are sufficiently specific and new incompatible uses trigger renewed information or consent. Withdrawal cannot always require deletion where another lawful basis or legal obligation applies, but this must be explained.

8.4 Transparency and platform terms

CITE identifies the absence of a platform privacy policy and terms of use as a known gap at M18. This is a priority before broad external use. The information should identify controllers and contact points, purposes and bases, categories, recipients, international transfers, retention, rights, automated functions, whether query and profile data improve services and how AI interaction and generated content are indicated.
Layered notices are recommended. A general platform notice should be supplemented by pilot-specific information for MathE, skills and other processing. Just-in-time notices should appear when a user enables personalisation, uploads data, submits content to an external model service or joins a study. Terms of use should not be used to obtain blanket consent or transfer statutory responsibilities to users.

8.5 Data subject rights

Respondents state that users can request access, rectification, erasure, restriction, objection and portability and can unsubscribe or delete conversation history. In a distributed architecture, rights require orchestration across AAI, logs, profiles, MoMa, pilot databases, model-service records and backups. CITE notes that requests must be implemented by individual service providers across intermediate storage layers.
The consortium should define a single intake route and an internal workflow that identifies the responsible controller, verifies identity, locates data, propagates the request, records exemptions and responds within the legal period. Users should be able to correct inferred interests and student or skill-profile information, disable personalisation and challenge consequential recommendations. Where data are held in backups, the response should explain the applicable restoration or deletion process.

8.6 Data minimisation and privacy by design

Positive evidence includes query-only discovery without persistent identifiers, content-based recommendation before personalisation, internal identifiers separated from external identity data, local storage of MathE data, role-based access and user control over conversation history. These measures should become default design requirements.
Areas requiring attention include broad user-profile schemas, optional device and IP information, persistent query histories, demographic fields in MathE and the possibility of profiling uploaded datasets containing personal data. The controller should justify each field, separate mandatory from optional data and avoid collecting attributes merely because they may be useful later. Gender, birth year and country should be evaluated for necessity and potential bias. Where analysis can use aggregated or pseudonymised performance data, direct identifiers should remain separate.

8.7 Accuracy

Accuracy applies to identity and profile data and to inferred attributes that may affect users. Users should be able to correct account and profile information. Inferred interests, skills or learning levels should be marked as inferences, dated, linked to source observations and subject to review. Outdated profiles should not continue to drive recommendations without checks.
For generated educational and analytical content, GDPR accuracy may apply where output is personal data, while general technical and ethical accuracy applies more broadly. The project should monitor hallucination, semantic SQL errors, false links and recommendation quality and should prevent corrected errors from recurring.

8.8 Storage limitation and retention

The evidence does not support a single project-wide retention period. CITE reports seven-day purging for transient logs but longer retention for authentication, audit and accounting. CNRS refers to deletion at project end for MathE data. ARC reports six months for component-level query logs and seven days for logs sent to the DataGEMS Monitoring component, with longer retention for critical security events; its updated response nevertheless identifies final retention validation as an open point. IPB refers to active-account or service-necessity retention and secure deletion.
A retention schedule should differentiate: active account data; authentication and security logs; service diagnostics; query and conversation history; user profiles; research datasets; consent records; study data; evaluation results; backups; and published or archived research outputs. Each period should have a purpose, trigger, owner, deletion method and exception. Technical configuration should enforce the schedule and deletion should be testable.

8.9 Security

The reports describe protected servers, authentication, authorisation, bearer tokens, JWT validation, role-based access, gateway controls, encrypted transport, logging, backups, SSH and regular maintenance. CITE provides AAI and internal identifiers. D5.1 describes audited, gateway-mediated access to the Airflow-based workflow orchestrator. These are positive architectural controls.
Further evidence would be beneficial on encryption at rest, key management, least-privilege reviews, secrets management, vulnerability scanning, dependency patching, penetration testing, backup restoration, incident response, breach assessment and subprocessors. Security should be risk-based and documented under Article 32. Pilot data should not be moved to shared or external services without confirming access, purpose and transfer conditions.

8.10 DPIA and risk assessment

A Data Protection Impact Assessment (DPIA) is required where processing is likely to result in high risk, taking account of nature, scope, context and purpose. Relevant indicators may include systematic evaluation or profiling, automated decisions with significant effects, sensitive data, vulnerable persons, large-scale monitoring, novel technology, data matching and barriers to rights.
CNRS, ZHAW and ARC report DPIA activity. IPB reports that a MathE DPIA is planned but not yet conducted. These statements may concern different institutions, datasets or components. The consortium should create a DPIA register identifying processing, controller, screening decision, document date, approval, residual risk and review trigger. MathE student profiling, rich user profiling, query-log reuse and future workforce analytics should receive specific screening. A DPIA should be completed before high-risk processing begins or materially changes, not only after pilot deployment.

