# DataGEMS

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




Report

# D1.6
# D1.4a Short policy brief on the expected impact from the project

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** | D1.6 |
| **Title** | D1.4a Short policy brief on the expected impact from the project |
| **Dissemination level** | PU |
| **Due Date** | 31/6/2026 |
| **Actual Delivery Date** | 24/7/2026 |
| **Lead Beneficiary** | ARC |
| **Other beneficiaries** | |
| **Editor** | Christina Contero |
| **Authors** | |
| **Contributors** | |
| **Reviewer** | Themis Palpanas |
| **Workpackage(s)** | WP1 |
| **Keywords** | Expected impact, Policy, FAIR data, EOSC, Adoption, Use cases, Sustainability |
| **Version** | 1.0 |
| **Version information** | Number of pages: 38<br>Last file save: 24/07/2026 1:47 PM<br>File revision: 391<br>File name: D1.6_M18_D1.4a Short Policy Brief on the Expected Impact from the Project |


### 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 | Christina Contero |
| | | | |
| | | | |

## Document Tables

### Table of Tables

Table 1 Recommendations ..... 29

### Table of Figures

Figure 1 **DataGEMS M18 capability-to-impact pathway.** The figure illustrates how the technical and organisational foundations established at M18 are expected to generate wider impact, subject to appropriate governance, interoperability, accessibility and sustainability conditions and verification through the four project use cases. ..... 20

## Table of Contents

Disclaimer ..... 2
Document Information ..... 3
Copyright Information ..... 3
Document Version ..... 4
Document Tables ..... 5
Table of Contents ..... 6
Abstract ..... 8
1 Introduction ..... 10
1.1 Context ..... 10
1.2 Relationship with D1.8 ..... 10
2 Project background and M18 context ..... 11
2.1 Formal scope and objectives ..... 11
2.2 Capability areas ..... 11
2.3 Use cases and pilots ..... 11
3 Purpose and scope of the report ..... 12
4 Methodology and sources reviewed ..... 13
4.1 Source hierarchy ..... 13
4.2 Evidence grading ..... 13
4.3 Questionnaire coverage and limitations ..... 13
4.4 Cross-document review ..... 14
5 Legal, ethical and policy framework for impact ..... 15
5.1 FAIR and EOSC alignment ..... 15
5.2 Human-centred data access ..... 15
5.3 Responsible AI as an adoption condition ..... 15
5.4 Data governance and reuse policy ..... 16
6 M18 platform capability baseline ..... 17
6.1 Dataset profiling and legal metadata ..... 17
6.2 In-dataset and cross-dataset discovery ..... 17
6.3 User assistance, recommendations and explanations ..... 18
6.4 Data and model management ..... 18
6.5 Security, access and operational foundations ..... 19
6.6 Pilot and evaluation foundations ..... 19
7 Cross-cutting expected impact ..... 21
7.1 Research and technological impact ..... 21
7.2 Economic and innovation impact ..... 21
7.3 Societal impact and inclusion ..... 21
7.4 Environmental impact ..... 22
7.5 Governance and trust as multipliers of impact ..... 22

8 Sector and use-case assessment ..... 23
8.1 Higher education: MathE ..... 23
8.1.1 Capability and maturity ..... 23
8.1.2 Expected benefits ..... 23
8.1.3 Challenges and policy implications ..... 23
8.2 Lifelong learning and skills matching ..... 24
8.2.1 Capability and maturity ..... 24
8.2.2 Expected benefits ..... 24
8.2.3 Challenges and policy implications ..... 24
8.3 Weather and climate ..... 25
8.3.1 Capability and maturity ..... 25
8.3.2 Expected benefits ..... 25
8.3.3 Challenges and policy implications ..... 25
8.4 Language data infrastructure ..... 25
8.4.1 Capability and maturity ..... 25
8.4.2 Expected benefits ..... 26
8.4.3 Challenges and policy implications ..... 26
9 Adoption barriers and enabling conditions ..... 27
9.1 Metadata and data quality ..... 27
9.2 Interoperability and controlled loss ..... 27
9.3 Skills and organisational capacity ..... 27
9.4 Trust, accountability and contestability ..... 27
9.5 Infrastructure and environmental cost ..... 27
10 Public-private cooperation, ecosystem development and sustainability ..28
11 Policy recommendations and follow-up actions ..... 29
11.1 Recommendations for the DataGEMS consortium ..... 29
11.2 Recommendations for data providers and research infrastructures ..... 30
11.3 Recommendations for policymakers and funders ..... 30
11.4 Recommendations for prospective public and private adopters ..... 30
12 Policy Brief ..... 31
13 Conclusions ..... 32
Appendix I Capability-impact-recommendation matrix ..... 33
Appendix II Proposed M36 impact evidence framework ..... 35
Appendix III List of acronyms and abbreviations ..... 37
Appendix IV References ..... 38
Project sources ..... 38
Official policy and legal sources ..... 38

## Abstract

This deliverable provides the first interim policy assessment of the expected impact of DataGEMS at Month 18 (M18). DataGEMS is developing a next-generation data discovery and management ecosystem intended to support the full data lifecycle, from sharing and storage to profiling, discovery, analysis and reuse, across structured, unstructured, historical and real-time data. Its formal objectives include reducing the effort required to discover and validate data, improving the precision of discovery, linking FAIR datasets, delivering an operational open-source platform and supporting adoption through EOSC and diverse user communities.
The assessment distinguishes three evidential levels. Demonstrated impact is limited to results for which the reviewed material contains direct evidence of an implemented output, user interaction or measured indicator. Early indications of impact are supported by prototypes, pilot preparation, co-design or partner-reported benefits but require broader evaluation. Anticipated impact reflects the Grant Agreement's expected pathways and the credible potential of capabilities that remain under development.
At M18, the strongest demonstrated progress concerns the technical and organisational foundations for impact. The reviewed reports describe a coherent architecture for semantic dataset profiling, structured and unstructured data discovery, natural-language interaction, dataset packaging, query assistance, provenance-based explanations, a unified MoMa representation and role-based platform access. The pilot requirements translate these capabilities into concrete tasks in higher education, lifelong learning, weather and language data infrastructure. Questionnaire respondents also report deployable or API-ready components, the first MathE pilot implementation, a first language-user pilot implementation, a first weather pilot implementation, relational conversion of ERA5 data and a common evaluation framework.
The principal policy value of DataGEMS lies in lowering the cost and expertise threshold for working with heterogeneous data. Rich profiles can make datasets easier to understand before access or reuse. Natural-language and cross-dataset discovery can reduce dependence on exact dataset names or specialist query languages. Dataset linking, packaging and reusable analytical patterns can support combination and repeatable analysis. Query disambiguation, source grounding, provenance and explanations can improve user confidence and reduce misinterpretation. These capabilities are relevant to researchers, SMEs, public authorities, educators, data stewards and sectoral decision-makers, although the precise benefits will depend on implementation quality, representative evaluation, governance and sustainable service operation.
The MathE use case has the clearest evidence of direct user engagement and pilot preparation, including a focus group of students and teachers and reported testing of adaptive learning and content-generation workflows. Its anticipated benefits include more targeted learning resources, reduced teacher preparation effort, support for students in the learning process, and better insight into common learning difficulties. These benefits depend on pedagogical validation, protection of student data, accessible design and preventing automated recommendations from becoming opaque or determinative. The lifelong learning use case could reduce manual effort in matching skills, roles and educational offerings, but its specific datasets were not yet fixed in D7.1 and questionnaire evidence on current user testing is limited. Its policy value therefore remains primarily anticipated and should be evaluated separately for voluntary self-development and organisational employment contexts. The weather use case has a strong public-interest pathway through easier access to climate and meteorological information, geospatial enrichment and near-real-time anomaly detection. Its impact depends on communicating data provenance, uncertainty and appropriate use, particularly for public protection or planning decisions. The language use case shows early technical progress and could materially reduce

the effort required to explore multilingual historical corpora, but must manage licensing, metadata heterogeneity, computational cost and bias towards higher-resource languages or better-represented genres.

