DataGEMS

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



DMP

D1.4
Data Management Plan

Public

3/7/2025

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 Coordinator 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 Coordinator 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	DMP
Identifier	D1.4
Title	Data Management Plan
Dissemination level	Public
Due Date	30/6/2025
Actual Delivery Date	3/7/2025
Lead Beneficiary	ATHENA RC
Other beneficiaries	All partners
Editor	Stavroula Eleftheraki
Authors	Stavroula Eleftheraki, Giorgos Papanikos, Georgios Kakaletris, Mike Xydas, Eleni Petra, Effie Zafeirakopoulou, Ana Pereira, Vasiliki Kotroni, Friedemann Schwenkreis, Thora Hagen, Cristina Contero Almagro
Contributors	
Reviewer	Sihem Amer-Yahia, Silviu Maniu
Workpackage(s	WP1 – Management and Coordination
Keywords	Data Management, Innovation Management, Intellectual Property Rights
Version	V1.0
Version information	Number of pages: 56 Last file save: 03/07/2025 5:44 AM File revision: 66 File name: D1.4_M6_Data_Management_Plan.docx

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	3/7/2025	Submitted to EU	Stavroula Eleftheraki

Document Tables

Table of Tables

Table 1: Dataset for Assessing Mathematics Learning - UC1 (Higher Education) 25
Table 2: MathE Database - UC1 (Higher Education) 26
Table 3: MathE Integration Techniques Resources - UC1 (Higher Education) 27
Table 4: ESCO Dataset - UC2 (Lifelong Learning) 28
Table 5: Era5land Dataset - UC3 (Weather) 29
Table 6: MeteoStations Climate Dataset - UC3 (Weather) 30
Table 7: MeteoStations Online Dataset - UC3 (Weather) 30
Table 8: Geodata Dataset - UC3 (Weather) 31
Table 9: EncycNet Dataset - UC4 (Language Data Infastructure) 32
Table 10: Britannica Dataset - UC4 (Language Data Infastructure) 33
Table 11: Diderot Dataset - UC4 (Language Data Infastructure) 33
Table 12: Wikipedia Dataset - UC4 (Language Data Infastructure) 34

Table of Figures

Figure 1: Argos 11

Table of Contents

Copyright Information 3
1 Introduction 8
1.1 Background and Purpose of the Deliverable 8
1.2 Overview of the Deliverable 8
2 Data Management Policies 10
2.1 Internal Consortium Policy 10
2.1.1 Data Management Responsible 11
2.1.2 Data Ownership and Responsibilities 12
2.1.3 Long-Term Preservation and Archiving 12
2.1.4 Data Security 13
2.2 Ensuring FAIR Data Compliance 14
2.2.1 Making Data Findable 15
2.2.2 Making Data Accessible 15
2.2.3 Ensuring Interoperability 15
2.2.4 Enhancing Reusability 16
3 Source Code & Executable Software Policy 17
4 Reports and Publications Policy 19
5 Ethical Considerations and Privacy Policy 20
5.1 GDPR and Privacy Compliance 20
5.2 Ethics and Machine Learning 21
5.3 Consent and User Rights 21
6 Overview of Data Types and Dataset Sources in DataGEMS 23
7 Conclusions 35
Appendix I Data Management Plan Information 36
Appendix II References 55

Abstract

This deliverable presents the first Data Management Plan (DMP) of the DataGEMS project, outlining the framework for managing data and research outputs across the project lifecycle. It sets out how the project will ensure data is handled responsibly, securely, and in alignment with FAIR principles and Horizon Europe's Open Science guidelines. The plan defines roles and responsibilities, policies for data access, licensing, preservation, and ethical compliance, while addressing ethical and legal considerations. It covers a diverse set of datasets collected or reused across four domain-specific use cases (higher education, lifelong learning, weather, and language data) and describes mechanisms for integrating external datasets and user-uploaded content. DataGEMS' DMP will be continuously updated to reflect project developments and will contribute to the long-term accessibility and reuse of project outcomes through EOSC-compliant repositories. The final version of the DMP (Deliverable D1.5) will be submitted in Month 34 of the project.

1 Introduction

1.1 Background and Purpose of the Deliverable

DataGEMS (Data Discovery Platform with **G**eneralized **E**xploratory, **M**anagement, and **S**earch Capabilities) aims to simplify and enhance the discovery, management, and use of open datasets across different domains. By developing a novel platform, DataGEMS offers intelligent tools that support data profiling, metadata enrichment, search, and exploration, with the goal of making open data more **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR) [29]. The project brings together expertise in data management, data science, machine learning, and knowledge engineering to address the challenges of scattered, incomplete, and heterogeneous open data.

This deliverable corresponds to the work performed under Task 1.5, "Data and Research output Management Planning", of Work Package 1 (Management and Coordination). **It defines how DataGEMS will manage its data, software, workflows, and other research outcomes throughout the project lifecycle.** Following best practices for FAIR data management, this Data Management Plan (DMP) lays out our strategy for responsible handling of research results, including data storage, access, sharing, licensing, preservation, and protection.

Effective data and research output management are critical to ensure the quality, transparency, and long-term impact of DataGEMS outcomes. This DMP also supports the project's commitment to Open Science [9] by promoting open access to publications, open sourcing of software where appropriate, and clarification of authorship and intellectual property rights. DMP will be continuously updated and finalized by the end of the project, with the objective to deposit it in EOSC-compliant repositories such as Zenodo [30] or OpenAIRE [22], contributing to the European Open Science ecosystem through its integration into the OpenAIRE Graph [17] that is the authoritative source for the European Open Science Cloud (EOSC).

1.2 Overview of the Deliverable

The document is structured as follows:

- **Section 2** details the internal data governance policies adopted by the consortium, including roles and responsibilities, data ownership, licensing, security measures, and preservation strategies.
- **Section 3** addresses the management of source code and software components, emphasizing open-source practices, versioning, licensing, and long-term archiving.
- **Section 4** sets out the project's policy on reports and publications, promoting open access and alignment with FAIR and Open Science principles, as well as AI ACT compliance.
- **Section 5** covers ethical considerations, privacy protection, GDPR compliance, and procedures for informed consent during user studies. More details will be provided in the upcoming deliverable D1.8 Ethics Compliance (M18) and its updated version D1.9 Ethics Compliance (M36).
- **Section 6** provides a detailed overview of dataset types and sources in DataGEMS.
- **Section 7** offers concluding remarks and points toward the ongoing evolution of the DMP throughout the project.

Appendix I provides a detailed breakdown of dataset management practices and FAIR compliance strategies tailored to each of the project's four pilot use cases (i.e., specific application scenarios focused on higher education, lifelong learning, weather data, and language data infrastructure).

This DMP serves both as a practical guide for consortium partners and as a living document that will be updated regularly to reflect new datasets, evolving practices, and alignment with Horizon Europe expectations and EOSC integration pathways.

2 Data Management Policies

This section outlines the Data Management Strategy adopted by DataGEMS to ensure all project-related data is Findable, Accessible, Interoperable, and Reusable (FAIR). This strategy combines internal policies with established Open Data guidelines [19] and platforms.

2.1 Internal Consortium Policy

As DataGEMS develops innovative tools for dataset discovery, profiling, and reuse, it must also address the complexities of managing diverse and heterogenous data sources in a responsible and consistent manner. These challenges include handling licensing terms, ensuring proper access control, and planning for long-term data preservation. To meet these needs, the project has established an internal data management policy that promotes adherence to FAIR principles and reflects the requirements of Horizon Europe Open Science guidelines.

For the management, authoring, distribution, and deposition of DataGEMS DMP, the ARGOS service [21], offered freely by OpenAIRE, will be used. DataGEMS will not only comply with the Open Access publication policy and the provisioning of a DMP but also adopts a robust path to support the FAIR research data principle; this deliverable being the first step in that direction.

The DataGEMS' DMP adheres to the European Commission guidelines for FAIR data and covers the full lifecycle of the data being gathered, processed and/or generated by the project by collecting information on:

- the handling of (research) data during and after the end of the project,
- the particular datasets to be collected, processed and/or generated and their nature,
- the methodology, technologies and standards that will be applied for data packaging, handling, description, sharing etc.,
- the licensing / restrictions on data managed by the project,
- the policies with respect to data curation and quality,
- the policy with respect to data preservation beyond the project duration.

ARGOS is an open, extensible service designed to simplify the management, validation, monitoring, and maintenance of DMPs. Provided by OpenAIRE, it enables researchers, project managers, and other stakeholders to generate actionable DMPs that are interoperable across infrastructures and consistent with data owners' intentions and responsibilities. Figure 1 shows a screenshot of the ARGOS plaform.

To support the adoption of the platform, DataGEMS has been working closely with OpenAIRE experts since the start of the project. This ongoing collaboration includes live presentations, online training sessions, and continuous guidance on using the platform effectively.