8.11 International transfers and external services

The consortium includes ZHAW in Switzerland. The European Commission recognises Switzerland as providing adequate protection, so EU-to-Switzerland transfers within the scope of the adequacy decision do not require additional Article 46 safeguards. The actual transfer, role and security arrangements must still be documented.
ARC reports use of Hugging Face models, AWS Bedrock and a SCAYLE LLM service. D3.1 also refers to external APIs such as Wikidata and OpenStreetMap. The consortium should map whether personal data, confidential dataset content, prompts or identifiers are transmitted,

where processing occurs, which provider terms apply, whether data are retained or used for training and which transfer mechanism is relied upon. External service calls should use minimised prompts and, where feasible, local or EU-hosted models for sensitive content.

8.12 Anonymisation, pseudonymisation and re-identification

IPB's replacement of names with IDs and separate mapping is pseudonymisation. CNRS's non-correlated dictionaries and hashing may be pseudonymisation unless the mapping is destroyed and re-identification is not reasonably likely. Hashing names without secret salt may be vulnerable to dictionary attacks; CNRS reports entropy-based methods intended to reduce this risk. ZHAW's statement that all data are fully anonymised should be supported by a documented assessment of data fields, uniqueness, linkage and recipient knowledge.
Student performance, country, school, birth year, topic history and timestamps may permit singling out even after names are removed. Re-identification risk should be assessed across combined datasets and outputs. Public release should use stronger aggregation or anonymisation than internal controlled research use.

8.13 Data sharing and governance

D2.1's legal and ethical metadata schema can support data sharing by recording controllers, licences, recipients, purposes, retention and safeguards. It should be integrated with access policy and not merely displayed. The onboarding workflow should prevent publication of restricted data and should enforce conditions on derived profiles, embeddings and analytical outputs where necessary.
Data-sharing arrangements should distinguish raw data, metadata, generated profiles, model inputs, evaluation data and publications. A data provider may allow profiling but not redistribution or model training. Rights and responsibilities should be explicit and machine-readable where possible.

9 Free flow of non-personal data and wider data governance

DataGEMS is designed for both personal and non-personal data. For non-personal data, the free-flow Regulation discourages unjustified localisation restrictions and supports portability and availability to authorities. The project should avoid unnecessary location constraints while allowing security, research and contractual requirements that are objectively justified.
The Data Act's switching and interoperability policy is relevant to the platform's cloud and model dependencies. DataGEMS should maintain exportable datasets, profiles, logs and analytical patterns in commonly used formats, document APIs and dependencies and provide exit procedures. This reduces lock-in and supports the project's EOSC sustainability objectives.
The Open Data Directive and Data Governance Act may be relevant where public-sector or protected data are reused. They do not displace the GDPR, confidentiality, intellectual property or sectoral law. Dataset-level legal metadata should record the specific reuse regime rather than labelling all public-source data as open.

10 Artificial intelligence assessment

10.1 Provisional AI inventory

The evidence supports the following provisional inventory:

Table 1 Provisional Inventory

Function	Technique or model evidence	Likely project role	Main legal and ethical considerations
Structured and unstructured profiling	LLM-generated keywords, summaries and descriptions; OCR; statistical profiling	Provider/deployer of a profiling AI system, depending integration	Personal data in uploaded content, accuracy, copyright, transparency of inferred metadata
Cross-dataset discovery	Embeddings, semantic retrieval, re-ranking and LLM question enrichment	Provider of integrated retrieval system; deployer of underlying models	Ranking bias, query logs, source grounding and vendor data flows
In-dataset discovery and text-to-SQL	LLM semantic parsing, query planning and geospatial entity extraction	Provider/deployer	Semantic errors, security of generated queries, explainability and human validation
Query disambiguation	LLM suggestions and self-reasoning with human selection	Provider/deployer	User control, misleading explanations, query data and AI-interaction disclosure
Query and dataset recommendation	Deep learning, metadata embeddings and possible future user history	Provider/deployer	Profiling, fairness, transparency, opt-out and future DPIA
User profiling	LLM or deep-learning inference of interests and preferences	Provider/deployer	Purpose, legal basis, accuracy, contestability and persistent surveillance
MathE MCQ generation	Llama and Qwen or other open-source LLMs	Provider of integrated educational content-generation system; deployer of underlying model	Accuracy, pedagogical validation, Article 50, copyright and potential Annex III use
MathE quiz recommendation	Reinforcement learning, random forest, decision trees and clustering	Provider/deployer	Intended purpose, student profiling, fairness and possible education high-risk classification

Lifelong learning skill matching	Profiling, ranking and recommendation planned	Provider/deployer	Employment context, discrimination, explainability and possible Annex III classification
Weather anomaly detection and media extraction	Time-series anomaly detection, NLP and NER planned	Provider/deployer	False alerts, source quality, operational over-reliance and personal data in media
Language profiling and retrieval	Multilingual LLMs, embeddings, summarisation and semantic-change analysis	Provider/deployer by technical partner	Language bias, copyright, cultural context, compute and provenance

This inventory should be validated by component owners and versioned. It should identify model name and version, provider, hosting, training or fine-tuning, input and output data, intended purpose, affected persons, project role, applicable instructions and change history.