Cross-cutting adoption risks are material. These include variable metadata quality, the difficulty of harmonising highly heterogeneous datasets without information loss, unequal access for users with limited digital or domain literacy, misinterpretation of AI-assisted outputs, insufficiently transparent use conditions, computational cost and the absence at M18 of consolidated evidence on platform terms, long-term service governance and business or sustainability arrangements. The use of large language models and agent orchestration is itself identified by respondents as requiring careful framing. These concerns do not negate the expected impact; they define the enabling conditions for achieving it.

The policy recommendations are therefore directed towards concrete actors. The consortium should establish a capability-to-impact evidence register, publish minimum metadata and licence requirements for onboarding, integrate uncertainty and provenance into user-facing outputs, adopt use-case-specific human oversight rules, test accessibility and multilingual usability and align the M36 evaluation with the Grant Agreement KPIs while adding sector-specific outcome measures. Data providers should supply machine-readable rights, provenance, quality and update information. EOSC and public infrastructure actors should support interoperable onboarding, service portability and sustainable hosting. Policymakers should promote reusable legal and ethical metadata, fund data stewardship and evaluation capacity and avoid treating natural-language access as a substitute for domain competence in high-consequence settings. Public and private adopters should use staged pilots, defined accountability and contestability mechanisms rather than adopting recommendations as autonomous decisions.

The M36 follow-up should test whether the project has moved from technical readiness and early pilot evidence to verified improvement in discovery time, precision, validation effort, user success, reuse and cross-sector adoption. It should also confirm the final set of datasets and user groups, measure distributional effects and document the arrangements for maintaining services, software, metadata and governance after project funding ends.

# 1 Introduction

## 1.1 Context

Data discovery remains a practical bottleneck even where data are nominally open or accessible. Potential users may not know that a relevant dataset exists, may not be able to interpret its schema or licence, or may lack the technical skills needed to query and combine it. DataGEMS addresses this problem by developing an integrated platform for profiling, discovering, linking, analysing and reusing heterogeneous datasets through user-oriented mechanisms including natural language, recommendations, analytical patterns and explanations. The formal project description emphasises FAIRness, openness and reuse and the need to bridge data providers and data consumers across the data lifecycle.

The present deliverable is the first of two policy briefs. D1.6 establishes an M18 baseline and identifies credible pathways by which project capabilities may generate economic, societal, environmental, technological and research value. It is not a final impact evaluation and does not assume that a prototype, architecture or planned pilot has already produced a durable outcome. D1.7 at M36 should test the findings in this report against completed pilots, usage evidence, quality metrics and sustainability arrangements.

## 1.2 Relationship with D1.8

D1.6 considers privacy, data protection and responsible AI only where they affect adoption, trust, usability or the realisation of impact. The detailed interim assessment of ethical and data compliance matters is contained in D1.8. This separation avoids treating compliance analysis as a general impact discussion while recognising that responsible governance is an enabling condition for DataGEMS to be used at scale.

# 2 Project background and M18 context

## 2.1 Formal scope and objectives

DataGEMS is a 36-month Horizon Europe Research and Innovation Action running from 1 January 2025 to 31 December 2027. The consortium is coordinated by ARC and includes beneficiaries and associated partners from Greece, Italy, Portugal, France, Germany, Spain, Poland and Switzerland. The platform is intended to handle structured, unstructured, real-time and historical data and to combine data profiling, discovery, analysis, guidance and data and model management in a common ecosystem.

The Grant Agreement defines four connected objectives. The first concerns advanced discovery and exploration tools, with indicators for discovery time, successful user requests and discovery precision. The second concerns profiling and linking for discoverability, interoperability and validation, with indicators for validation time and FAIR datasets linked. The third concerns an open-source, operational and sustainable ecosystem, including technical maturity, uptime, software components and FAIR data provision. The fourth concerns integration with EOSC, uptake by communities and support for SMEs and third parties under transparent conditions.

The expected impact pathways in the Description of the Action include scientific and technological value through reusable methods and software, economic value through improved access to data and new services and societal value through easier access to evidence and lower barriers for non-technical users. The Grant Agreement also recognises barriers involving data literacy, reproducibility, interoperability, heterogeneity and language.

## 2.2 Capability areas

The M18 architecture can be understood through four capability areas. Dataset profiling generates structural, statistical, semantic, legal and ethical metadata for structured, unstructured and streaming data. Dataset discovery supports natural-language exploration within datasets, retrieval across datasets and the packaging of complementary datasets. User assistance supports query recommendation, query disambiguation, related-dataset recommendation and explanations. The data and model management layer provides the MoMa representation, analytical patterns, catalogues, user profiles, workflow orchestration and polyglot querying.

These areas are designed to work together. Profiling supplies the information used to index and interpret datasets. Discovery identifies datasets or evidence. Assistance clarifies user intent, recommends next steps and explains results. MoMa and analytical patterns represent and coordinate the data, models and operators used to complete tasks. The architecture is therefore more than a search interface: its policy relevance depends on whether these components reduce friction across the full discovery and analysis process.

## 2.3 Use cases and pilots

The Grant Agreement and D7.1 identify four use cases: higher education through MathE, lifelong learning and skills matching, weather and climate and language data infrastructure. D7.1 states that 20 pilots are planned, five per use case, and that pilots will increase in complexity through co-creation and evaluation. The use cases cover different data modalities, user expertise, access conditions and decision contexts, providing a useful test of replicability.

# 3 Purpose and scope of the report

D1.6 assesses how DataGEMS capabilities may create value for identified sectors and stakeholders and which policy or governance conditions would improve that value. It covers the platform capabilities evidenced at M18, the four project use cases, cross-sector adoption conditions, public-private cooperation, sustainability, scalability and indicators for the M36 follow-up.
The assessment does not measure final social or economic returns and does not validate every technical claim through independent testing. Partner questionnaire responses are treated as evidence of the position reported by the named respondent, not automatically as a consortium-wide conclusion. Expected outcomes in the Grant Agreement are treated as targets or pathways rather than achievements. Legal and ethical matters are addressed proportionately where they affect impact and adoption, with detailed analysis reserved for D1.8.

# 4 Methodology and sources reviewed

## 4.1 Source hierarchy

The Grant Agreement was treated as the primary source for formal project scope, objectives, tasks, use cases, deliverables and expected impacts. The responses to the D1.6 questionnaire, designed by the External Ethics Advisor and distributed among the partners, were used as evidence of the M18 position reported by CNRS, IDS and ZHAW. D2.1, D3.1, D4.1, D5.1, D7.1 and D7.7 were used for the technical architecture, operational flows, pilot requirements and initial pilot implementation and evaluation results. The project website was used only for general public-facing background. Official EU legal and policy sources were reviewed to frame policy recommendations.

## 4.2 Evidence grading

The report uses three evidence categories:

- **Demonstrated at M18** means that the reviewed evidence identifies an implemented output, available prototype, user interaction or reported quantitative result.
- **Early indication** means that architecture, prototype work, co-design, initial evaluation or partner reporting supports a plausible benefit but wider evidence is not yet available.
- **Anticipated or potential impact** means that the pathway is consistent with the Grant Agreement and technical design but remains dependent on implementation, deployment and evaluation.

This grading avoids presenting technical completion as societal impact. This is because a profiling service can be operational without yet showing that researchers discover data faster or a recommendation function can exist without yet demonstrating fair or useful recommendations across user groups. The M36 assessment should connect component performance to user and organisational outcomes.