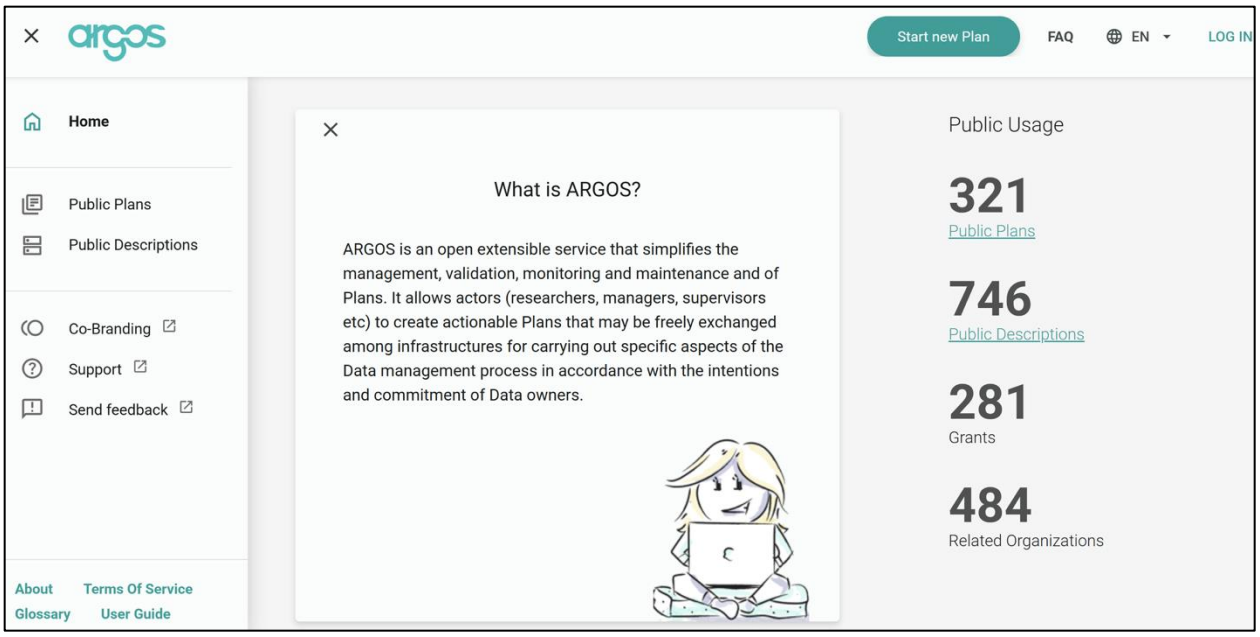


Figure 1: Argos

Argos is based on OpenCDMP [23], an open-source software platform for actionable, machine-readable Data Management Planning. It has been developed in alignment with the specifications and developments of the RDA Working Group on maDMPs.

2.1.1 Data Management Responsible

DataGEMS does not assign a single individual to handle all aspects of Data Management. Instead, responsibilities are distributed among key project members, under the coordination of WP5 Leader, Prof. Matteo Lissandini (UVR). In his role as the Data Management Manager, he oversees and coordinates all activities related to data ownership and fosters awareness among partners regarding their data management responsibilities.

In addition, data management planning for the project's four Use Cases is assigned to four designated project members, each responsible for their respective area:

- Higher Education: Prof. Ana Pereira (IPB)
- Weather: Dr. Vassiliki Kotroni (NOA)
- Lifelong Learning: Prof. Dr. Friedemann Schwenkreis (DHBW)
- Language Data Infrastructure: Thora Hagen (IDS)

Responsibilities concerning software will be led by the Technical Manager of the project, Prof. Dr. Kurt Stockinger (ZHAW).

Finally, data management planning for research publications will be coordinated by the Project Coordinator, Dr. Georgia Koutrika (ARC).

2.1.2 Data Ownership and Responsibilities

As part of the Data Management Plan, DataGEMS will establish an inventory of results that will collect, document, and store all research outcomes of the project.

DataGEMS Inventory of Results will play a critical role in the following areas:

- **Documentation and Tracking of Research Outcomes.** All research outcomes of project partners will be systematically recorded to ensure full traceability, facilitate internal coordination, and provide a comprehensive overview of project achievements across WPs.
- **Innovation Opportunity Analysis and Exploitation Planning.** The Innovation Manager, in collaboration with the WP leaders, will preliminarily analyze each service, based on the information collected and mapped in the DataGEMS Inventory. For each solution, tailored innovation management and Intellectual Property Protection (IPR) strategies may be explored. These strategies will outline the main steps for preparing the exploitation of the solution, identifying potential business models, and defining the main actions for IPR. Further details can be found in **Deliverable 1.2 "Innovation Management and Sustainability Plan"**.
- **Access Policies and Licensing Strategy.** WP leaders are responsible for identifying the core competences associated with each technology. Based on this input, the innovation manager and DataGEMS' decision bodies will determine the appropriate level of dissemination and protection. The possible categories are as follows:
  - **Public**: This means that it can be shared with the external stakeholders and users. Public information about the technologies does not diminish the product's value or compromise the competitive advantage of the associated technology.
  - **Protected by Non-disclosure Agreement (NDA)**: Certain core competence elements may be shared with external stakeholders under the terms of an NDA. While this information is important for fostering collaboration and unlocking added value, it forms a key part of the technology's competitive edge and must remain confidential.
  - **Protected by IPR**: Decisions regarding IPR protection will be made in consultation with the relevant project partners, the Technical Manager, and the Project Coordinator.

2.1.3 Long-Term Preservation and Archiving

The DataGEMS long-term preservation policy covers both research artefacts, and their associated metadata generated or managed throughout the project lifecycle. The strategy for long-term preservation and archiving is built on the following key principles:

- All data and metadata will be preserved in formats compatible with the supported solutions for at least the contractual duration of the project. An additional five (5) years of extended preservation is targeted as best practice. Special provisions will be made for datasets subject to retention limits or requiring periodic ethical review (particularly those from the higher education and lifelong-learning use cases). These measures aim to account for long-term risks such as re-identification or data misuse.
- Data and metadata repositories such as Zenodo (an open-access repository that allows researchers from all disciplines to share and preserve their research outputs, regardless of size or format) are expected to be engaged for long-term preservation and archiving on a best-effort basis. These repositories should also ensure transparency regarding the strategies and practices they employ to support this objective.

- Data and metadata managers (i.e., the persons listed above) must adopt open standards (such as widely accepted protocols and formats) that minimize the risk of obsolescence and ensure long-term usability.
- The security policies of all tools used for preservation must support at minimum the most restrictive access levels required to protect the stored data and metadata.
- Data managers are responsible for securing the necessary rights to preserve and share data, and for ensuring these rights are properly enforced within the preservation system. They must also ensure that the preservation system's protection mechanisms meet at least the minimum requirements set by the copyright owners of the data. License documentation is part of the preserved metadata to avoid legal ambiguities in the future.

The use of additional depositing or cataloguing services is encouraged for all project partners, provided these services can enhance the long-term preservation of data and metadata, and fully comply with the data protection requirements established by the respective data owner or manager. Particular attention must be paid to the lawful basis for processing personal data, as well as to the technical and organizational measures in place to protect it, such as the use of pseudonymization when full anonymization is not feasible. DataGEMS partners recognize the high importance of selecting trusted and certified repositories that adhere to FAIR principles and establish standards for long-term preservation.

2.1.4 Data Security

Operating within the framework of Open Science and FAIR data principles, data security is sometimes mistakenly perceived as less critical. In reality, it is a fundamental pillar of responsible data management; one that presents additional challenges due to the diverse and overlapping policy requirements across disciplines. This results in a broad landscape of options that must be carefully aligned with the specific needs of a project like DataGEMS, which operates within a focused scope and limited timeframe. Accordingly, and based on (a) the requirements posed by the pilots, (b) the general importance of data security, and (c) the technological relevance in the project, in the following we present the cornerstones of data security policy in DataGEMS.

Data confidentiality. Data confidentiality applies to both data and metadata, although the policies governing them can differ significantly. The inclusion of ethical review procedures and redaction guidelines for metadata (especially internal/test datasets or metadata involving vulnerable groups) will be carefully considered. DataGEMS strongly suggests the **open publication of metadata** for any data that may be reused, whether within or beyond the project, and regardless of the access level granted to the data itself. However, temporal or internal reference data may differ from this rule. In such cases, the metadata must at least be registered in the project's internal research product catalogue, and (depending on specific policies) may also be made openly accessible to third parties.

In addition to metadata provision, it is essential that a declarative, comprehensible license is provided. If targeting open access under restrictions, the use of a Creative Commons 4.0 license variant is recommended. However, data owners are free to select any existing license or define a new one that better suits their needs. It is also critical that a clear statement of data ownership accompanies the license. While actionable licensing is preferred, it is acknowledged that many systems may not yet support this capability effectively.

Regarding data access, DataGEMS promotes open access as the default policy. However, since much of the data involved stems from prior or ongoing work and may be subject to restrictive licensing, the final decision rests with the data owner. Therefore, owners may choose between private, restricted, or open access (with or without specific limitations). In case data are kept private, they may not be accessed by other users inside or outside the DataGEMS ecosystem. Restricted use may have different meanings depending on the developed solution. Open access on the other hand, can be provided also under restrictions, as for instance a requirement for user registration or limited use of data, etc. DataGEMS will embed a transparent and well-documented justification process for all data access decisions. This process will help ensure that restrictions are applied consistently, lawfully, and with appropriate reasoning.

Another aspect of data protection option is the establishment of an embargo period on fresh data. This helps protect the interests of the data owner while still enabling future reuse by the research community. In open access contexts, common embargo periods range from 6 to 12 months, with up to 24 months in some cases.