10.2 Prohibited practices

No reviewed evidence indicates intended use of prohibited practices such as manipulative subliminal techniques causing significant harm, social scoring by public authorities, prohibited biometric categorisation, untargeted facial-image scraping or emotion recognition in education or workplaces. The project should maintain this boundary. In particular, it should not infer emotions from students, workers or users without a separate legal and ethical assessment; certain education and workplace emotion-recognition uses are prohibited.

10.3 High-risk classification

The CNRS questionnaire states that MathE is high-risk because education is a high-risk area. IPB states that no formal AI Act assessment has been completed and that the system recommends content without determining access or significantly affecting rights. The latter reflects the correct need for intended-purpose analysis.
For MathE, optional content generation, hints and teacher-approved recommendations may be outside Annex III. Classification should be revisited if the system evaluates learning outcomes, determines a student's educational level, controls progression, assigns access or monitors prohibited behaviour during tests. Even outside Annex III, transparency, human oversight and data protection remain required.
For lifelong learning, self-directed training discovery is likely outside the employment high-risk category. Use by HR to recruit, select, promote, terminate, allocate tasks based on behaviour or characteristics, or monitor and evaluate performance may fall within Annex III. The use case should maintain separate product descriptions and data flows for learner support and employer decision support.
Other discovery, profiling and assistance components generally appear to be lower-risk on current intended purposes, but integration into a high-risk system can affect obligations and supply-chain responsibilities. A reusable component should provide documentation enabling downstream classification.

10.4 Research and development context

The AI Act contains exclusions and adaptations for scientific research and for research, testing and development before placement on the market or putting into service, subject to conditions and without circumventing other law. DataGEMS should not rely on a broad "research exemption" for operational pilots or services used in real settings. The transition from internal research to pilot, public service or integration into MathE should be recorded, and classification should be reassessed when intended purpose or deployment changes.

10.5 AI literacy

Article 4 already applies. The reviewed evidence does not show a consortium-wide AI-literacy programme. The requirement is role- and risk-based. Developers need training on model limitations, data governance, security, documentation and the AI Act. Teachers and pilot facilitators need to understand generated-content risks, oversight and student communication. Data stewards need to understand profiling, rights metadata and model-service data flows. Users should receive concise guidance on AI interaction, errors and appropriate reliance. The consortium should record training audiences, content, attendance and updates. Reading model documentation alone may be insufficient, particularly for education and employment-related functions.

10.6 Transparency and explainability

DataGEMS has strong positive design evidence: source-grounded answers, relevance scores, human-in-the-loop disambiguation, SQL provenance and analytical-pattern explanations. Users can inspect retrieved documents and teachers review generated questions.
ARC reports that users can inspect retrieved documents, compare the properties of seed and recommended datasets and view how coverage, diversity and compatibility contribute to dataset-packaging scores. Generated MCQs and chatbot interactions should be clearly identified as AI-assisted. The system should distinguish model-generated explanations from faithful causal or provenance information. Explanations should disclose relevant datasets, important transformations, uncertainty and how to correct or challenge the output.
Article 50 is scheduled to apply from 2 August 2026. Interactive systems should inform users that they are interacting with AI unless this is obvious. Providers of systems generating synthetic content may need machine-readable marking where technically feasible. Deployers publishing AI-generated text on matters of public interest may have labelling duties, subject to exceptions. Educational MCQs and internal analytical summaries require case-specific assessment; voluntary adoption of clear labelling is recommended even where the precise duty does not apply.

10.7 Human agency and oversight

Positive controls include teacher validation, user selection between disambiguation options, assistive rather than legally determinative search and the ability to correct or disable profiles as described by respondents. These controls should be formalised by component: who reviews, what information they receive, when they can override, what happens after override and how repeated errors are addressed.
Human oversight should be meaningful. A teacher cannot validate hundreds of generated questions without time, expertise and an interface showing source material. A user cannot challenge a ranking without an explanation and alternatives. Oversight effectiveness should be tested, not assumed from the presence of a human in the workflow.