## 4.3 Questionnaire coverage and limitations

Three organisations provided identifiable D1.6 responses: CNRS, IDS and ZHAW. Their roles provide useful coverage of discovery, user assistance, profiling, pilot coordination and the language use case, but the questionnaire does not contain direct D1.6 responses from all use-case providers, platform operators or exploitation partners. The absence of a response is not treated as a deficiency. Where necessary, D7.1 and the architecture reports were used to establish the planned position.

There are certain elements that should be observed as limitations that may potentially affect the strength of conclusions. Many reported benefits concern prototypes or expected functionality rather than comparative measurements. Some KPI values are supplied by a single respondent and were not accompanied by calculation methods or source datasets. The Grant Agreement KPIs are ambitious project-level targets, but the questionnaire does not yet provide a consolidated M18 KPI dashboard. Information on costs, business models, environmental footprint, service-level performance, accessibility testing and external adoption is incomplete. These matters are therefore framed as areas for M36 verification rather than presumed gaps.

## 4.4 Cross-document review

The evidence was checked for consistency across the two M18 reports. D1.6 focuses on capability, use, benefit, adoption and policy. D1.8 focuses on legal duties, ethical principles, safeguards, uncertainty and accountability.

# 5 Legal, ethical and policy framework for impact

## 5.1 FAIR and EOSC alignment

DataGEMS is positioned within the EOSC policy environment and its emphasis on findable, accessible, interoperable and reusable research outputs. The project does not merely publish data; it seeks to improve the metadata, interfaces and workflows through which data can be understood and used. D2.1 adopts and extends the Croissant and Schema.org approach to describe distributions, record sets, access information and derived profiles. D5.1 provides a common MoMa representation and catalogue for datasets, models, profiles and analytical patterns. These design choices support interoperability if the extensions are documented, stable and exportable.

FAIRness should not be equated with unrestricted openness. Some project datasets are private, restricted or contain personal data. The policy objective is therefore appropriately framed as making data as open as possible and as closed as necessary, with machine-readable access conditions, purpose restrictions, licences and safeguards. The draft legal and ethical metadata schema in D2.1 is particularly relevant because it could enable users and automated services to identify copyright status, licence type, allowed uses, restrictions, personal-data status, controller, purpose, retention and safeguards before reuse.

## 5.2 Human-centred data access

Natural-language access and stepwise guidance can broaden participation, but only where the interface preserves meaningful user control. Users need to know which datasets were used, what transformations occurred, whether a query was disambiguated and how uncertainty affects an answer. D3.1 and D4.1 describe source-grounded retrieval, human-in-the-loop disambiguation, provenance and analytical-pattern explanations. These capabilities align with a human-centred policy approach because they can make system outputs inspectable rather than merely convenient.

The policy implication is that accessibility should be defined broadly. It includes plain-language interaction, multilingual capability, disability accessibility, understandable licensing, predictable response times and support for users who lack data engineering expertise. Natural-language interfaces can lower one barrier while creating another if users cannot judge the quality of an answer. Domain-specific explanation and uncertainty communication are therefore part of accessibility, not optional additions.

## 5.3 Responsible AI as an adoption condition

Several DataGEMS functions use or plan to use machine learning, embeddings, reinforcement learning or large language models for profiling, retrieval, ranking, recommendation, text-to-SQL, question generation and assistance. The AI Act is already partly applicable. As at 26 June 2026, the obligations concerning prohibited practices and AI literacy have applied since 2 February 2025 and governance and general-purpose AI model obligations since 2 August 2025. Most remaining provisions are scheduled to apply from 2 August 2026 under the original Regulation. Parliament approved an agreed AI Omnibus amendment on 16 June 2026, but formal Council adoption was still outstanding and scheduled for 29 June 2026. The agreed high-risk dates should therefore be treated as highly probable planning assumptions, not yet as formally enacted law at the research cut-off.

For impact policy, the central point is not classification alone. Trustworthy use requires documented intended purpose, suitable data, human oversight, transparency, accuracy and monitoring. These controls can increase adoption by universities, public authorities and data holders that would otherwise be reluctant to rely on AI-assisted discovery. Conversely, overclaiming accuracy or presenting generated content without clear labelling could undermine confidence in the wider platform.

## 5.4 Data governance and reuse policy

The EU data framework includes the GDPR, the free-flow regime for non-personal data, the Data Governance Act, the Data Act, the Open Data Directive and copyright and database law. The Data Act has applied since 12 September 2025 and reinforces policy expectations around fair access, switching and portability for data-processing services. While not every provision will apply directly to DataGEMS, service portability, open interfaces and avoidance of lock-in are relevant to sustainability. Policy recommendations should therefore favour exportable profiles, documented APIs, transparent dependency on external models and clear rights in datasets, software and derived outputs.

# 6 M18 platform capability baseline

## 6.1 Dataset profiling and legal metadata

D2.1 describes structured profiling for CSV, Excel and relational databases, unstructured profiling for text and PDF, near-real-time profiling for evolving time series and dataset linking based on metadata and structural relationships. Profiles can include schema, descriptive statistics, language, keywords, summaries, data-quality information and links between datasets. The technical value is a consistent layer that can help users determine relevance and help other DataGEMS components index, query and combine data.

At M18, this capability represents demonstrated technical progress rather than demonstrated user impact. ZHAW reports that core profiling components are developed and that work has produced a deployable platform prototype, a relational form of ERA5 data, automated classification of MathE questions and linguistic feature extraction for historical encyclopaedias. IDS reports API readiness of key language components and reduced effort in textual retrieval and automated feature generation, with a first interface accessible to researchers. ARC has developed the structured dataset profiler, automatically discovering and inferring statistical data features to reduce time needed for user familiarization with the dataset. These are early indications that profiling can reduce data familiarisation and preparation effort.

The draft legal and ethical metadata schema is a distinctive policy contribution. If implemented consistently, it could convert legal and ethical conditions that are usually scattered across licences, notices and institutional records into discoverable metadata. This would support responsible reuse and could become reusable beyond the four pilots. Its impact depends on governance: fields must be accurate, attributable to a data steward, updated when conditions change and not treated as automated legal advice.

## 6.2 In-dataset and cross-dataset discovery

In-dataset discovery is designed to answer questions over relational, textual and image data. For relational data, natural-language questions may be translated into SQL, supported by dataset profiles and query disambiguation. Cross-dataset discovery indexes profiles and content-derived representations, retrieves candidate datasets and re-ranks them by relevance. Dataset packaging seeks combinations of datasets that cover a user need while limiting redundancy and improving compatibility.

The expected impact pathway is direct: better retrieval can reduce search time and improve the chance that a user finds data able to answer a question. The design also addresses a known limitation of keyword search by considering content, semantic similarity and the user's analytical intent. Evidence-backed answers and relevance scores can help users inspect results. The main challenges are semantic error, false confidence in generated SQL, bias in ranking, latency and the need for evaluation datasets that represent realistic user needs.

CNRS reports first implementations of in-dataset and cross-dataset operators and a unified backend for MoMa, analytical patterns and operators. ZHAW reports initial Text-to-SQL development, including translation of place names into coordinates for weather queries. ARC reports initial implementation and development of search and retrieval capabilities, including an LLM-based filtering step for increased recall, as well as the integration to the dataset ingestion pipeline which automatically enables search capabilities over any newly uploaded datasets. While these are credible early indications of value, they are not yet evidence of the Grant

Agreement target reductions in discovery time. However, ZHAW has demonstrated increases in precision over state-of-the-art baselines

## 6.3 User assistance, recommendations and explanations

D4.1 describes four assistance functions: query recommendation, query disambiguation, dataset recommendation and analytical-pattern explanations. Query recommendation may use the current query, prior queries and a user profile. Disambiguation combines dataset context, LLM reasoning and human confirmation. Initial dataset recommendation is metadata-based, with future personalisation contemplated. Explanations use provenance and operator annotations to show how analytical patterns produce results.