Data dependability. Another aspect that data security policy in DataGEMS refers to data trustworthiness. This is addressed with policies in several directions:

- Data integrity needs to be safeguarded both for data and metadata by storage service and catalogues and, if possible, further augmented by author seals.
- Authorship shall be well registered and tacked so that repudiation or any other relevant false claims are prevented.
- Versioning of data / metadata is ensured so that consumers may precisely address the artefact in question.
- Metadata description is adequate to support introspection of data and accurate consumption of the latter.
- Provenance is tracked adequately to support comprehension of data origins, generation procedures and ownership respect.
- Strong referencing of data is provided so that any further claims may be routed appropriately.
- Long term preservation is enabled allowing data to be dependable in the future.

2.2 Ensuring FAIR Data Compliance

This section of the Data Management Plan aims to address the following questions:

- Will the produced data and those used in the project be discoverable through metadata, and identifiable and accessible using standard identifiers (e.g., DOIs)?
- Will provisions be made to enable the re-use of data?
- Will clear versioning be provided to track changes over time?
- What metadata will be produced?

In response to these questions, the FAIR Data Management policy is articulated through the following principles and implementation plan.

2.2.1 Making Data Findable

The DataGEMS platform is specifically designed to improve the findability of datasets by providing advanced browsing and keyword search functionalities. During the project, both existing datasets and newly generated datasets will be integrated into the platform. All datasets used and produced will be documented with comprehensive and standardized metadata to ensure clarity, traceability, and compliance with the FAIR principles. This approach will facilitate dataset discovery and reuse by both human users and automated systems. We will follow the steps detailed below:

- Most datasets used in the project will be sourced from existing open repositories. When reusing such datasets, we will retain metadata annotations from the original repository wherever possible. This includes recording the source, any persistent identifiers (e.g., DOIs), and the data and time of data acquisition. The original search terms, if available, will also be preserved.

- All datasets will be versioned to clearly distinguish between the original form and any processed versions (e.g., filtered). Metadata for such datasets will document the transformation process, including the methods, tools, and parameters used. A human-readable file (e.g., README.md) will accompany these datasets, detailing the entire transformation pipeline to ensure transparency and reproducibility.

- Zenodo [30] will serve as one of the project's trusted repositories for dataset deposition. Its metadata schema complies with that of DataCite [4], enriched where appropriate. Metadata records are indexed in both Zenodo's search engine and DataCite's registry as part of DOI assignment. Each dataset published via Zenodo will receive a globally unique and persistent identifier (PID), ensuring long-term accessibility and citability.

- For datasets involving sensitive content, a metadata review process will be conducted before publication to ensure appropriate handling (e.g., description, access policies, etc.).

2.2.2 Making Data Accessible

For datasets reused from external repositories, the DataGEMS portal will provide direct references to the original sources. For datasets generated within the project and are intended to be public, datasets will be deposited in Zenodo [30], using standard formats such as CSV, JSON, or PDF, ensuring broad compatibility and long-term accessibility. Zenodo guarantees persistent access to deposited content, with datasets remaining available for the lifetime of the repository.

Prior to granting open access to any dataset and its metadata, a careful assessment of data protection requirements and other legal constraints (e.g., licensing terms, intellectual property rights, and privacy considerations) will be conducted to ensure full regulatory compliance.

2.2.3 Ensuring Interoperability

To support interoperability, the project will rely on open standards, structured metadata, and widely supported formats for both datasets and metadata.

Regarding metadata, Zenodo uses JSON Schema as its internal representation and supports export to commonly used formats such as Dublin Core [6] or MARCXML. Where needed,

interoperability will be enhanced through the provision of dedicated APIs for importing and exporting content in standard formats (e.g., XML and JSON) and using standard data models. Metadata for datasets will follow an enriched version of the MLCommons Croissant vocabulary, which aims to standardize how Machine Learning datasets are described to improve discovery and usability across tools and platforms [31]. Croissant builds upon the Schema.org Dataset type.

As for the datasets themselves, only widely supported formats (e.g., CSV, JSON, and PDF) will be used to ensure broad accessibility and ease of reuse.

2.2.4 Enhancing Reusability

The public content within the DataGEMS will be available for download and re-use with no restrictions or embargo. The content will be available under permissive licenses, (CC-BY 4.0, CC-0 or comparable) but certain conditions (e.g. Non-commercial use, or NC) and/or exceptions may also apply. In Zenodo, the code is open source and built on the foundation of the Invenio digital library [15] which is also open source. The work-in-progress, open issues, and roadmap are shared openly in GitHub. All meta data is openly available under CC0 license, and all open content is openly accessible through open APIs when necessary. License is one of the mandatory terms in Zenodo's metadata and is referring to an Open Definition license [20]. Data downloaded by the users is subject to the license specified in the metadata by the uploader. To support responsible reuse, DataGEMS will include clear licensing information, and where applicable, will provide disclaimers and/or a code of conduct to guide third-party users. This will help ensure ethical use, attribution, and compliance with intended usage conditions.

Open Access Policy. Scientific publications, expected from almost all project activities, set another field of regulation in the project. DataGEMS aims for 100% Open Access, as mandated by Horizon Europe. DataGEMS currently has a target of 50% of publications to be filed under Gold Open Access approach and the other 50% will be filed under Green Open Access (self-archived with the support of OpenAIRE repositories). The project will reserve financial resources for achieving open access in cases where this is not the de facto situation.

3 Source Code & Executable Software Policy

DataGEMS recognizes that its software is not only a set of tools for helping users, but also a key part of sharing and advancing research knowledge. By ensuring open access to its source code, DataGEMS supports the core values of Open Science: enabling others to repeat, reproduce, and build upon its results, and contributing to a more open and trustworthy research community. As such, Free and Open-Source Software (FOSS) licensing is encouraged in the project, with public code releases adopting permissive licenses such as MIT License, Apache 2.0 License [1], or EUPL [8], which are widely used and reduce potential conflicts with external or proprietary components. If licensing conflicts arise, the selected license may be revisited during the project.

DataGEMS adopts a centralized, open, and structured approach to software management using a source code management platform for all its source code that is licensed under an Open-Source License. The selected platform is GitHub [13].

DataGEMS suggests that all such Open-Source code is hosted under the dedicated DataGEMS GitHub Organization [5], each component deposited into a separate repository. Partners may opt in for managing Open-Source code in other repositories given that those comply with the processes of the project, for building software as well as access to the public. In this case, a partner may also opt for mirroring code under DataGEMS GitHub Organization. Repository structures are standardized, and developers may be assisted by a shared Cookiecutter [2] template to ensure consistent, well-documented, and maintainable codebases across all software components. Each public repository will include a LICENSE file clearly stating the terms of use. Note that proprietary or early-stage components may remain private until they mature for public release.

Furthermore, as it has been stated above, to ensure long-term preservation and open access, stable software releases will be archived in Zenodo; GitHub repositories will be connected to Zenodo, enabling the automatic generation of Digital Object Identifiers (DOIs) for each release. This approach not only guarantees persistent citation of software artifacts but also integrates them into the OpenAIRE Graph [17], strengthening DataGEMS' contribution to the European Open Science ecosystem. As such, software deposits in Zenodo [28] will include at a minimum:

- The source code (excluding binaries and sensitive information),
- A README file (with package name, description, contributors, links),
- A LICENSE file (with copyright and licensing).

Deposits will be packaged in common archive formats (e.g., zip, tar.gz). The granularity of deposits will aim for self-contained, usable modules rather than partial fragments. It is expected that software deposits will occur at major milestones, once services are considered mature enough to support reproducible research. At least one final deposit per Open-Source component will be made before the end of the project. Through this combined use of GitHub for code change management and versioning and Zenodo for preservation, DataGEMS ensures its software outputs are compliant with FAIR principles, supporting the broader Open Science objectives of transparency, sustainability, and accessibility.

Software development practices in DataGEMS follow widely recognized best practices to ensure quality, reproducibility, and sustainability:

- Code reviews are available for all pull requests, helping by having changes peer-reviewed before integration.
- Semantic versioning is applied systematically to manage software releases and track evolution over time.
- Testing practices include basic smoke tests, unit tests, and regression tests are applied to ensure that new changes do not break existing functionalities.

It must be noted that security best practices are also applied in source code management, with sensitive information managed externally via various techniques (e.g. environment variables and secret management tools).

Besides Open-Source software, other software components may be included in DataGEMS software platform. Those may include pre-existing components of partners licensed under different models or 3rd party software components that may not follow the Open-Source paradigm. Those may be distributed in the context of DataGEMS under executable or linkable forms, such as technology-dependent libraries, containerized tools or accessible service APIs.

Licensing of those executable components of partners will allow their unrestricted use in the context of their purpose inside the project and its products, unless otherwise explicitly presented and reasoned. One such case includes virtual computational power and storage provisioning services provided by 3rd parties, or platforms provided on top of such virtual resources.

Docker-compliant packaging will be used to containerize applications, supporting consistent deployment across development, staging and operation environments.

Details on the project's software development and contribution practices are detailed in **Deliverable 6.1 “Platform Development Manual”**.

4 Reports and Publications Policy

DataGEMS will produce a wide range of written materials, including scientific publications, project reports (e.g., management, technical, and dissemination), training materials (e.g., slides and tutorials), technical documentation (e.g., API descriptions and system specifications), and project dissemination artifacts (e.g., brochures, posters, newsletters, and blog posts).

All public-facing written materials will be made openly accessible through the project's website and/or trusted repositories such as Zenodo (or institutional platforms) compliant with OpenAIRE harvesting standards. The latter approach ensures long-term preservation, assignment of persistent identifiers (DOIs), and integration into the OpenAIRE Graph [17], enhancing discoverability across the European Open Science ecosystem. Additionally, selected dissemination content may be shared via the project's social media channels to broaden outreach and engagement. Confidential deliverables, on the other hand, will be submitted only through the EC project management platform and will not be made publicly accessible.