10.8 Technical robustness, accuracy and safety

The architecture includes health checks, monitoring, error rates, latency and user feedback. Text-to-SQL design includes regeneration after errors, but D3.1 recognises that semantic errors remain challenging. Cross-dataset discovery plans precision, recall and NDCG evaluation. These are sound technical foundations.

The consortium should define acceptance thresholds by use. A research search tool may tolerate lower recall than a weather safety-support function. Generated SQL should execute under least-privilege database accounts and use safeguards against injection, destructive statements and excessive queries. Model and dataset versions should be logged to reproduce outputs. Robustness testing should include adversarial prompts, ambiguous queries, missing data, language variation and out-of-distribution inputs.

10.9 Bias, fairness and accessibility

No consistent fairness assessment was provided. CNRS leaves bias and fairness largely unanswered; IPB refers to topic, difficulty and performance-based recommendation; IDS identifies high-resource language bias; ARC states that sensitive-attribute profiling is not performed. The absence of sensitive attributes does not guarantee fairness because proxies and dataset imbalance remain.

Each use case should define relevant fairness harms and tests. MathE should examine error and recommendation distribution across topics, prior attainment, country and other justified groups. Lifelong learning should examine occupational and language bias and prevent protected characteristics or proxies from driving adverse employment outcomes. Language tools should report performance by language, period and genre. Interfaces should follow accessibility requirements and be usable with assistive technologies.

10.10 Accountability and documentation

The project needs a common AI governance file containing inventory, intended purpose, provider/deployer role, classification, data sources, model providers, evaluation, oversight, incidents, changes and user information. Lower-risk systems also benefit from this record. An ALTAI-based assessment can organise the ethical evidence without being presented as legal certification.

Questionnaire responses identify DPOs or advisors at several organisations, but ethical oversight and escalation are not consistently allocated. The External Ethics Advisor can support independent review, but operational responsibility remains with the consortium and relevant controllers, providers and deployers.

11 Use-case compliance assessment

11.1 MathE

The information reviewed indicates a lawful and ethical pathway is feasible, but the processing is among the highest-priority areas for M18 follow-up. Student identity, demographic and performance data support adaptive recommendations. Consent materials and DPO involvement are reported, but legal-basis mapping, controller roles, participant age, DPIA coverage, retention and data-sharing documentation should be verified. Coded IDs are pseudonymisation, and direct identifiers should remain separated from analytical data.
Generated MCQs require teacher validation, source traceability and error reporting. Recommendations should remain supportive and should not determine access, grades or progression without a separate high-risk and educational assessment. Students should be informed about profiling, automated recommendations, data used and their ability to seek human support or alternative materials. Accessibility and different learning abilities should be included in evaluation.

11.2 Lifelong learning

The current evidence is insufficient for a definitive processing assessment because datasets and deployment context remain under development and no direct D1.8 response was identified. The project should first determine whether pilots use synthetic, volunteer, employee or organisational data. Voluntary learner self-assessment should be separated from employer-led workforce evaluation.
Where employee data are processed, consent may be inappropriate due to imbalance. Controllers should assess legitimate interests, contract, collective or employment-law bases and national rules. Workers should receive transparent information, access and correction, and meaningful human review. No adverse employment decision should be based solely on DataGEMS output. AI Act classification should be documented before any HR decision-support deployment.

11.3 Weather

Environmental and geospatial data are mainly non-personal, but account, query, log and media-extraction data may be personal. Real-time restricted station data require access and security governance. External geospatial APIs should receive minimised queries and no unnecessary user identifiers.
Ethical focus should be on accuracy, uncertainty and appropriate reliance. Outputs intended for civil protection or planning should display source, timestamp, resolution, method and limitations. The pilot should avoid representing research anomaly detection as an official warning service. A direct NOA data-flow and governance confirmation would strengthen D1.9.

11.4 Language data infrastructure

The current historical corpora are reported to be open, public domain or licensed and not to contain data about living individuals. This lowers GDPR risk but does not eliminate copyright, database-right, cultural bias or inaccurate metadata concerns. Original licences and rights reservations should be recorded at item or dataset level. Public availability should not be treated as proof of lawful reuse.

Restricted uploads require a clearer governance model than "user responsibility" alone. The platform should request rights and personal-data declarations, restrict sharing by default, provide a takedown or correction route and prevent use for training or publication unless authorised. Multilingual bias and information loss during format harmonisation should be documented and evaluated.

12 Governance, security and accountability

12.1 Current positive practices

The evidence indicates DPO consultation at CNRS/University of Grenoble Alpes, IPB and other respondents, a privacy advisor at IDS, CITE's DPO and assigned officer, protected servers, role-based access, federated identity, internal identifiers, consent materials, teacher validation, mailing lists or helpdesk routes for escalation, provenance explanations and draft legal and ethical metadata. ARC additionally reports institutional DPO support and DPIA activity, use of Argos for the DataGEMS DMP, access to the Amnesia anonymisation tool and planned use of Zenodo for suitable project output. These are valuable foundations and should be recognised.