These capabilities could materially improve adoption by non-specialists. Disambiguation can prevent a system from silently choosing a meaning for ambiguous terms. Explanations can support learning, audit and error correction. Recommendations can expose related resources that users would not otherwise encounter. However, assistance may also steer users towards familiar or popular datasets, reinforce profile-based assumptions or obscure the distinction between a system suggestion and an evidence-based conclusion. The policy standard should therefore be assistive rather than determinative: users should be able to view, reject and correct recommendations and to continue without personalisation where feasible.

CNRS reports an implemented service for provenance explanations for SQL queries and plans to extend explanations beyond SQL. This is demonstrated technical progress. User-level benefits, such as improved comprehension or reduced error rates, remain to be measured.

ARC's material recommendation system shows demonstrated technical progress, with online evaluation and expert validation indicating that it can correctly match a substantial share of questions to relevant educational materials. Potential benefits for students and educators are promising, but impacts on learning outcomes, engagement, and workload reduction remain to be measured.

## 6.4 Data and model management

D5.1 presents MoMa as a unified property-graph representation for datasets, metadata, machine-learning models, user profiles, tasks, operators and analytical patterns. It also describes a task orchestrator, polyglot query engine, data catalogue, user-profile component and workflow orchestrator. This layer is intended to decouple storage and low-level management from discovery and analytics and to enable reusable workflows across heterogeneous sources.

The impact pathway is technological and organisational. A common representation can reduce repeated integration work and make analytical processes reusable, inspectable and optimisable. Analytical patterns can preserve not only a result but the sequence of operators, data and models used to produce it. This can support reproducibility and service composition. CNRS reports that the need for a unifying model and platform has been validated through cross-fertilisation between use-case providers and technical partners.

The long-term value of MoMa will depend on documentation, interoperability with external standards, versioning and the ability to export objects and workflows without dependence on a single deployment. A project-specific model can enable rapid development, but policy value at ecosystem level requires stable mappings to broadly used standards and governance for extensions.

## 6.5 Security, access and operational foundations

The technical reports describe REST APIs, JSON Web Token authentication, role-based access control, protected services, health monitoring, audit logs and gateway-mediated access. D5.1 describes an internal workflow orchestrator based on Apache Airflow and integration with DataGEMS authentication and authorisation infrastructure. D3.1 describes bearer-token validation and role checks for discovery services.
These controls are not impact outcomes in themselves, but they are prerequisites for data holders to contribute restricted data and for public bodies or companies to use the platform. The architecture also supports differentiated access rather than a binary open/closed model. At M36, the consortium should complement architectural descriptions with operational evidence on uptime, incident management, access reviews, performance and user support.

## 6.6 Pilot and evaluation foundations

D7.1 maps user communities, datasets, user stories and development tasks for each use case. It states that pilots will be co-created and progressively evaluated, with qualitative and quantitative indicators and feedback into technical work packages. CNRS reports a unified evaluation framework. D7.7 provides the initial implementation evaluation results of the first five pilots deployed. These pilots were selected for early deployment due to their technical and operational maturity, active stakeholder involvement, and coverage of diverse domains, namely Higher Education MathE use case, Weather use case and Language use case. Together, they provide initial evidence on the performance, usability, and adaptability of DataGEMS services in realistic operational settings. Across the five pilots, the DataGEMS components were successfully exercised in realistic operational scenarios. These included adaptive question recommendation, educational material retrieval, AI-assisted MCQ generation and validation, natural language querying of meteorological data, and feature-based exploration of text corpora. The evaluation also confirmed the relevance of the three-dimensional assessment framework
The strongest policy opportunity is to connect evaluation across three levels: component performance, user experience and sectoral outcome. For example, retrieval precision is a component measure, time saved is a user or organisational measure and improved climate planning is a sectoral outcome. D1.7 should avoid aggregating these levels into a single claim.

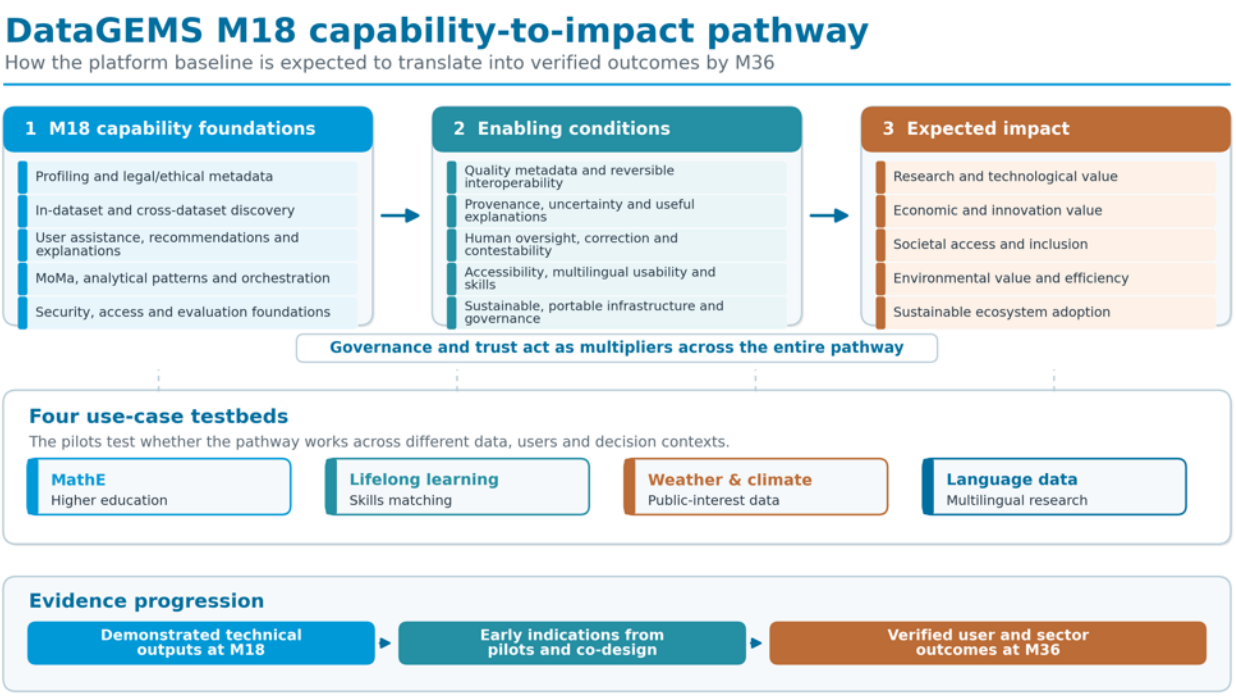



*Figure 1* ***DataGEMS M18 capability-to-impact pathway.*** *The figure illustrates how the technical and organisational foundations established at M18 are expected to generate wider impact, subject to appropriate governance, interoperability, accessibility and sustainability conditions and verification through the four project use cases.*

# 7 Cross-cutting expected impact

## 7.1 Research and technological impact

The most immediate expected research impact is a reduction in the engineering effort needed to locate, understand and combine data across formats. Standardised profiles, content-derived metadata and dataset links can help researchers identify data beyond their existing disciplinary repositories. Natural-language and multimodal discovery can support exploratory work where the user does not know the exact schema or dataset name. Analytical patterns can capture repeatable workflows and provenance, increasing the potential for reproducibility.
The project also has methodological value. Its components address difficult research problems including semantic profiling, cross-dataset retrieval, text-to-SQL, multilingual discovery, real-time anomaly profiling and provenance explanations. Open-source release and reusable benchmarks could allow these contributions to be tested and extended by other projects. CNRSZHAW, and ARCreport cross-use-case technical fertilisation, which is an early indication that generic components are emerging rather than isolated pilot applications.
Risks to research impact include overfitting to four use cases, incomplete benchmarking, reliance on proprietary model endpoints, insufficient documentation and performance that varies by language or data modality. The project should therefore publish evaluation protocols, negative results, model and dependency versions and examples where the system does not provide a reliable answer.