Scientific publications will be submitted to peer-reviewed journals and conferences, with a strong preference for open access venues whenever possible. In line with Horizon Europe Open Science guidelines [10], publications will aim to be licensed under a Creative Commons Attribution (CC BY) or equivalent open license, enabling unrestricted access, reuse, and redistribution with proper attribution. DataGEMS will actively seek to publish in journals or conferences that support open access, either directly (Gold Open Access) or through self-archiving of accepted manuscripts (Green Open Access), in support of the project's KPIs. In all cases, DataGEMS will pursue the highest possible level of openness to maximize dissemination, reuse, and long-term visibility of its research outputs. Publications will include standard metadata, including the title, authors, affiliations, project acknowledgments, funding references (DataGEMS Grant Agreement number), Digital Object Identifiers (DOIs), and author ORCID identifiers (if available). Proper citation practices [26] will be followed to ensure traceability and recognition of contributions.

For materials with potential social impact, an ethical review process will be applied to identify and mitigate risks such as biases, harmful content, or the reinforcement of stereotypes. This process will also promote inclusivity in language and examples, and ensure that accessibility considerations are addressed, including for individuals with disabilities. These safeguards will help ensure that public-facing content aligns with the project's values of diversity, equity, and responsible communication.

5 Ethical Considerations and Privacy Policy

5.1 GDPR and Privacy Compliance

This section of the Data Management Plan addresses the following questions:

- Are there any ethical or legal issues that may affect data sharing?
- Do questionnaires (or other data collection tools) clearly ask for participants' consent to share and preserve their personal data over the long term?

DataGEMS processes data that may fall under the definition of personal data as established by the *EU General Data Protection Regulation (GDPR)*. According to GDPR, personal data is any information relating to an identified or identifiable natural person (data subject), whereby *"an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"*.

Research conducted within DataGEMS extends beyond the academic sector. The DataGEMS platform supports collaboration across a broad and diverse research ecosystem that includes researchers, software developers, entrepreneurs, commercial entities, and government bodies. These stakeholders may interact with the platform or contribute data, making data governance and privacy compliance particularly important. In alignment with the European Data Protection Supervisor [11], DataGEMS recognises that the processing of personal data for the purposes of academic expression encompasses:

1. Processing that is directly linked to the academic freedom to share knowledge.
2. The unrestricted distribution of research findings, e.g., through publications.
3. The sharing of data and methods with peers for the exchange of views and opinions.

While the recognition of academic freedom and openness is essential, this approach alone does not ensure compliance with GDPR, as its specific requirements should still be fully implemented. DataGEMS will maintain appropriate safeguards and observe data protection requirements throughout the data lifecycle. This includes adherence to key GDPR principles such as transparency, data minimisation, and purpose limitation. These measures will ensure that personal data is processed lawfully and responsibly, upholding GDPR compliance while supporting open and responsible research across diverse sectors.

In response to the second question, DataGEMS takes a comprehensive approach to GDPR compliance, embedding data protection principles throughout the project's lifecycle and across all relevant activities. In line with transparency requirements, DataGEMS developed a privacy policy during the first six months, outlining key safeguards such as data subjects' rights and communication channels. DataGEMS also plans to implement governance, technical, and procedural measures appropriate to the level of risk to ensure adequate protection of personal data. An External Ethical Advisor has been appointed to further assist with identifying and documenting the appropriate legal bases for all instances of personal data processing in accordance with Article 6 of the GDPR, including consent, legitimate interest, and public interest

where applicable. Data Protection Impact Assessments (DPIAs) will be conducted for high-risk processing activities, particularly in pilots involving potentially sensitive information or vulnerable populations. Core principles such as data minimisation, storage limitation, and purpose limitation will be applied systematically. DataGEMS is also designing mechanisms to uphold the rights of data subjects, including access, rectification, erasure, objection, and data portability and procedures to fulfil them. Taken together, these measures demonstrate that GDPR compliance in DataGEMS is not merely a matter of documentation but is operationalised through concrete practices that support legal, ethical, and responsible data handling across diverse sectors.

5.2 Ethics and Machine Learning

DataGEMS recognizes the ethical considerations involved in the development and operation of AI data-driven services, both at the level of system design and real-world deployment. DataGEMS follows the European Commission's non-binding Ethics Guidelines for Trustworthy AI [7], adopting a human-centric approach that upholds European values and fundamental rights:

> *"The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoys a unique and inalienable moral status."*

These principles guide both the technical implementation and governance of the DataGEMS platform, ensuring that all components are developed and validated in line with the key requirements for trustworthy AI and responsible data services [16].

To move from principles to practice, DataGEMS will conduct targeted ethical assessments of its AI components throughout the development lifecycle. These will include risk-based evaluations of potential impacts on fundamental rights, fairness, and transparency. Based on the outcomes of these assessments, appropriate technical and organizational measures will be defined and applied, such as explainability tools, bias mitigation strategies, and human oversight mechanisms. These efforts aim to ensure that AI-enabled services developed within DataGEMS remain lawful, ethical, and aligned with EU values and societal expectations.

5.3 Consent and User Rights

While DataGEMS does not routinely process personal data, such data may be collected during pilot evaluations, user studies, and co-collaboration activities with the use case partners, particularly within **Work Package (WP) 7 (Co-Creation and Evaluation through Pilots)**.

WP7 is responsible for the design and implementation of four pilot use cases and associated user studies (Tasks 7.1–7.5), which may involve the collection of both quantitative and qualitative human-subject data such as user feedback and survey responses.

DataGEMS partners collecting personal data will act as data controllers for that specific activity, meaning they are responsible for deciding why and how the data is processed. Each partner will collaborate with their institutional **Data Protection Officer (DPO)** to develop appropriate informed

consent procedures. These procedures will include the preparation of clear, accessible, and GDPR-compliant consent forms that explicitly specify the purpose, scope, and duration of data collection. Consent will be obtained in a freely given, unambiguous, informed and specific manner, with clear options for withdrawal at any time. Transparency will be ensured through the appropriate notices and within the consent statement. A procedure for consent withdrawal will also be designed and implemented. All procedures will be coordinated with WP7 to ensure that ethical and legal standards are upheld across all pilot activities.

In addition to the pilot-specific procedures described above, the DataGEMS platform will also be accessible to external users. For this broader user base, a dedicated list of user consents will be compiled covering aspects such as account registration, profile data handling, and activity logging. This effort will ensure transparency and alignment with GDPR requirements and will be carried out in collaboration with the platform's technical team and the respective institutional DPOs.

Finally, logging is designed to support troubleshooting, security investigations, and usage accounting, while remaining GDPR compliant. Personal data is avoided unless strictly necessary, and where needed, user consent and transparency are ensured. Pseudonymization approaches are applied by using internal identifiers disjoint from federated identities. Retention periods are purpose-driven, minimized for troubleshooting, and extended where necessary for audit, accounting or security, in line with data protection principles.

6 Overview of Data Types and Dataset Sources in DataGEMS

The DataGEMS platform is designed to support the discovery, profiling, enrichment, and exploration of open datasets from diverse domains (e.g., higher education). The datasets handled within DataGEMS can be categorized into three main types:

- **Structured**:
  Data with a well-defined, explicit schema describing the structure and types of data elements, typically represented in tabular forms.
- **Semi-structured**:
  Data that contains some organizational structure (e.g., using hierarchies), but without a strict schema like structured data.
- **Unstructured**:
  Data with no predefined organizational structure or consistent formatting (e.g., images).

Datasets are ingested, processed, and displayed through the DataGEMS platform's components, including data profilers, metadata enrichment tools, and exploratory search interfaces.

The **DataGEMS Dataset Catalogue** integrates datasets from three main sources:

- **Use Case (UC) Provider Datasets:** Identified by project partners in accordance with the platform's four core use cases:
  - Higher Education: IPB (Instituto Politécnico de Bragança)
  - Lifelong Learning: DHBW (Baden-Wuerttemberg Cooperative State University)
  - Weather: NOA (National Observatory of Athens)
  - Language Data Infrastructure: IDS (Leibniz-Institut für Deutsche Sprache)

  These datasets will be developed, tested, and used for pilot evaluation purposes and are summarized in the tables that follow. Key details include licensing, format, access level, and any specific handling requirements (e.g., privacy or ethical constraints). Note that some of these datasets are private and may only be partially visible or used internally, with selective public releases following anonymization and curation.
- **User-Uploaded Datasets via Onboarding**: External users can contribute datasets through an onboarding process. During this process, users are guided through a structured metadata submission form that ensures each dataset is described according to FAIR principles and aligned with the DMP policies defined in this deliverable.
- **External Catalogue Integration:** DataGEMS can also integrate datasets from external sources, such as the OpenAIRE Research Catalogue. Integration with these external sources is permitted only if they meet the necessary DMP and FAIR data requirements.

Datasets from UC providers and those uploaded via the onboarding process are physically hosted on DataGEMS infrastructure, specifically within SCAYLE [27], the consortium's primary infrastructure provider. Datasets in relational formats are stored in a PostgreSQL [25] environment managed within SCAYLE. All other formats, such as files, documents, and non-relational formats, are stored securely in OpenCAYLE [24] (object storage solution also hosted by SCAYLE).