12.2 Need for project-wide coordination

Institution-level controls do not automatically provide end-to-end accountability across a distributed platform. A project ethics and data governance group should maintain the processing and AI inventories, review new pilots and model services, track DPIAs and agreements, monitor actions and coordinate incident escalation. This function should include technical, use-case, security, data-protection and ethics expertise and should report material issues to project governance and the External Ethics Advisor.

12.3 Incident and ethical escalation

Respondents refer to helpdesk tickets, mailing lists and institutional routes, but no common procedure was supplied. The consortium should define security incident, personal data breach, AI incident and ethical concern categories; contact points; triage; evidence preservation; notification assessment; partner coordination; decision records; and lessons learned. The procedure should account for the 72-hour GDPR supervisory-authority notification period where a personal data breach is likely to risk rights and freedoms.

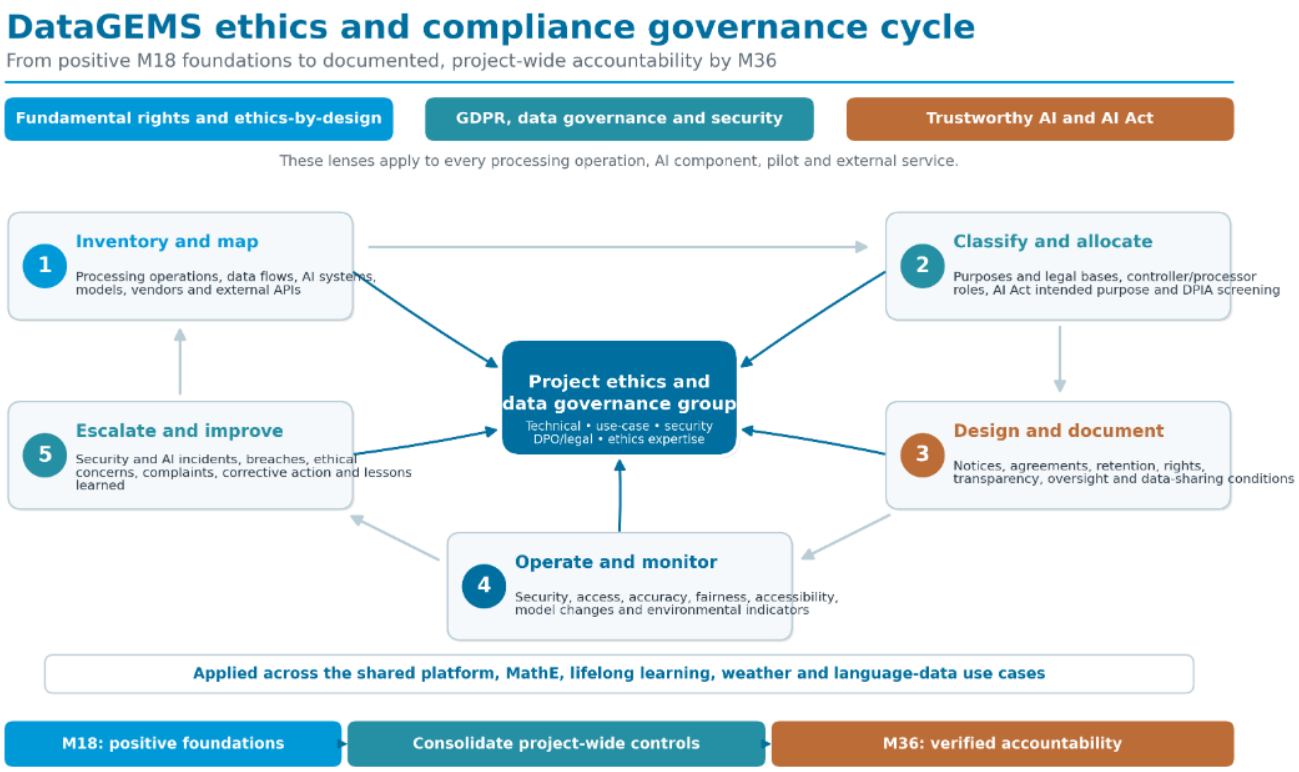


Figure 1 ***DataGEMS ethics and compliance governance cycle.*** *The figure illustrates a continuous process for identifying and mapping processing operations and AI systems; allocating legal roles and classifying risks; designing and documenting safeguards; monitoring implementation and escalating incidents or ethical concerns. The cycle should be coordinated through a multidisciplinary project ethics and data governance function and applied across the shared platform and all four use cases.*