## 7.2 Economic and innovation impact

Economic value may arise from lower data acquisition and preparation costs, faster discovery of relevant evidence, reuse of existing public or private datasets and new services built on DataGEMS components. SMEs could benefit because they often lack dedicated data engineering teams. The Grant Agreement anticipates links between company data and open scientific data and support for SMEs and third parties through EOSC.
At M18, there is not yet sufficient evidence to quantify cost savings, revenue, new products or SME adoption. The relevant demonstrated outputs are reusable software and platform components. The economic pathway should therefore be tested through task-based pilots: time required to onboard a dataset, time to answer a representative question, number of manual steps, compute and hosting cost, user support required and the value of datasets or combinations discovered. A small number of well-designed before-and-after studies would be more persuasive than broad adoption claims.
Economic sustainability also requires clarity on licensing, service responsibility and dependencies. Open-source code does not by itself guarantee a sustainable service. A viable post-project model must specify which components will be maintained, by whom, at what service level, under which licence and how infrastructure and support costs will be covered.

## 7.3 Societal impact and inclusion

DataGEMS may broaden access to evidence by allowing users to express needs in natural language and by providing guidance through complex datasets. This can benefit educators, journalists, planners, civil-protection actors, cultural institutions and smaller research groups. The weather and education use cases are particularly visible public-interest examples. The language use case may improve access to cultural and historical resources across languages. The lifelong learning use case may help individuals identify skills and training opportunities.

Inclusion is not automatic. Natural-language performance may be better in high-resource languages. Interfaces may still require data literacy to interpret outputs. Personalised systems may work less well for users with limited history. People with disabilities may encounter barriers unless accessibility is tested. Smaller organisations may lack compute or staff to operate local models. The project should measure who benefits, who needs additional support and whether default settings produce materially different outcomes across user groups.

## 7.4 Environmental impact

The weather use case can support climate research, risk awareness and planning by making historical, real-time and geospatial data easier to explore. DataGEMS may also reduce duplicated computation and storage where reusable profiles, cached results and analytical patterns allow work to be repeated efficiently. These are plausible positive pathways.
At the same time, profiling large corpora, generating embeddings and using LLMs can consume substantial computing resources. IDS identifies high computational cost for multilingual LLM-based processing as a risk. The project materials do not provide an M18 environmental footprint or energy metrics. Environmental claims should therefore remain qualified. The consortium should monitor model size, inference volume, caching, hardware utilisation and the benefit obtained per compute-intensive task. Lightweight and local models should be preferred where they meet the required quality and security level.

## 7.5 Governance and trust as multipliers of impact

A data platform creates impact only if data holders and users trust its handling of data and outputs. The draft legal and ethical metadata schema, role-based access, human-in-the-loop disambiguation, provenance and explanation are positive foundations. They can reduce uncertainty over reuse and allow users to inspect how answers were produced. D1.8 nevertheless identifies material areas for consolidation, including controller and processor roles, platform privacy information, retention rules, DPIA scope and AI inventory. Addressing these matters is likely to increase adoption and reduce later redesign.

# 8 Sector and use-case assessment

## 8.1 Higher education: MathE

### 8.1.1 Capability and maturity

The MathE use case combines student guidance, adaptive multiple-choice question recommendation, supplementary resource discovery, teacher analytics and administrator validation. D7.1 identifies students, teachers and MathE administrators and describes datasets containing student responses, user profiles, questions, topics, difficulty levels and pedagogical materials. It reports a focus group of 20 students and five teachers and specifies user stories for chatbot guidance, adaptive questions, generated resources, teacher dashboards and metadata-rich content generation.
CNRS reports open-source language models for MCQ generation from MathE PDFs and reinforcement-learning approaches for quiz composition. It also reports a feedback platform through which MathE partners review generated questions and inject corrections. The respondent describes the current stage as development and pilot, with MathE students and teachers included in the reported pilot-engagement count. ZHAW reports automatic categorisation of mathematics questions into a taxonomy. ARC reports an implemented material recommendation system for MathE questions, with online evaluation as well as expert validations from selected topic-subtopic pools.

### 8.1.2 Expected benefits

For students, the principal anticipated benefit is more responsive access to exercises, hints and materials aligned with topic, difficulty and prior performance. For teachers, the platform may reduce the effort required to search for resources, compose quizzes and identify common learning difficulties. For administrators, automated metadata and validation workflows may help expand and maintain the content base. At institutional level, the use case could provide evidence on how AI-assisted content discovery and generation can augment rather than replace pedagogical judgement.
The early evidence supports technical feasibility and user engagement, but not yet improved learning outcomes. D1.7 should distinguish engagement, time saved and pedagogical outcome. Recommended measures include time required to prepare a quiz, teacher acceptance and correction rates for generated MCQs, student comprehension, error and hallucination rate, distribution of recommendations by difficulty and topic, accessibility, dropout and learning gains where methodologically appropriate.

### 8.1.3 Challenges and policy implications

Generated questions can be mathematically wrong, ambiguous or pedagogically inappropriate even where language is fluent. Adaptive recommendations may narrow exposure, misclassify ability or create undue reliance on historical performance. Student data require careful protection, particularly if users may include persons under 18 or if profiles are reused for research. The use case should therefore maintain teacher approval before generated MCQs enter the operational bank, show learners when content is AI-generated or recommended, provide a route to report errors and avoid using recommendation scores as autonomous assessment or access decisions.

Policy recommendation for higher-education institutions: adopt a documented pedagogical assurance protocol that defines which outputs require teacher validation, how students are informed, how errors are corrected and how recommendation performance is reviewed across different learner groups. Procurement or adoption decisions should require evidence of educational value and accessibility, not only technical accuracy.

## 8.2 Lifelong learning and skills matching

### 8.2.1 Capability and maturity

The lifelong learning use case aims to compare individual or organisational skill models with current or future skill needs and to identify relevant educational offerings. D7.1 identifies workers, managers, human-resource executives, skill researchers and trainers. It notes that specific datasets had not yet been determined and describes relevant categories such as ESCO, organisational skill models, educational offerings, skill-demand datasets and sector taxonomies. Planned scenarios cover worker-focused and manager-focused gap analysis in English, German and French.
The CNRS questionnaire provides a more developed M18 capability description, including structured and unstructured profiling, restricted uploads, cross-dataset discovery, skill-gap analysis and training recommendations using ESCO, O*NET and other resources. This response is useful evidence of the technical concept, but the D1.6 questionnaire does not include a direct response from the lifelong-learning use-case provider on current user testing, fairness or organisational effects. Accordingly, the impact is mainly anticipated.

### 8.2.2 Expected benefits

Individuals may gain a clearer view of skills associated with a desired role and relevant learning opportunities. Managers may gain a structured overview of organisational capabilities and future needs. Training providers may improve the discoverability of offerings through richer metadata and standard taxonomies. Researchers may analyse changes in skill demand across sources and time.
The main economic pathway is reduction of manual mapping between inconsistent role, skill and course descriptions. The main societal pathway is more accessible upskilling and reskilling. These benefits depend on data quality, current taxonomies, understandable recommendations and the ability to challenge a result. A skills profile is not a complete representation of a worker's capability, potential or circumstances.