To meet its Key Performance Indicator (KPI) of supporting at least 300 datasets, DataGEMS will continuously expand its catalogue over the course of the project. Beyond its UC-specific scope,

the platform is designed to serve a broader community of users, including researchers, data scientists, public institutions, and others interested in exploring and reusing open datasets. To address this objective, as previously mentioned, DataGEMS will also incorporate external datasets from diverse domains, e.g., cultural heritage.

Data profiling activities will generate descriptive metadata and support basic quality assessment to ensure the usability and reliability of all datasets (internal and external) integrated into the platform. These descriptions, profiling results, and metadata will be progressively produced and maintained through the activities of **Work Package 2 (Dataset Profiling).**

Generated Data. In addition to these datasets displayed in the platform, DataGEMS also produces new types of data artifacts as part of its enrichment and reasoning processes. These include Machine Learning (ML) models and Analytical Patterns. These new types constitute instances of the MoMa model (MoMa stands for unified Model for data Management and discovery and represents a semantic graph that integrates and connects all assets in the DataGEMS ecosystem.). These generated data will also be described according to FAIR principles and aligned with the DMP policies defined in this deliverable. Additionally, the selection, use, and deployment of ML models will comply with the EU AI ACT and associated ethical and legal requirements.

UC1: Higher Education

Table 1: Dataset for Assessing Mathematics Learning - UC1 (Higher Education)

Property	Description
Name	Dataset for Assessing Mathematics Learning in Higher Education
Lead Partner	IPB
Description	This dataset was extracted from the MathE platform, an online educational platform developed to support mathematics teaching and learning in higher education. It contains 546 student responses to questions on several mathematical topics. Each record corresponds to an individual answer and includes the following features: Student ID, Student Country, Question ID, Type of Answer (correct or incorrect), Question Level (basic or advanced based on the assessment of the contributing professor), Math Topic (broader mathematical area of the question), Math Subtopic, and Question Keywords. The data spans from February 2019 to December 2023.
License	CC0 1.0
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	CSV
Dataset type	Structured (plain tabular)
Dataset size	1.1 MB
Accessible at	https://dados.ipb.pt//dataset.xhtml?persistentId=doi:10.34620/dadosipb/PW3OWY

Table 2: MathE Database - UC1 (Higher Education)

Property	Description
Name	MathE Database
Lead Partner	IPB
Description	The MathE database is a structured relational dataset derived from the MathE platform. It integrates information across multiple dimensions, including user profiles, student performance, questions, learning materials, and educational metadata (e.g., topics of questions).
License	N/A
Availability	This dataset is private due to the absence of formal consent from all question authors and the presence of sensitive student information.
Special content handling requirements	This dataset includes metadata associated with students (e.g., name, country, etc.). Any use of this data must involve appropriate anonymization to ensure that no individuals can be directly or indirectly identified. Compliance with GDPR and ethical data handling principles is required.
Data format	SQL
Dataset type	Structured (relational)
Dataset size	10.7 MB
Accessible at	N/A

Table 3: MathE Integration Techniques Resources - UC1 (Higher Education)

Property	Description
Name	Integration/Integration Techniques Questions MathE Database
Lead Partner	IPB
Description	MathE is a mathematical platform developed under the MathE project (mathe.ipb.pt) for higher education. The dataset associated with the Integration topic and the Integration techniques subtopic has 61 questions, the correct answer and two incorrect answers with common errors made by students, keywords associated and material association. The database is organised as follows: Question ID, topic name; subtopic name; question; Correct answer, Incorrect answer 1, incorrect answer 2, keywords, Videos, Material.
License	CC0 1.0
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	CVS and pdf folder
Dataset type	Unstructured
Dataset size	127 MB
Accessible at	N/A

Table 4: MathE Pedagogic Resources - UC1 (Higher Education)

Property	Description
Name	Pedagogic Resources MathE Database
Lead Partner	IPB
Description	MathE is a mathematical platform developed under the MathE project (mathe.ipb.pt) for higher education. The dataset is based on the 410 pdf documents organized by keywords in a cvs file.
License	CC0 1.0
Availability	Public
Special content handling requirements	N/A (fully open data).
Data format	Pdf and cvs file
Dataset type	Unstructured
Dataset size	127 MB
Accessible at	mathe.ipb.pt

UC2: Lifelong Learning

Table 4: ESCO Dataset - UC2 (Lifelong Learning)

Property	Description
Name	ESCO
Lead Partner	DHBW
Description	The European Skills, Competences, Qualifications, and Occupations (ESCO) dataset functions as a standardized classification system for occupations, skills, and qualifications across the European labor market. It acts as a structured dictionary, defining skills and linking them to relevant job roles and qualifications.
License	CC-BY 4.0
Availability	Public
Special content handling requirements	N/A, provided for public use by the ESCO project.
Data format	Has various formats but the full dataset is available only in RDF: SKOS-RDF (ttl), ODS, and CSV
Dataset type	Semi-structured (RDF) and structured (ODS and CSV)
Dataset size	~1.2 GB
Accessible at	ESCO Service Portal

UC3: Weather

Table 5: Era5land Dataset - UC3 (Weather)

Property	Description
Name	Era5land
Lead Partner	NOA
Description	A global atmospheric reanalysis dataset produced by the European Centre for Medium-Range Weather Forecast's (ECMWF) and has data available from 1950, providing a consistent view of the evolution of land variables. It has an enhanced resolution of 0.1° x 0.1°, while the temporal frequency of the model output is hourly.
License	Copernicus data licence
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	GRIB or NetCDF4
Dataset type	Structured
Dataset size	Depending on selected variables, spatial coverage, and time period; can reach petabytes for full global coverage over the entire time span (1950–present).
Accessible at	https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=download

Table 6: MeteoStations Climate Dataset - UC3 (Weather)

Property	Description
Name	MeteoStations Climate
Lead Partner	NOA
Description	Daily values of 53 meteorological stations in Greece. Values include temperature, wind, precipitation, relative humidity.
License	CC-BY-SA 4.0 [3]
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	CSV
Dataset type	Structured (plain tabular)
Dataset size	~50 MB
Accessible at	https://data.climpact.gr/dataset/497dc26d-45e0-4ad5-b8f3-5f8890f65129

Table 7: MeteoStations Online Dataset - UC3 (Weather)

Property	Description
Name	MeteoStations Online
Lead Partner	NOA
Description	Live data collected from a network of ~550 meteorological stations across Greece. The dataset is updated every 10 minutes and includes key measurements such as temperature, wind, precipitation, and surface pressure.
License	N/A
Availability	Private. The organization does not provide open access to the dataset, as the data are not yet quality controlled. Access may be granted upon request after quality control procedures have been applied.
Special content handling requirements	Access to data is restricted. Any further use or dissemination of the data must be coordinated with NOA in advance.
Data format	GeoJSON
Dataset type	Semi-structured
Dataset size	~10 MB
Accessible at	-

Table 8: Geodata Dataset - UC3 (Weather)

Property	Description
Name	Geodata
Lead Partner	NOA
Description	This dataset includes geospatial information used to derive complex indices related to the geographic characteristics of different areas. It comprises: (a) A high resolution Digital Elevation Model (DEM) released by NASA, with a spatial resolution of 30 m. (b) land use datasets (CORINE 5.1.1), offering a pan-European inventory of land cover and land use across 44 thematic classes, ranging from broad forested areas to specific land types, such as vineyards.
License	N/A
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	GeoTIFF
Dataset type	Structured
Dataset size	~3GB (for 2-3 European countries)/ ~200 GB (for the whole European domain)
Accessible at	https://stratus.meteo.noa.gr/data/DataGEMs/

UC4: Language Data Infrastructure

Table 9: EncycNet Dataset - UC4 (Language Data Infastructure)

Property	Description
Name	EncycNet
Lead Partner	IDS
Description	This datasets comprises digitized versions of 22 historical German encyclopedias from the early 18th to early 20th century, which includes over 49,300,000 word tokens. The dataset is accompanied by a set of images from the encyclopedias and a knowledge graph that was constructed from a portion of the encyclopedia dataset.
License	CC BY 4.0
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	XML (text), PNG (images), RDF (graph)
Dataset type	Semi-structured (XML), unstructured (images), structured (RDF graph)
Dataset size	757 MB (text only), 7.7 GB (including pictures), 160 MB (graph)
Accessible at	https://zenodo.org/records/4554112 (text and images) https://doi.org/10.5281/zenodo.10219192 (graph)

Table 10: Britannica Dataset - UC4 (Language Data Infastructure)

Property	Description
Name	19th century knowledge project
Lead Partner	IDS
Description	This dataset contains machine-readable transcriptions of historical editions of the Encyclopedia Britannica, currently the 7th edition (1830-1842, 21 volumes) and 9th edition (1875-1889, 25 volumes). It provides structured encyclopedic articles covering scientific, historical, philosophical, and cultural topics from a 19th-century perspective.
License	CC BY 4.0
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	XML or TXT
Dataset type	Semi-structured (XML) or unstructured (plain text)
Dataset size	988 MB
Accessible at	https://tu-plogan.github.io/source/r_releases.html

Table 11: Diderot Dataset - UC4 (Language Data Infastructure)

Property	Description
Name	Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers
Lead Partner	IDS
Description	A digitized version of Diderot's encyclopedia (1751-1772), extracted from Wikisource.
License	CC BY-SA 4.0
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	JSON
Dataset type	Semi-structured
Dataset size	117 MB
Accessible at	https://github.com/ThoraHagen/languagedatagems/releases/tag/v1.0

Table 12: Wikipedia Dataset - UC4 (Language Data Infastructure)

Property	Description
Name	Wikipedia
Lead Partner	IDS
Description	A subset of wikipedia entries in English, German, and French that overlap with the EncycNet corpus (EncycNet dataset). Each entry is linked to corresponding EncycNet corpus IDs.
License	CC BY-SA 4.0
Availability	Public
Special content handling requirements	N/A (fully open data)
Data format	JSON
Dataset type	Semi-structured
Dataset size	2 GB
Accessible at	https://github.com/ThoraHagen/languagedatagems/releases/tag/v1.0

7 Conclusions

The DataGEMS Data Management Plan (DMP) aims to ensure responsible, transparent, FAIR-compliant, GDPR compliant, and AI ACT-compliant handling of research data, software, and research outcomes throughout the lifecycle of the project. The DMP establishes clear governance structures, assigning data management responsibilities across the consortium members and use case leaders, while emphasizing compliance with Horizon Europe Open Science policies.