13 Recommendations and action plan

Table 2 Recommendations and action plan

Priority	Action	Responsible function	Evidence/rationale	Target
High	Create and approve a project-wide record of processing operations and data-flow map, including current and planned processing.	WP1 coordination with component owners, use-case providers and DPOs	Current evidence is distributed and several answers remain unresolved.	Initial version within three months; quarterly review.
High	Allocate controller, joint-controller and processor roles for each operation and put Article 26 or 28 arrangements in place where required.	Relevant institutions with legal/DPO support	Roles cannot be inferred from work packages and affect notices, rights and security duties.	Before broader external pilots or data exchange.
High	Publish a layered platform privacy notice and terms of use and prepare pilot-specific notices.	Platform operator, WP1 and controllers	CITE reports these are not yet available.	Before public or external user onboarding.
High	Establish a purpose and legal-basis matrix for accounts, logs, queries, profiles, research, communication and each pilot.	Controllers and DPOs	Current responses list multiple bases without purpose mapping.	Alongside the processing inventory.
High	Approve and technically implement a differentiated retention schedule.	Platform/security owners and controllers	Seven days, six months, project end and active-account periods are reported for different or uncertain data.	Configure and test before scale-up.
High	Verify DPIA screening and status in a DPIA register; complete or update assessments for high-risk operations.	Controllers and DPOs	CNRS, ZHAW and ARC report DPIA activity; IPB plans a MathE DPIA; supporting documents and exact scopes were not supplied.	MathE, profiling and workforce screening before further pilot phases.

High	Create a versioned AI inventory and classification record, including intended purpose, provider/deployer role, model/vendor and Article 50 analysis.	Technical leads, WP1 and ethics/legal support	AI use is broad and classifications in responses are inconsistent.	Initial inventory by 30 September 2026; update on material change. This internal deadline should not delay any applicable AI Act transparency obligations.
High	Implement role-based AI literacy and record completion.	Consortium management and partner employers	Article 4 already applies.	Begin immediately and refresh annually or on major changes.
High	Formalise rights handling across AAI, profiles, logs, pilot databases and backups.	Controllers and service owners	Rights are promised but distributed implementation is not documented.	Test with a simulated request before external pilots.
High	Map external model/API data flows, contracts, retention, training use, subprocessors and transfers.	Component owners, procurement/legal and security	AWS Bedrock, Hugging Face, SCAYLE and external APIs are identified.	Before sending personal, confidential or restricted content.
Medium	Adopt a consortium anonymisation and pseudonymisation standard with re-identification assessment templates.	DPO/ethics group and data stewards	Terminology is inconsistent and coded data remain personal.	Apply to all pilot releases and public datasets.
Medium	Complete ALTAI-style assessments and use-case fairness plans for material AI systems.	AI component owners and use-case leads	No consistent HLEG assessment or fairness metrics are evidenced.	Prior to final pilot evaluation.
Medium	Make AI use, sources, uncertainty, generated content	UI, component and pilot teams	Positive explanation foundations exist but coverage is incomplete.	Include in usability and acceptance testing.

	and contestability visible in interfaces.			
Medium	Establish a formal incident and ethical-escalation procedure with breach and AI-incident paths.	Security, DPOs, project management and ethics advisor	Current routes are mainly mailing lists or helpdesk tickets.	Approve and exercise before M30.
Medium	Define rights and access controls for derived profiles, embeddings, benchmarks and analytical patterns.	Data stewards and IP/legal leads	Reuse conditions for derived artefacts are not consistently specified.	Integrate with onboarding and legal metadata.
Medium	Monitor compute and environmental indicators for LLM and multilingual processing.	Technical leads and infrastructure providers	Environmental impact is not measured and computational cost is identified as a risk.	Baseline by M30 and report at M36.

14 Matters to revisit at M36

D1.9 should verify, through documents and operational evidence:

- the final record of processing activities and role allocation for platform and pilots;
- publication and accuracy of privacy notices, terms and consent or information materials;
- implemented retention configuration and deletion tests for each data category;
- DPIA screening decisions, completed assessments and mitigation status;
- Article 26 and 28 arrangements and data-sharing terms;
- international transfer records and third-party model-service due diligence;
- response to at least one simulated or real data-subject request across distributed services;
- incident and breach procedure exercises and any incidents or near misses;
- the AI inventory, intended-purpose statements and classification decisions after final pilot design;
- implementation of Article 4 AI literacy and evidence of role-specific training;
- readiness for Article 50 transparency and any later AI Omnibus amendments in force at M36;
- performance, fairness and accessibility results by use case and user group;
- teacher correction rates, student challenge routes and whether MathE functions affect assessment or progression;
- intended use and safeguards for lifelong-learning outputs in any employment setting;
- rights, licensing and provenance controls for language corpora and user uploads;
- uncertainty and operational-use labels for weather services;
- whether "anonymised" datasets have documented re-identification assessments;
- environmental and resource metrics for major AI workflows; and
- the governance, ownership and security arrangements for post-project operation.

