# DataGEMS

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




Report

# D7.1
# Pilot Requirements Report

PU

13/10/2025

## Disclaimer

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

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

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

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



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

## Document Information

| | |
|---|---|
| **Type** | Report |
| **Identifier** | D7.1 |
| **Title** | Pilot Requirements Report |
| **Dissemination level** | PU |
| **Due Date** | 30/09/2025 |
| **Actual Delivery Date** | 13/10/2025 |
| **Lead Beneficiary** | NOA |
| **Other beneficiaries** | IPB, DHBW, IDS, ARC, CNRS |
| **Editor** | Vassiliki Kotroni |
| **Authors** | Vassiliki Kotroni (NOA), Antonis Bezes (NOA), Vassilis Vafeiadis (NOA), Florbela Fernandes (IPB), Ana I. Pereira (IPB), Gabriel A. Leite (IPB), Friedemann Schwenkreis (DHBW), Thora Hagen (IDS), Georgia Koutrika (ARC), Sihem Amer Yahia (CNRS), Sarab Boussouar (CNRS) |
| **Contributors** | |
| **Work Packages** | WP7 |
| **Keywords** | Use cases, user communities, datasets and products, user stories, requirements mapping, tasks |
| **Version** | 1.0 |
| **Version information** | Number of pages: 37<br>Last file save: 13/10/2025<br>File revision: 66<br>File name: D7.1_Pilot_requirement_report.docx |


### Copyright Information

License: Creative Commons Attribution (CC-4.0-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 | 13/10/2025 | Submitted to EU | Vassiliki Kotroni |
| | | | |
| | | | |

## Document Tables

### Table of Tables

Table 1. Description of End-User Communities for HIGHER EDUCATION Use Case ..... 12
Table 2 Description of Datasets for HIGHER EDUCATION Use Case. ..... 12
Table 3. Description of User Stories for HIGHER EDUCATION Use Case. ..... 14
Table 4. Description of Requirements for HIGHER EDUCATION Use Case. ..... 15
Table 5. Mapping of Tasks to Requirements for HIGHER EDUCATION Use Case. ..... 15
Table 6. Mapping of Tasks to Scenarios for HIGHER EDUCATION Use Case. ..... 16
Table 7. Description of End-User Communities for LIFELONG LEARNING Use Case ..... 17
Table 8. Description of Datasets for LIFELONG LEARNING Use Case. ..... 18
Table 9. Description of User Stories for LIFELONG LEARNING Use Case. ..... 18
Table 10. Description of Requirements for LIFELONG LEARNING Use Case. ..... 19
Table 11. Mapping of Tasks to Requirements for LIFELONG LEARNING Use Case. ..... 20
Table 12. Mapping of Tasks to Scenarios for LIFELONG LEARNING Use Case. ..... 20
Table 13. Description of End-User Communities for WEATHER Use Case. ..... 21
Table 14. Description of Datasets for WEATHER Use Case. ..... 22
Table 15. Description of User Stories for WEATHER Use Case. ..... 25
Table 16. Description of Requirements for WEATHER Use Case. ..... 25
Table 17. Mapping of Tasks to Requirements for WEATHER Use Case. ..... 27
Table 18. Mapping of Tasks to scenarios for WEATHER Use Case. ..... 28
Table 19. Description of End-User Communities for LANGUAGE DATA INFRASTRUCTURE Use Case. ..... 29
Table 20. Description of Datasets for LANGUAGE DATA INFRASTRUCTURE Use Case. ..... 30
Table 21. Description of User Stories for LANGUAGE DATA INFRASTRUCTURE Use Case. ..... 34
Table 22. Description of Requirements for LANGUAGE DATA INFRASTRUCTURE Use Case. ..... 34
Table 23. Mapping of Tasks to Requirements for LANGUAGE DATA INFRASTRUCTURE Use Case. ..... 36
Table 24. Mapping of Tasks to Scenarios for LANGUAGE DATA INFRASTRUCTURE Use Case. ..... 36

### Table of Figures

Figure 1. Scheme of the interconnection among WPs in the frame of DataGEMS. ..... 9