Throughout the project, DataGEMS addresses the complexity of managing diverse data sources (structured, semi-structured, and unstructured) collected or integrated from internal use cases, external sources, and user-contributed datasets. The plan specifies detailed policies for data ownership, access control, licensing, long-term preservation, and ethical compliance, including GDPR adherence for any personal data collected during pilot activities.

DataGEMS leverages trusted infrastructures for storage, preservation, and publication, including SCAYLE, Zenodo, and GitHub, ensuring persistent identification, interoperability, and discoverability of data assets. Metadata schemas extend established standards such as schema.org, supporting machine-readable indexing and enhancing the reuse potential of datasets beyond the project's immediate scope.

Scientific publications will be submitted to peer-reviewed journals and conferences, with a strong preference for Open Access venues whenever possible. For software and data, Open Access is also the general objective; however, legal, ethical, or technical constraints may in some cases require specific licensing terms or access limitations. In all cases, the project aims to ensure maximum transparency, reproducibility, and responsible management of research outputs.

Finally, it is emphasized that this DMP represents a living document. As the project evolves, additional datasets will be integrated, FAIR-compliant processes will mature, and, when applicable, best practices with respect to the AI ACT, will continue to be adopted in alignment with Horizon Europe and EOSC guidelines. In the next phase of the project, the ARGOS platform will be adopted to further formalize, maintain, and update this DMP, ensuring long-term accessibility, traceability, and alignment with current best practices for data management.

Appendix I Data Management Plan Information

This appendix contains the detailed data management questionnaires completed for each of the DataGEMS Use Cases (UCs). The structure of these questionnaires follows the Horizon Europe Data DMP template [14] and addresses key aspects of research data handling, including data sources, FAIR principles, data access, interoperability, reuse, security, and ethical considerations.

The information provided in these questionnaires complements the general data management policies described in the main body of this DMP and ensures that the specific characteristics and needs of each UC are fully documented in compliance with the Horizon Europe requirements.

In the next phase of the project and in future versions of this deliverable, the ARGOS platform will be adopted to further support the formalization, versioning, and maintenance of these Data Management Plans in alignment with Horizon Europe standards.

I.1 UC1: Higher Education

I.1.1 Data Summary

Will you re-use any existing data and what will you re-use it for?
Yes. The MathE platform's existing data (e.g., student responses, usage logs, learning outcomes) will be reused to develop and validate data exploration services.

What types and formats of data will the project generate or re-use?

- CSV
- SQL dump (platform database)
- PDF (educational material)

What is the purpose of the data generation or re-use and its relation to the objectives of the project?
To develop and evaluate data analytics services for exploration, profiling, and semantic enrichment within the DataGEMS platform and enable personalized learning scenarios in the Higher Education pilot.

What is the expected size of the data that you intend to generate or re-use?
All available on MathE database (~ 150 MB).

What is the origin/provenance of the data, either generated or re-used?
All data originates from the MathE platform, managed by IPB.

To whom might your data be useful ('data utility'), outside your project?
Researchers and data scientists.

I.1.2 FAIR Data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?
Yes, openly published datasets will be assigned DOIs via Zenodo.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?
DataGEMS dataset metadata schema extends the MLCommons Croissant [12] schema which in turn extends the Dataset entity of schema.org. The purpose of the Croissant schema is to make datasets easier to discover and use.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?
Yes, appropriate keywords describing the learning content.

Will metadata be offered in such a way that it can be harvested and indexed?
Yes. Metadata will be OAI-PMH [18] compatible and machine-readable through Zenodo or institutional repositories.

Making data accessible

Will the data be deposited in a trusted repository?
Yes, public datasets will be deposited in Zenodo. Restricted datasets will be managed via IPB's institutional repository.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?
One dataset has already been deposited in IPB's institutional repository, which assigns DOIs and supports public access. However, other datasets have not yet been formally deposited in any public repository. The project is currently exploring further arrangements for dataset deposition in Zenodo, focusing on those that are public and pseudonymized. Sensitive datasets will remain hosted internally at IPB under restricted access, subject to GDPR compliance.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?
Yes, Zenodo provides DOIs and landing pages for all deposits.

Will all data be made openly available? If certain datasets cannot be shared, explain why.
Only pseudonymized datasets will be made public. Personally identifiable or sensitive data will not be shared due to GDPR compliance.

If an embargo is applied, specify why and how long this will apply.
At present, no embargo has been applied to the public datasets deposited. For future datasets, embargoes may be considered on a case-by-case basis to allow for additional processing, anonymization, or publication preparation.

Will the data be accessible through a free and standardized access protocol?
Data in DataGEMS will be accessible through our dataset catalogue. Access and downloading of data will be allowed for free if the user has logged in via EGI Check-In.

If there are restrictions, how will access be provided to the data?

Access-controlled datasets will be available upon request and subject to approval by the responsible use case provider: Ana Pereira (apereira@ipb.pt).

How will the identity of the person accessing the data be ascertained?
The identity of the person accessing the data will be ascertained through the mandatory requirement of the user logging in through EGI Check-In.

Is there a need for a data access committee?
Yes, a committee coordinated by IPB will review restricted data requests.

Will metadata be made openly available and licensed under a public domain dedication (CC0), as per the Grant Agreement?
Metadata will be openly licensed under CC0 and publicly available when possible. For datasets that cannot be made publicly available the metadata will follow the same license and publishing rules.

How long will the data remain available and findable? Will metadata remain available after data is no longer accessible?
To be defined. Metadata will remain accessible even if the data are no longer hosted.

Will documentation or reference to software needed to access or read the data be included?
Yes, where applicable, references to open-source tools or exported data schema will be included.

Making data interoperable

What data and metadata vocabularies, standards, formats, or methodologies will you follow?
The ML Commons Croissant standard which builds upon the Dataset entity of schema.org.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
For each data type, specific metadata instances will be structured as data assets under the Croissant schema. Should any extensions or project-specific elements become necessary, they will be carefully aligned with existing standards and openly published to support reuse, refinement, and interoperability.

Will your data include qualified references to other data?
Yes. Datasets will include links to related course materials.

Increase data re-use

How will you provide documentation needed to validate and reuse the data?
Each dataset will be accompanied by README files and schema descriptions.

Will your data be made freely available in the public domain? Will it be licensed using standard reuse licenses?
Where legally and ethically possible, data will be published under CC-BY-NC. Where legally and ethically possible, metadata will be under CC0.

Will the data be usable by third parties, particularly after the project ends?
Yes. Data will be structured and documented to support long-term reuse.

Will the provenance of the data be thoroughly documented?
Yes. Provenance information will be part of the metadata and project deliverables.

Describe data quality assurance processes.
To be defined.

I.1.3 Other Research Outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Will other research outputs (software, protocols, etc.) be generated?
Yes. These outputs will be documented and made publicly available when possible.

I.1.4 Allocation of Resources

What will the costs be for making data or other research outputs FAIR compliant in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?
N/A

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).
N/A

Who will be responsible for data management in your project?
The person responsible for the MathE data management is Ana Pereira (apereira@ipb.pt).

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long).
N/A

I.1.5 Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?
Will be stored in secure managed storage for a limited period.

Will the data be safely stored in trusted repositories for long term preservation and curation?

Yes, Zenodo and IPB's repository both meet these criteria.

I.1.6 Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?
Yes. The use of student data requires informed consent and strict pseudonymization to comply with GDPR.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?
Yes.

I.1.7 Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?
N/A

I.2 UC2: Lifelong Learning

I.2.1 Data Summary

Will you re-use any existing data and what will you re-use it for?
We will re-use existing data from other projects as well as from associate partners. This data will be used to generate models as well as for analysis and as a basis for the generation of sets of result documents.

What types and formats of data will the project generate or re-use?
The set of types and formats will be determined as part of the project.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?
To develop and evaluate data analytics services for exploration, profiling, and semantic enrichment within the DataGEMS platform as well as the generation of use case specific solutions.

What is the expected size of the data that you intend to generate or re-use?
The size of generated and re-used data will be determined as part of the project.

What is the origin/provenance of the data, either generated or re-used?
The origin/provenance of re-used data will be determined as part of the project. The origin of generated data are the systems of DataGEMS.

To whom might your data be useful ('data utility'), outside your project?
End-user communities will be identified on behalf of the pilot definitions (see for instance D7.1). At this point five different groups habe been identified: researchers, HR executives, managers, workers and trainers.