15 Conclusions

The information reviewed indicates that DataGEMS has incorporated several important ethical and compliance safeguards at M18, including human validation, access controls, protected storage, provenance, explanations, DPO involvement and an emerging legal and ethical metadata scheme. No material incident or clear evidence of non-compliance was reported.
The current position should nevertheless not be characterised as an unconditional finding of compliance. Processing, role allocation, notices, retention, DPIA status, model-service flows and AI classification are not yet consolidated across the project, and some questionnaire answers remain provisional or inconsistent. These are manageable issues at the M18 stage if addressed before pilot expansion and operational use.
The recommended approach is to convert distributed partner practices into project-wide accountability: a processing inventory, role map, DPIA and retention register, platform transparency, rights workflow, AI inventory, AI literacy and documented human oversight. D1.9 should then assess not only whether these documents exist, but whether controls operate effectively in completed pilots and sustainable services.

Appendix I Provisional processing and role map

Operation	Data	Possible actors/roles to confirm	Key documents or actions
Account and federated identity	Name, email, external and internal identifiers, session data	Platform operator/controller; CITE processor or controller for defined security purposes	Privacy notice, role decision, Article 28 if processor, retention and rights workflow
Query processing and conversation history	Free-text query, session ID, result, feedback	Service/pilot controller; technical processors; possible separate controller for improvement	Purpose/basis, warning on personal data, retention, benchmark-reuse decision
Access, security and accounting logs	Identifiers, timestamps, IP/device where enabled, events	Service operator/controller or processor; security controller purpose	Data minimisation, retention schedule, restricted access and breach procedure
User profiling and recommendation	ORCID, role, interests, history and inferred preferences	Platform or pilot controller; recommender provider/processor	Explicit enablement, transparency, correction/disable, DPIA screening and fairness
Dataset onboarding and profiling	Uploaded content, metadata, inferred profiles, legal metadata	Data provider/controller; platform processor or joint controller depending purposes	Upload declaration, Article 28/26, access, rights, special-category controls, derived-data rights
MathE research and personalisation	Student identity, demographics, responses, performance and recommendations	IPB/MathE and research partners as controllers/joint controllers; technical processors	DPIA, legal basis, consent/information, age, pseudonymisation, teacher oversight and deletion
Lifelong learning pilot	Individual/organisational skills, roles and offerings	Use-case provider/employer/learner service controller; technical partners	Intended purpose, employment-law analysis, high-risk assessment, rights and fairness
Weather pilot	Environmental data, restricted station data, user queries/logs and possible media content	NOA and platform roles to confirm	Data-flow map, access terms, media-source and personal-data assessment, uncertainty communication
Language pilot and uploads	Historical corpora, licences, user uploads and profiles	IDS/data owner/controller;	Rights metadata, upload terms, TDM basis,

		platform processor; technical AI providers	restricted access and takedown
Dissemination and events	Contact details, subscriptions and attendance	Each organising partner/controller	Consent or other basis, notice, unsubscribe, retention and processor terms
External LLM and APIs	Prompts, content, metadata and technical identifiers	DataGEMS controller/provider; cloud/model/API processors or independent controllers	Vendor terms, data use/training, location, transfer mechanism, minimisation and security

Appendix II AI Act classification worksheet

For each component, the responsible owner should record:

1. System name, version and technical description.
2. Intended purpose and prohibited or excluded purposes.
3. Provider, deployer, importer, distributor or product-manufacturer role, including the role for the underlying GPAI model.
4. Persons and groups affected and the decision or output produced.
5. Whether the system matches an Annex I product or an Annex III intended purpose.
6. Whether Article 6(3) applies and whether profiling of natural persons prevents reliance on an exception.
7. Whether Article 50 interactive, generated-content, deepfake or public-interest text duties apply.
8. Data protection role, legal basis and DPIA link.
9. Human oversight, override and contestability.
10. Accuracy, robustness, cybersecurity, bias and accessibility evidence.
11. Model, data and vendor documentation and change controls.
12. AI literacy required for developers, operators, domain experts and users.
13. Date, approver and review trigger, including movement from research to pilot or operation.