### 8.2.3 Challenges and policy implications

The same technical function can have very different consequences depending on context. A voluntary tool used by a worker to explore training is lower consequence than a system used by an employer to rank workers, allocate opportunities or make promotion or redundancy decisions. Employment use may introduce discrimination, power imbalance and AI Act high-risk implications. The project should separate these intended purposes in design, testing and documentation.
Policy recommendation for employers and public employment services: use DataGEMS outputs as decision support for training and workforce planning, not as sole evidence of suitability or performance. Provide employees with access to the data and assumptions used, a route to correct skills information and meaningful human review. Evaluate recommendations for indirect discrimination across occupational, language, gender, age and disability groups where data and law permit.

## 8.3 Weather and climate

### 8.3.1 Capability and maturity

The weather use case combines ERA5-Land reanalysis, historical and near-real-time station data, elevation, land cover and population data. D7.1 identifies climate scientists, civil-protection authorities, urban and regional planners, sectoral decision-makers, educators and journalists. The planned functionality includes natural-language querying, geospatial processing, derived climate indices, overlays and anomaly detection.
ZHAW reports conversion of ERA5 data into a relational database and initial Text-to-SQL functionality that resolves place names to coordinates. CNRS's questionnaire describes a staged weather data query service extending to geospatial enrichment, CERRA, historical stations, near-real-time anomaly detection and exploratory extraction from media reports. The architecture reports provide a worked example of translating a natural-language weather question into geospatial information and SQL. D7.7 confirms that the first Weather Data Query Service pilot reached a functional end-to-end stage and underwent initial testing by 15 participants, five from ZHAW and ten from NOA, while detailed structured user feedback and broader performance and platform-level testing are still required.

### 8.3.2 Expected benefits

For scientists, the platform may reduce preparation time and make it easier to combine meteorological and contextual datasets. For planners and sectoral users, natural-language and map-based access may make climate indicators more usable. For civil protection, near-real-time anomaly information could improve situational awareness. For educators and journalists, source-grounded summaries and visualisations may support climate literacy.
The expected societal impact is significant because weather and climate data affect public safety, infrastructure, tourism, energy, construction and adaptation. The project should nevertheless distinguish exploratory insight from operational warning. A research prototype should not be represented as an authoritative alerting service without validated data feeds, thresholds, availability and organisational responsibility.

### 8.3.3 Challenges and policy implications

Meteorological outputs are sensitive to resolution, station coverage, missing data, model assumptions, spatial interpolation and the distinction between reanalysis, observation and forecast. Natural-language responses can omit these limitations. Geospatial joins may appear precise while combining datasets with different scales and update cycles. Media extraction introduces additional source-quality risks.
Policy recommendation for public authorities and sectoral adopters: require every decision-support output to expose source, observation period, spatial resolution, update time, method and uncertainty or known limitations. Establish clear labels for exploratory, research and operational products. Where outputs support safety-related decisions, integrate them into existing expert and institutional protocols rather than bypassing them.

## 8.4 Language data infrastructure

### 8.4.1 Capability and maturity

The language use case targets computational linguists, digital humanists, lexicographers, educators and domain experts working with multilingual historical encyclopaedic corpora and related resources. D7.1 describes tasks for feature-based corpus views, domain retrieval, semantic-change analysis, accessible knowledge transformation and human-in-the-loop

research. The underlying data include German, English and French corpora in TEI XML, JSON, text and image formats.
IDS reports UC4P1 at pilot stage, UC4P2 and UC4P3 in development and later pilots in design. It reports API-ready components, efficient textual retrieval, automated feature generation and a first interface available to researchers. It also reports gold-dataset work and a CLARIN-DataGEMS workshop in February 2026. The respondent identifies no personal data from living persons in the current pilot corpora and states that contemporary Wikipedia entries without historical counterparts were excluded for pilot alignment. ARC reports the implementation and deployment of the text retrieval functionality that is used for UC4P1 through the Cross-Dataset Discovery service, the implementation of an LLM-based filtering step that ensuring that retrieved texts are relevant to the user's question and achieving high recall (evaluated on the dataset created by IDS), and the continuous development and extension of the retriever for upcoming pilot releases. It also reports the integration of Cross-Dataset Discovery to the project's dataset ingestion pipeline, which will allow text retrieval to operate over any dataset uploaded in the DataGEMS platform in future pilot releases.

### 8.4.2 Expected benefits

The principal research benefit is lower effort in collecting, normalising and comparing corpora across languages and periods. Multilingual profiling and common pipelines can help smaller teams and cultural institutions conduct analyses that would otherwise require specialised coding and data-engineering capacity. Semantic-change analysis and accessible transformation can support digital humanities and education. Better metadata and rights information can improve reuse of public-domain and openly licensed resources.
The evidence supports early technical and user-interface progress. Broader impact remains dependent on evaluation with researchers from different disciplines and levels of technical skill. D1.7 should measure time to complete representative corpus tasks, retrieval quality by language, information loss during conversion, accessibility, correction rates and reuse outside the initial encyclopaedic corpora.

### 8.4.3 Challenges and policy implications

IDS identifies heterogeneity of formats and project-specific annotation, inconsistent metadata quality, potential information loss during harmonisation, limited adoption by non-technical users and computational cost. It also identifies bias from over-representation of high-resource languages, genres, demographics and periods. Users may upload data with unclear copyright or reuse conditions.
Policy recommendation for language-data infrastructures and cultural institutions: adopt common metadata and provenance profiles while preserving access to original annotations; document any loss introduced by conversion; provide machine-readable licences and text-and-data-mining conditions and evaluate retrieval separately by language, period and genre. Responsibility for rights clearance should not be shifted to users through a generic warning alone: onboarding should request evidence and restrict publication where rights remain uncertain.

# 9 Adoption barriers and enabling conditions

## 9.1 Metadata and data quality

Automated discovery is only as reliable as the profiles, source data and links it uses. Missing licences, ambiguous column names, outdated taxonomies and inconsistent units can propagate through ranking and analysis. DataGEMS should establish minimum onboarding fields and quality flags, distinguish user-supplied from inferred metadata and allow data stewards to approve or correct profiles. Automated summaries should retain links to source fields and indicate confidence where appropriate.

## 9.2 Interoperability and controlled loss

A common representation enables cross-domain work but can flatten domain-specific meaning. The language use case explicitly anticipates controlled annotation loss in exchange for uniform analysis. Similar trade-offs arise in weather resolution and skill taxonomies. The platform should preserve original formats, record transformations and expose mappings rather than replacing domain structures. Interoperability should be reversible where feasible.

## 9.3 Skills and organisational capacity

Natural-language interfaces reduce coding requirements but do not eliminate the need for data literacy. Users need to understand relevance scores, uncertainty, licences, personal-data restrictions and the difference between correlation and explanation. DataGEMS adoption should therefore include role-specific guidance for data stewards, researchers, teachers, public officials and SMEs. The project can contribute reusable training materials through EOSC and sectoral networks.

## 9.4 Trust, accountability and contestability

Users and data holders need clear responsibility for source data, profiling, model outputs, access decisions and corrections. DataGEMS should display responsible organisations and contact routes, retain meaningful provenance, and support user feedback and challenge. Explanations should be evaluated for usefulness, not merely generated. A technically detailed trace may not be understandable to a teacher or planner; layered explanations are preferable.

## 9.5 Infrastructure and environmental cost

Local hosting can improve control but may be difficult for smaller organisations. External model services can lower deployment cost but introduce transfer, confidentiality, dependency and pricing concerns. The platform should support a documented choice between local and external models, publish resource requirements and avoid coupling core functions to a single provider. Caching, smaller models and selective use of generative functions can improve cost and environmental performance.