I.2.2 FAIR Data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?
Yes, openly published datasets will be assigned DOIs via Zenodo.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?
DataGEMS dataset metadata schema extends the MLCommons Croissant[1] schema which in turn extends the Dataset entity of schema.org. The purpose of the Croissant schema is to make datasets easier to discover and use.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?
Yes, appropriate keywords describing the learning content.

Will metadata be offered in such a way that it can be harvested and indexed?
Yes. Metadata will be OAI-PMH compatible and machine-readable through Zenodo.

[1] https://github.com/mlcommons/croissant

Making data accessible

Will the data be deposited in a trusted repository?
Yes, public datasets will be deposited in Zenodo.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?
This needs to be done as part of the project.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?
Yes, Zenodo provides DOIs and landing pages for all deposits.

Will all data be made openly available? If certain datasets cannot be shared, explain why.
So far, all datasets provided in the context of this use case are openly available. If additional datasets are introduced later that contain sensitive or third-party content, appropriate access restrictions will be applied in accordance with legal and ethical requirements.

If an embargo is applied, specify why and how long this will apply.
At this point we have not identified any needs for an embargo concept in the context of the use case.

Will the data be accessible through a free and standardized access protocol?
Data in DataGEMS will be accessible through our dataset catalogue. Access and downloading of data will be allowed for free if the user has logged in via EGI Check-In.

If there are restrictions, how will access be provided to the data?
So far, there are no restrictions.

How will the identity of the person accessing the data be ascertained?
The identity of the person accessing the data will be ascertained through the mandatory requirement of the user logging in through EGI Check-In.

Is there a need for a data access committee?
Yes, there is a need to a data access committee.

Will metadata be made openly available and licensed under a public domain dedication (CC0), as per the Grant Agreement?
Metadata will be openly licensed under CC0 and publicly available when possible. For datasets that cannot be made publicly available the metadata will follow the same license and publishing rules.

How long will the data remain available and findable? Will metadata remain available after data is no longer accessible?
To be defined. Metadata will remain accessible even if the data are no longer hosted.

Will documentation or reference to software needed to access or read the data be included?

Yes, where applicable, references to open-source tools or exported data schema will be included.

Making data interoperable

What data and metadata vocabularies, standards, formats, or methodologies will you follow?
The ML Commons Croissant standard which builds upon the Dataset entity of schema.org.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
For each data type, specific metadata instances will be structured as data assets under the Croissant schema. Should any extensions or project-specific elements become necessary, they will be carefully aligned with existing standards and openly published to support reuse, refinement, and interoperability.

Will your data include qualified references to other data?
Data in skills models will reference data contained in competency taxonomies.

Increase data re-use

How will you provide documentation needed to validate and reuse the data?
Each dataset will be accompanied by README files and schema descriptions.

Will your data be made freely available in the public domain? Will it be licensed using standard reuse licenses?
Where legally and ethically possible, data will be published under CC-BY-NC. Where legally and ethically possible, metadata will be under CC0.

Will the data be usable by third parties, particularly after the project ends?
Yes. Data will be structured and documented to support long-term reuse.

Will the provenance of the data be thoroughly documented?
Yes. Provenance information will be part of the metadata and project deliverables.

Describe data quality assurance processes.
To be defined.

I.2.3 Other Research Outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Will other research outputs (software, protocols, etc.) be generated?
Yes. These outputs will be documented and made publicly available when possible.

I.2.4 Allocation of Resources

What will the costs be for making data or other research outputs FAIR compliant in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?
N/A

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).
N/A

Who will be responsible for data management in your project?

While overall data management responsibilities are coordinated at the project level, each use case designates a responsible contact to ensure consistent implementation of the DMP. For the Lifelong Learning use case, Prof. Dr. Friedemann Schwenkreis (DHBW) oversees data management activities, including compliance with FAIR principles and GDPR, coordination of data deposition, and alignment with project-wide data governance.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long).
N/A

I.2.5 Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?
Will be stored in secure managed storage for a limited period.

Will the data be safely stored in trusted repositories for long term preservation and curation?
Yes, Zenodo.

I.2.6 Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?
This needs to be evaluated on behalf of the project.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?
We do not plan any questionnaires at this point.

I.2.7 Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?
Needs to be determined on behalf of the project.

I.3 UC3: Weather

I.3.1 Data Summary

Will you re-use any existing data and what will you re-use it for?
Yes. We will reuse several well-established meteorological datasets:

- ERA5-Land: Re-used for calculating standardized climate indices (e.g., heatwaves, heavy rainfall) over historical periods and for supporting multi-decadal trend analysis and spatial overlays (Open Data)
- MeteoStations Climate Dataset (NOA's historical station data): Re-used for national-scale validation and local indices generation (Open Data)
- MeteoStations Online (live data from NOA's network): Re-used to enable real-time querying and semantic summarization (Restricted Access Data)
- Geodata: Re-used to enrich weather and climate data analysis with spatial context (e.g., exposure of population to climate extremes) (Open Data)

What types and formats of data will the project generate or re-use?

- Structured: CSV, NetCDF, GRIB (climate observations, reanalysis)
- Semi-structured: GeoJSON (live streaming sensor data)
- Raster: GeoTIFF (geospatial elevation and land cover data)

What is the purpose of the data generation or re-use and its relation to the objectives of the project?
To enable **intelligent, FAIR, and user-friendly exploration of climate and weather data**, directly aligned with DataGEMS' overarching goals of enhancing data discovery, management, and semantic querying across domains.

What is the expected size of the data that you intend to generate or re-use?
~300 GB

What is the origin/provenance of the data, either generated or re-used?

- ERA5-Land: European Centre for Medium-Range Weather Forecasts (ECMWF) via the Copernicus Climate Change Service (C3S). Access: Public, via the Copernicus Climate Data Store (CDS).
- NOA datasets: National Observatory of Athens (NOA), METEO Unit. Access: Provided via a service developed by NOA or linked from the Climpact Project. The real time data sets will be accessed through an API.
- Geodata: NASA SRTM (via Earthdata); European Environment Agency (Copernicus Land Monitoring Service); European Environment Agency (based on census and LUCAS data);

To whom might your data be useful ('data utility'), outside your project?

- Environmental scientists and meteorologists
- Civil protection services
- Climate Scientists and Researchers
- Data infrastructures handling time-series and geospatial data

I.3.2 FAIR Data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?
Yes. Public datasets already have DOIs. For subsets created within the project, they will be deposited to Zenodo.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?
DataGEMS dataset metadata schema extends the MLCommons Croissant[2] schema which in turn extends the Dataset entity of schema.org. The purpose of the Croissant schema is to make datasets easier to discover and use.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?
Yes. Examples include: “meteorological data”, “ERA5”, “NOA climate”, “geospatial”, “DEM”, “weather station data”.

Will metadata be offered in such a way that it can be harvested and indexed?
Yes. Metadata will be OAI-PMH compatible and machine-readable through Zenodo.

Making data accessible

Will the data be deposited in a trusted repository?
Yes. Derived datasets will be deposited in Zenodo. Original datasets remain hosted on official portals (Copernicus, NOA).

Have you explored appropriate arrangements with the identified repository where your data will be deposited?
Yes. Zenodo supports large datasets with associated metadata and DOIs. Access to real-time NOA data will be restricted to the DataGEMs system.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?
Yes, DOIs provided by Zenodo are resolvable and indexed.

Will all data be made openly available? If certain datasets cannot be shared, explain why.
Most datasets are public. However, real-time NOA sensor data is private due to the lack of quality control and institutional policy.

If an embargo is applied, specify why and how long this will apply.
N/A

Will the data be accessible through a free and standardized access protocol?
Data in DataGEMS will be accessible through our dataset catalogue. Access and downloading of open data, for this use case, will be allowed for free if the user has logged in via EGI Check-In.

If there are restrictions, how will access be provided to the data?

[2] https://github.com/mlcommons/croissant

DataGEMs system will have access to restricted data through an API. External users will not have access to these data.

How will the identity of the person accessing the data be ascertained?
The identity of the person accessing the data will be ascertained through the mandatory requirement of the user logging in through EGI Check-In.

Is there a need for a data access committee?
Not at this stage.

Will metadata be made openly available and licensed under a public domain dedication (CC0), as per the Grant Agreement?
Metadata will be openly licensed under CC0 and publicly available when possible. For datasets that cannot be made publicly available the metadata will follow the same license and publishing rules.

How long will the data remain available and findable? Will metadata remain available after data is no longer accessible?
To be defined. Metadata will remain accessible even if the data are no longer hosted.

Will documentation or reference to software needed to access or read the data be included?
Yes, where applicable, references to open-source tools or exported data schema will be included.

Making data interoperable

What data and metadata vocabularies, standards, formats, or methodologies will you follow?
The ML Commons Croissant standard which builds upon the Dataset entity of schema.org.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
For each data type, specific metadata instances will be structured as data assets under the Croissant schema. Should any extensions or project-specific elements become necessary, they will be carefully aligned with existing standards and openly published to support reuse, refinement, and interoperability.

Will your data include qualified references to other data?
N/A

Increase data re-use

How will you provide documentation needed to validate and reuse the data?
Each dataset will be accompanied by README files and schema descriptions.

Will your data be made freely available in the public domain? Will it be licensed using standard reuse licenses?

All open data, for this use case, have their own licensing.

Will the data be usable by third parties, particularly after the project ends?
Data will be structured and documented to support long-term reuse.

Will the provenance of the data be thoroughly documented?
Yes. Provenance information will be part of the metadata and project deliverables.