Appendix III List of acronyms and abbreviations

Acronym	Meaning
AAI	Authentication and Authorisation Infrastructure
AI	Artificial intelligence
AI Act	Regulation (EU) 2024/1689
ALTAI	Assessment List for Trustworthy Artificial Intelligence
API	Application programming interface
ARC	Athena Research Center
CNRS	Centre national de la recherche scientifique
DGA	Data Governance Act
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
EDPB	European Data Protection Board
EEA	European Economic Area
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable and Reusable
GDPR	General Data Protection Regulation
GPAI	General-purpose artificial intelligence
HLEG	High-Level Expert Group on Artificial Intelligence
IDS	Leibniz Institute for the German Language
IPB	Polytechnic Institute of Bragança
JWT	JSON Web Token
LLM	Large language model
M18	Month 18
M36	Month 36
MCQ	Multiple-choice question
MoMa	Model of Metadata and data
NLP	Natural language processing
PU	Public dissemination level
RBAC	Role-based access control
R&D	Research and development
RoPA	Record of processing activities
UC	Use case
WP	Work package
ZHAW	Zurich University of Applied Sciences

Appendix IV References

I.1 Project sources

[1] DataGEMS, Grant Agreement No 101188416, including Annex 1 Description of the Action.
[2] DataGEMS, Template and partner responses questionnaire for D1.6.
[3] DataGEMS, Template and partner responses questionnaire for D1.8.
[4] DataGEMS, D2.1 Dataset Profiling Design Report, 31 October 2025.
[5] DataGEMS, D3.1 Dataset Discovery Architecture Report, 31 October 2025.
[6] DataGEMS, D4.1 User Assistance Architecture Report, 31 October 2025.
[7] DataGEMS, D5.1 Data Storage, Management and Access Architecture Report, 31 October 2025.
[8] DataGEMS, D7.1 Pilot Requirements Report, 13 October 2025.
[9] DataGEMS project website, https://datagems.eu/ (accessed 26 June 2026).

I.2 EU legislation and official guidance

[10] Charter of Fundamental Rights of the European Union, https://eur-lex.europa.eu/eli/treaty/char_2012/oj/eng.
[11] Regulation (EU) 2016/679 (General Data Protection Regulation), https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.
[12] European Data Protection Board, Guidelines 07/2020 on the concepts of controller and processor in the GDPR, final version, 7 July 2021, https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-072020-concepts-controller-and-processor-gdpr_en.
[13] 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, 18 December 2024, https://www.edpb.europa.eu/documents/opinion-of-the-board-art-64/opinion-282024-on-certain-data-protection-aspects-related-to_en.
[14] European Data Protection Board, Guidelines 01/2025 on Pseudonymisation, consultation version, https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-012025-pseudonymisation_en.
[15] Regulation (EU) 2018/1807 on a framework for the free flow of non-personal data in the European Union, https://eur-lex.europa.eu/eli/reg/2018/1807/oj/eng.
[16] Regulation (EU) 2022/868 (Data Governance Act), https://eur-lex.europa.eu/eli/reg/2022/868/oj/eng.
[17] Regulation (EU) 2023/2854 (Data Act), applicable from 12 September 2025, https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng.
[18] Directive (EU) 2019/1024 on open data and the re-use of public sector information, https://eur-lex.europa.eu/eli/dir/2019/1024/oj/eng.
[19] Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market, https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng.
[20] Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[21] European Commission, Guidelines on the definition of an AI system, 6 February 2025, https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application.
[22] European Commission, AI Literacy - Questions & Answers, https://digital-strategy.ec.europa.eu/en/faqs/ai-literacy-questions-answers (accessed 26 June 2026).
[23] European Commission, Draft guidelines on the classification of high-risk AI systems, 19 May 2026, https://digital-strategy.ec.europa.eu/en/library/draft-commission-guidelines-classification-high-risk-ai-systems.
[24] European Commission, Code of Practice on Transparency of AI-Generated Content, 10 June 2026, https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content.
[25] European Commission, General-Purpose AI Code of Practice, https://digital-strategy.ec.europa.eu/en/policies/ai-code-practice.
[26] European Parliament, AI Act: EP approves simplification measures and "nudifier" app ban, 16 June 2026, https://www.europarl.europa.eu/news/en/press-room/20260611IPR45207/ai-act-ep-approves-simplification-measures-and-nudifier-app-ban.
[27] Council of the European Union, ST 9834/1/26 REV 1, 12 June 2026, https://data.consilium.europa.eu/doc/document/ST-9834-2026-REV-1/en/pdf.
[28] European Commission High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI, 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[29] European Commission High-Level Expert Group on AI, Assessment List for Trustworthy Artificial Intelligence, 2020, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.
[30] Directive (EU) 2019/882 on accessibility requirements for products and services, https://eur-lex.europa.eu/eli/dir/2019/882/oj/eng.
[31] Directive (EU) 2016/2102 on accessibility of public sector websites and mobile applications, https://eur-lex.europa.eu/eli/dir/2016/2102/oj/eng.