# 10 Public-private cooperation, ecosystem development and sustainability

DataGEMS has a credible role as shared infrastructure between universities, research infrastructures, public authorities and companies. Data holders can contribute expertise and controlled datasets; technical partners can develop reusable services; EOSC can provide discovery, federation and sustainability channels; public authorities and SMEs can test real-world tasks. This model is valuable where governance prevents private advantage from undermining openness or public-interest use.

Public-private pilots should define contributions, rights and post-pilot access before data are shared. Agreements should cover permitted purposes, confidentiality, publication, derived metadata, model training, security, liability, exit and deletion. Data providers should retain the ability to correct metadata and revoke access where the legal basis ends. Publicly funded generic components and non-sensitive outputs should be released under clear open licences where consistent with the Grant Agreement and third-party rights.

Long-term sustainability requires a service portfolio rather than an assumption that all prototypes will continue. The consortium should identify which components are core EOSC-facing services, which are research demonstrators and which are use-case-specific integrations. For each retained component, it should specify owner, maintainer, code repository, licence, hosting, support, security patching, data stewardship, service levels and financing. The Grant Agreement includes a target of three years' service sustainability after the project; D1.7 should assess the evidence supporting that target.

Scalability should also be measured qualitatively. A component may scale computationally but fail to scale organisationally because onboarding requires intensive manual data cleaning or rights review. Replicability indicators should therefore include time and skills needed to onboard a new domain, number of bespoke mappings, accuracy before and after domain adaptation, support effort and the ability of an external organisation to deploy or call the service using public documentation.

# 11 Policy recommendations and follow-up actions

## 11.1 Recommendations for the DataGEMS consortium

*Table 1 Recommendations*

| Priority | Recommendation | Evidence and purpose | Timing |
|---|---|---|---|
| **High** | Maintain a capability-to-impact evidence register linking each pilot task to component version, user group, baseline, metric, result and limitation. | M18 evidence is stronger on technical outputs than user or sectoral outcomes. The register will support credible D1.7 claims and prevent planned benefits being reported as achieved. | Establish in M19-M21 and update quarterly. |
| **High** | Implement minimum onboarding metadata for provenance, licence, access, quality, update frequency, personal-data status and responsible data steward. | D2.1 provides a strong draft legal and ethical schema; adoption depends on accurate, complete and maintained fields. | Core fields before wider external onboarding; enrich through M36. |
| **High** | Make provenance, source, transformation, uncertainty and known limitations visible in user-facing results. | Misinterpretation is a cross-use-case risk, particularly for generated educational content and weather outputs. | Integrate into pilot interfaces and evaluation. |
| **High** | Define use-case-specific human oversight and contestability rules. | The consequences of recommendations differ between education, employment-related skills use and exploratory research. | Before pilot expansion involving external users. |
| **Medium** | Evaluate accessibility, multilingual performance and distributional effects as explicit quality dimensions. | Natural-language and multilingual access are impact pathways but may reproduce language, disability or digital-skills barriers. | Include in M18-M36 evaluation plans. |
| **Medium** | Track compute, latency, external-model dependency and approximate energy indicators for intensive services. | IDS identifies computational cost; environmental benefit is currently unmeasured. | Begin with the most compute-intensive profiling and LLM workflows. |
| **Medium** | Publish deployment profiles for local, federated and managed-service operation, including export and exit procedures. | SMEs and public bodies need realistic resource and governance information and protection against lock-in. | Draft by M30 and validate before D1.7. |

## 11.2 Recommendations for data providers and research infrastructures

Data providers should supply structured metadata on provenance, temporal and spatial coverage, update cycles, quality, licences, restrictions and responsible contacts. Where an automated profile is used, the provider should be able to review and correct it. Original data and annotation structures should be preserved alongside harmonised representations. Providers of restricted datasets should define permitted purposes and whether profiling, embeddings, benchmark creation or model training are allowed separately.
Research infrastructures, including EOSC and CLARIN-related actors, should support mapping between DataGEMS profiles and established community standards, persistent identifiers, vocabularies and rights expressions. They should provide test environments for cross-node access, portability and identity federation and should consider shared stewardship services for organisations that cannot maintain rich metadata alone.

## 11.3 Recommendations for policymakers and funders

Policymakers should support investment in data stewardship, metadata quality and evaluation, not only in AI interfaces. Funding conditions can encourage machine-readable licences, reusable legal and ethical metadata, open evaluation resources and maintenance plans. Public-sector adoption guidance should require provenance and uncertainty for AI-assisted data analysis and should distinguish exploratory tools from systems used for safety, education assessment or employment decisions.
The EU and Member States should continue developing interoperable guidance on copyright, text and data mining, public-sector information, data protection and AI transparency in research infrastructures. Smaller organisations need practical templates and support to participate without transferring disproportionate compliance or technical burdens to individual researchers.

## 11.4 Recommendations for prospective public and private adopters

Adopters should start with bounded tasks, baseline current effort and define acceptable error and response conditions. They should retain domain experts, test outputs with representative users and document who is responsible for final decisions. Personalisation should be optional and proportionate. Adoption contracts should secure access to data and metadata exports, audit information, incident support and the ability to discontinue external model services without losing core data or workflows.

# 12 Policy Brief

The final policy brief should revisit the following questions:

- Have the Grant Agreement indicators for discovery time, successful requests, precision, validation time and FAIR linking been measured using published methods and representative tasks?
- How many pilots were completed, with which user groups, datasets and sample sizes, and how were participants recruited?
- Which impacts are demonstrated through comparative evidence rather than self-report or technical completion?
- Do benefits and error rates differ by language, technical expertise, disability, organisation size or use-case subgroup?
- How often do users accept, reject or correct disambiguation, recommendations and generated content?
- What is the teacher correction rate for generated MCQs and what evidence exists of learning or preparation-time benefits?
- Have lifelong-learning recommendations been tested separately for voluntary learner use and organisational workforce use?
- Are weather outputs accompanied by provenance, resolution, recency and uncertainty and are any operational users relying on them?
- How much effort and information loss are involved in onboarding a new language corpus or a new domain outside the four use cases?
- What are the compute, cost, latency, uptime and incident figures for the services intended for sustainability?
- Which components have confirmed owners, licences, repositories, hosting, support and financing after M36?
- What external EOSC, SME, public-authority or research-infrastructure adoption is evidenced and what barriers remain?

# 13 Conclusions

At M18, DataGEMS has established a credible technical and pilot foundation for improving data discovery and reuse. Its strongest distinctive features are the integration of rich profiling, semantic and natural-language discovery, dataset linking and packaging, reusable analytical patterns, user assistance and provenance within a common architecture. The project has early evidence of implemented components and user engagement, particularly in MathE and the language use case, but most economic, societal and environmental impacts remain prospective.

The pathway to impact is therefore conditional. DataGEMS will create durable public value if it demonstrates measurable improvements on real tasks, maintains domain and human oversight, provides transparent rights and quality metadata, supports diverse users and establishes a sustainable and portable service model. The recommendations in this report are intended to convert the M18 technical baseline into an evidence base for D1.7 and to ensure that adoption grows with, rather than ahead of, the project's capacity to explain, govern and maintain its outputs.