Describe data quality assurance processes.
To be defined.

I.3.3 Other Research Outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Will other research outputs (software, protocols, etc.) be generated?
Yes. These outputs will be documented and made publicly available when possible.

I.3.4 Allocation of Resources

What will the costs be for making data or other research outputs FAIR compliant in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?
N/A

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).
N/A

Who will be responsible for data management in your project?
Mr. Antonis Bezes (NOA).

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long).
N/A

I.3.5 Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?
Will be stored in secure managed storage for a limited period.

Will the data be safely stored in trusted repositories for long term preservation and curation?
Yes, Zenodo.

I.3.6 Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?
No. The datasets involve environmental measurements and do not include sensitive data.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?
Not applicable.

I.3.7 Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?
N/A

I.4 UC4: Language Data Infrastructure

I.4.1 Data Summary

Will you re-use any existing data and what will you re-use it for?
Yes. Several historical and modern encyclopedic corpora and lexical datasets are reused:

- EncycNet (22 digitized historical German encyclopedias with a knowledge graph)
- Britannica (machine-readable 19th-century editions)
- Diderot (digitized French Enlightenment encyclopedia)
- Wikipedia subset (English, German, and French aligned to EncycNet topics)

What types and formats of data will the project generate or re-use?

- XML and TXT (textual encyclopedic content)
- PNG (illustrations and scans)
- RDF (graph)

What is the purpose of the data generation or re-use and its relation to the objectives of the project?
To develop and evaluate data analytics services for exploration, profiling, and semantic enrichment within the DataGEMS platform and support Language Data Infrastructure pilot.

What is the expected size of the data that you intend to generate or re-use?
~12GB

What is the origin/provenance of the data, either generated or re-used?

- EncycNet: The data originates from historical German encyclopedias digitized by Directmedia Publishing and made available through platforms Zeno.org and the TextGrid Repository. It has since been curated and published openly on Zenodo.
- Britannica: This dataset was compiled from digitized editions of the *Encyclopædia Britannica* available through the HathiTrust Digital Library, Internet Archive, and Project Gutenberg. The digitization was carried out by various university and national libraries. The current version is maintained and accessible via GitHub.
- Diderot: The data is based on the French *Wikisource* edition of Diderot and d'Alembert's *Encyclopédie*, retrieved through scraping methods. Extracted and versioned releases are hosted on GitHub.
- Wikipedia: This dataset consists of selected entries from the English, German, and French Wikipedias. The content was sourced via the MediaWiki API and aligned with the EncycNet corpus; it is available on GitHub. Additional material comes from openly available Wikipedia dumps.

To whom might your data be useful ('data utility'), outside your project?
Researchers in digital humanities, lexicography, AI and NLP communities working on historical corpora, ontology alignment, or semantic web applications.

I.4.2 FAIR Data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?
Yes. All datasets are published on Zenodo or GitHub with persistent DOIs or release tags.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed?
DataGEMS dataset metadata schema extends the MLCommons Croissant[3] schema which in turn extends the Dataset entity of schema.org. The purpose of the Croissant schema is to make datasets easier to discover and use.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?
Yes. Metadata will include keywords such as “encyclopedia”, “digitized text”, and “historical corpus”.

Will metadata be offered in such a way that it can be harvested and indexed?
Yes. records are OAI-PMH compatible and machine-readable through Zenodo.

Making data accessible

Will the data be deposited in a trusted repository?
Yes. All datasets are already published on Zenodo or institutional GitHub repositories with stable release identifiers.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?
Yes. Zenodo has been confirmed as the primary repository for data deposition and is fully compliant with EOSC guidelines. Additionally, intermediate or supplementary data products may be deposited on GitHub repositories.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?
Yes. All Zenodo and GitHub releases have DOIs or persistent URLs.

Will all data be made openly available? If certain datasets cannot be shared, explain why.
Yes. All data is fully open and in the public domain or under CC-BY(-SA) 4.0 licenses.

If an embargo is applied, specify why and how long this will apply.
N/A

Will the data be accessible through a free and standardized access protocol?
Data in DataGEMS will be accessible through our dataset catalogue. Access and downloading of data will be allowed for free if the user has logged in via EGI Check-In.

If there are restrictions, how will access be provided to the data?
No access restrictions apply.

How will the identity of the person accessing the data be ascertained?

[3] https://github.com/mlcommons/croissant

The identity of the person accessing the data will be ascertained through the mandatory requirement of the user logging in through EGI Check-In.

Is there a need for a data access committee?
Not applicable.

Will metadata be made openly available and licensed under a public domain dedication (CC0), as per the Grant Agreement?
Metadata will be openly licensed under CC0 and publicly available when possible. For datasets that cannot be made publicly available the metadata will follow the same license and publishing rules.

How long will the data remain available and findable? Will metadata remain available after data is no longer accessible?
To be defined. Metadata will remain accessible even if the data are no longer hosted.

Will documentation or reference to software needed to access or read the data be included?
Yes, where applicable, references to open-source tools or exported data schema will be included.

Making data interoperable

What data and metadata vocabularies, standards, formats, or methodologies will you follow?
The ML Commons Croissant standard which builds upon the Dataset entity of schema.org.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
For each data type, specific metadata instances will be structured as data assets under the Croissant schema. Should any extensions or project-specific elements become necessary, they will be carefully aligned with existing standards and openly published to support reuse, refinement, and interoperability.

Will your data include qualified references to other data?
Yes, in particular to Wikidata/Wikipedia identifiers, some of which are already available. Wikidata linking then offers references to other standards such as GND or VIAF that could potentially be made explicit.

Increase data re-use

How will you provide documentation needed to validate and reuse the data?
Each dataset will be accompanied by README files and schema descriptions.

Will your data be made freely available in the public domain? Will it be licensed using standard reuse licenses?
Yes. Data is under CC BY 4.0 or CC BY-SA 4.0.

Will the data be usable by third parties, particularly after the project ends?
Yes. Data will be structured and documented to support long-term reuse.

Will the provenance of the data be thoroughly documented?
Yes. Provenance information will be part of the metadata and project deliverables.

Describe data quality assurance processes.
To be defined.

I.4.3 Other Research Outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Will other research outputs (software, protocols, etc.) be generated?
Yes. These outputs will be documented and made publicly available when possible.

I.4.4 Allocation of Resources

What will the costs be for making data or other research outputs FAIR compliant in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?
N/A

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).
N/A

Who will be responsible for data management in your project?
Thora Hagen, IDS (Leibniz Institute for the German Language)

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long).
N/A

I.4.5 Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?
Will be stored in secure managed storage for a limited period.

Will the data be safely stored in trusted repositories for long term preservation and curation?
Yes Zenodo and Github.

I.4.6 Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?
No. All data is public domain or openly licensed and does not include personal data.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?
Not applicable.

I.4.7 Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?
N/A

Appendix II References

[1] Apache License: https://www.apache.org/licenses/LICENSE-2.0

[2] Cookiecutter: https://www.cookiecutter.io/

[3] Creative Commons Licenses: https://creativecommons.org/licenses/

[4] DataCite's Metadata Schema: https://schema.datacite.org/

[5] DataGEMS GitHub: https://github.com/datagems-eosc

[6] Dublin Core Ontology: https://www.dublincore.org/resources/glossary/ontology/

[7] Ethics guidelines for trustworthy AI: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[8] EUPL: https://eupl.eu/

[9] European Commission: Directorate-General for Research and Innovation. (2015). Open innovation, open science, open to the world: a vision for Europe. Publications Office. https://data.europa.eu/doi/10.2777/061652.

[10] European Commission: European Innovation Council and SMEs Executive Agency. (2024). European IP helpdesk: your guide to open science in Horizon Europe. Publications Office of the European Union. https://data.europa.eu/doi/10.2826/943044.

[11] European Data Protection Supervisor, 2020. Preliminary Opinion on data protection and scientific research

[12] GitHub Croissant: https://github.com/mlcommons/croissant

[13] GitHub: https://github.com/

[14] Horizon Europe Data Management Plan Template. https://cohortcoordinationboard.eu/toolkit/horizon-europe-data-management-plan-template/

[15] Invenio digital library: https://invenio-software.org/

[16] MADIEGA, Tambiama André. "EU guidelines on ethics in artificial intelligence: Context and implementation." (2019).

[17] Manghi, Paolo, et al. "The OpenAIRE research graph data model." Zenodo (2019).

[18] Open Archives: https://www.openarchives.org/pmh/

[19] Open data: https://ec.europa.eu/digital-single-market/en/open-data

[20] Open Definition license: https://opendefinition.org/

[21] OpenAIRE Argos: https://argos.openaire.eu/

[22] OpenAIRE: https://openaire.eu

[23] OpenCDMP: https://github.com/OpenCDMP/OpenCDMP

[24] OpenCAYLE: https://www.scayle.es/manual/en/opencayle

[25] PostgreSQL: https://www.postgresql.org/

[26] Publications Office of the European Union and Jessop, P., Data citation – A guide to best practice, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2830/59387

[27] SCAYLE: https://www.scayle.es/

[28] Software Deposit: What to deposit, M. Jackson - Jisc [DOI: 10.5281/zenodo.1327325] [https://doi.org/10.5281/zenodo.1327325]

[29] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3.1 (2016): 1-9.

[30] Zenodo: https://zenodo.org/

[31] MLCommons, Croissant: A Metadata Format for ML-Ready Datasets, 2024, https://docs.mlcommons.org/croissant/docs/croissant-spec.html