# Appendix I Capability-impact-recommendation matrix

| Capability | Primary use cases and stakeholders | Expected pathway to impact | Main challenge | M18 evidence level | Priority policy response |
|---|---|---|---|---|---|
| **Structured, unstructured and real-time profiling** | All use cases; data stewards, researchers, platform services | Faster understanding, indexing and validation of heterogeneous data | Inferred metadata may be incomplete or wrong; compute cost | Demonstrated components; early user benefit | Steward review, confidence and provenance, minimum metadata |
| **Legal and ethical metadata** | All use cases; data providers and users | More responsible and efficient reuse decisions | Metadata may become stale or be treated as legal advice | Draft schema | Accountable steward, update process and clear limitations |
| **In-dataset discovery and Text-to-SQL** | Weather and other relational datasets; non-specialists and analysts | Lower technical barrier and faster content exploration | Semantic SQL errors and false confidence | Prototype/early implementation | Source display, validation and uncertainty, expert review for consequential use |
| **Cross-dataset discovery and packaging** | All use cases; researchers, SMEs, public bodies | Finds complementary resources and reduces search effort | Ranking bias, poor profiles and latency | Early implementation | Representative benchmarks, relevance explanations and diversity checks |
| **Query recommendation and disambiguation** | All use cases; especially non-technical users | Better expression of needs and guided exploration | Steering, profile bias and loss of user control | Architecture/prototypes | Optional use, visible alternatives, rejection and correction |
| **Dataset recommendation** | MathE, language and general platform | Increases reuse of related resources | Popularity or metadata bias | Initial metadata-based design | Explain why recommended and allow non- |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | personalised operation |
| **Analytical patterns and MoMa** | All use cases; developers and researchers | Reusable workflows, provenance and cross-component integration | Project-specific lock-in and versioning | Implemented foundations | Open specifications, export and external mappings |
| **Explanations and provenance** | MathE, weather and platform-wide | Trust, auditability, learning and error correction | Explanations may be technically correct but unusable | SQL explanation service reported | Layered, user-tested explanations linked to evidence |
| **Adaptive quiz and MCQ generation** | MathE; students, teachers, administrators | Personalised support and reduced content-preparation effort | Accuracy, pedagogical bias and student profiling | Pilot-stage early evidence | Mandatory teacher validation, transparency and outcome evaluation |
| **Skill-gap and training matching** | Lifelong learning; workers, managers, trainers | Reduced manual matching and targeted learning | Employment discrimination and opaque taxonomies | Mainly anticipated | Separate intended purposes, employee rights and fairness testing |
| **Weather geospatial analysis and anomaly detection** | Scientists, planners, civil protection and sectors | Faster climate insight and risk awareness | Uncertainty, scale mismatch and operational over-reliance | Functional first pilot and initial user testing | Product labels, provenance and integration with expert protocols |
| **Multilingual corpus exploration** | Researchers, cultural institutions, educators | Lower effort and broader multilingual research | Format loss, rights and high-resource language bias | Functional pilot and initial user evaluation | Preserve originals, language-specific evaluation and rights checks |

# Appendix II Proposed M36 impact evidence framework

| Impact question | Suggested indicator | M18 baseline | Evidence to collect for D1.7 |
|---|---|---|---|
| **Does DataGEMS reduce the effort required to find relevant data?** | Median time to first suitable dataset, successful-request rate and discovery precision on representative tasks | Architecture, prototypes and Grant Agreement targets are available; no consolidated comparative M18 result was supplied | Pre-defined task set, baseline comparator, system logs, user success criteria and error analysis by use case |
| **Does profiling improve understanding and validation?** | Time to understand or validate a dataset, profile correction rate and proportion of mandatory metadata complete | Profiling services and early use-case outputs are reported | Before-and-after steward study, sampled profile audit, corrections and confidence or provenance records |
| **Does the platform increase responsible reuse and interoperability?** | Datasets linked, exported or reused outside their original context, including successful external mappings | Croissant extensions, MoMa and dataset-linking foundations are documented | Reuse cases, export tests, external repository or EOSC integration evidence and documented information loss |
| **Are natural-language and recommendation functions useful and inclusive?** | Task completion, acceptance, rejection and correction rates, accessibility findings and performance by language and user group | Initial qualitative feedback and component prototypes are reported | Representative user testing, accessibility review, subgroup analysis and examples of failure or abstention |
| **Does MathE create educational value without weakening teacher control?** | Teacher preparation time, MCQ correction rate, student comprehension, learning gain where methodologically appropriate and complaint or error rate | Pilot implementation and direct participant engagement are reported; learning outcomes are not yet demonstrated | Teacher validation logs, controlled or quasi-experimental study where feasible, student feedback and accessibility analysis |
| **Does lifelong learning support voluntary and organisational users fairly?** | Skill-gap accuracy, explanation usefulness, training uptake and distribution of recommendations across relevant groups | Intended scenarios are defined but specific datasets and direct M18 user evidence remain limited | Separate evaluations for learner and employer contexts, taxonomy audit, contestability testing and indirect-discrimination analysis |

| | | | |
|---|---|---|---|
| **Are weather outputs fit for their stated decision context?** | Accuracy, latency, provenance completeness, uncertainty comprehension and documented reliance by operational users | ERA5 preparation, Text-to-SQL and staged functionality are reported | Domain validation, source and uncertainty labels, user studies and explicit distinction between exploratory and operational services |
| **Can language-data workflows scale beyond the initial corpora?** | Onboarding time, conversion loss, retrieval quality by language, compute cost and rights-clearance effort | UC4P1 is reported at pilot stage with later pilots in development or design | New-corpus onboarding studies, language-specific benchmarks, provenance records, licensing review and resource accounting |
| **Is the service sustainable and portable?** | Uptime, latency, incident rate, cost per task, deployment effort, component ownership and exit or export success | Architectural security and workflow controls are documented; consolidated operational and business evidence is incomplete | Service-level records, dependency register, deployment profiles, maintenance commitments, licences, repositories and financing plan |
| **Has wider ecosystem adoption occurred?** | Number and diversity of external adopters, repeated use, co-investment or public-private cooperation and barriers resolved | Early community and pilot engagement is reported, but external adoption is not yet evidenced consistently | Verified adoption cases, letters or agreements, usage records, adopter interviews and evidence of continued use after initial support |

# Appendix III List of acronyms and abbreviations

| Acronym | Meaning |
|---|---|
| AAI | Authentication and Authorisation Infrastructure |
| AI | Artificial intelligence |
| ALTAI | Assessment List for Trustworthy Artificial Intelligence |
| AP | Analytical pattern |
| ARC | Athena Research Center |
| CNRS | Centre national de la recherche scientifique |
| DPIA | Data Protection Impact Assessment |
| EOSC | European Open Science Cloud |
| FAIR | Findable, Accessible, Interoperable and Reusable |
| HLEG | High-Level Expert Group on Artificial Intelligence |
| IDS | Leibniz Institute for the German Language |
| IPB | Polytechnic Institute of Bragança |
| KPI | Key performance indicator |
| LLM | Large language model |
| M18 | Month 18 |
| M36 | Month 36 |
| MCQ | Multiple-choice question |
| MoMa | Model of Metadata and data |
| NOA | National Observatory of Athens |
| NLP | Natural language processing |
| PU | Public dissemination level |
| SME | Small and medium-sized enterprise |
| TRL | Technology readiness level |
| UC | Use case |
| WP | Work package |
| ZHAW | Zurich University of Applied Sciences |

# Appendix IV References

## 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, Dataset Profiling Design Report, 31 October 2025.
5. DataGEMS, Dataset Discovery Architecture Report, 31 October 2025.
6. DataGEMS, User Assistance Architecture Report, 31 October 2025.
7. DataGEMS, Data Storage, Management and Access Architecture Report, 31 October 2025.
8. DataGEMS, Pilot Requirements Report, 13 October 2025.
9. DataGEMS, Pilot Implementation Evaluation Report, 30 June 2026
10. DataGEMS project website, https://datagems.eu/ (accessed 26 June 2026).


## Official policy and legal sources


11. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.
12. European Commission, AI Act - regulatory framework and application timeline, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed 26 June 2026).
13. 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.
14. Regulation (EU) 2016/679 (General Data Protection Regulation), https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.
15. Regulation (EU) 2018/1807 on a framework for the free flow of non-personal data, 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), 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. 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.
20. European Commission High-Level Expert Group on AI, Assessment List for Trustworthy Artificial Intelligence (ALTAI), 2020, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.