## Table of Contents

**1 General Overview ..... 8**

1.1 Introduction ..... 8

1.2 Context and Document Structure ..... 9

**2 User Requirements ..... 11**

2.1 HIGHER EDUCATION (MathE) Use Case (UC1) ..... 11
2.1.1 End-User Communities ..... 11
2.1.2 Datasets and Products ..... 12
2.1.3 Use Case Scenarios ..... 13
2.1.4 User Requirements ..... 13
2.1.4.1 User Stories ..... 13
2.1.4.2 Requirements Priority Table ..... 14
2.1.5 Tasks ..... 15

2.2 LIFELONG LEARNING Use Case (UC2) ..... 17
2.2.1 End-User Communities ..... 17
2.2.2 Datasets and Products ..... 17
2.2.3 Use case Scenarios ..... 18
2.2.4 User Requirements ..... 18
2.2.4.1 User Stories ..... 18
2.2.4.2 Requirements Priority Table ..... 19
2.2.5 Tasks ..... 19

2.3 WEATHER Use Case (UC3) ..... 21
2.3.1 End-User Communities ..... 21
2.3.2 Datasets and Products ..... 21
2.3.3 Use Case Scenarios ..... 23
2.3.4 User Requirements ..... 24
2.3.4.1 User Stories ..... 24
2.3.4.2 Requirements Priority Table ..... 25
2.3.5 Tasks ..... 26

2.4 LANGUAGE DATA INFRASTRUCTURE Use Case (UC4) ..... 29
2.4.1 End-User Communities ..... 29
2.4.2 Datasets and Products ..... 30
2.4.3 Use Case Scenarios ..... 31
2.4.4 User Requirements ..... 33
2.4.4.1 User Stories ..... 33
2.4.4.2 Requirements Priority Table ..... 34
2.4.5 Tasks ..... 35

**3 Conclusion ..... 37**

## Abstract

This deliverable (D7.1) defines the initial set of pilot-specific requirements that will guide the co-creation, implementation, and evaluation of the DataGEMS platform under Work Package 7 (WP7): *Pilots and Co-Creation Activities*. WP7 aims to validate the platform's functionalities through real-world pilot deployments across four diverse use cases: **Higher Education (UC1), Lifelong Learning (UC2), Weather (UC3) and Language Data Infrastructure (UC4)**, involving multiple domains and user communities.

For each use case, this report outlines:

- The target user communities and their specific needs,
- The datasets and products required for use case scenarios,
- A description of use case scenarios
- A structured set of user requirements, articulated as user stories,
- And a task-based framework that translates these requirements into actionable development goals.

A key component of this deliverable is the mapping from user requirements to specific tasks, which ensures traceability between user needs and the system components to be developed and tested in later phases (D7.2–D7.5). Tasks act as an intermediate abstraction that bridges human-centered expectations and technical implementation, facilitating coordination with platform development (WP2–WP6) and laying the foundation for meaningful evaluation metrics.

By capturing the practical, domain-specific expectations of each use case, this deliverable ensures that the DataGEMS platform remains grounded in real-world contexts, supports explainable and FAIR data practices, and is adaptable across heterogeneous data environments. The outcomes of this report directly inform the design specifications (D7.2) and guide the iterative use case scenario and then pilot deployment strategy throughout the project's lifecycle.

# 1 General Overview

## 1.1 Introduction

The DataGEMS system is designed as a flexible, intelligent platform for data discovery and exploration across heterogeneous domains. As the project evolves, it introduces a range of use case scenarios, shaped by differences in data types, user expertise, and technological needs. This diversity makes evaluation a core concern: beyond verifying technical functionality, it is essential to assess whether the system delivers meaningful, usable, and explainable support for different user groups and domains.

To structure this evaluation, the project defines a set of use cases each comprising multiple scenarios involving concrete user communities, tasks, and datasets. These pilots are designed to test how well the DataGEMS components perform under diverse operational conditions.

The four selected use cases are:

- UC1 – HIGHER EDUCATION (MathE),
- UC2 – LIFELONG LEARNING,
- UC3 – WEATHER,
- UC4 – LANGUAGE DATA INFRASTRUCTURE.

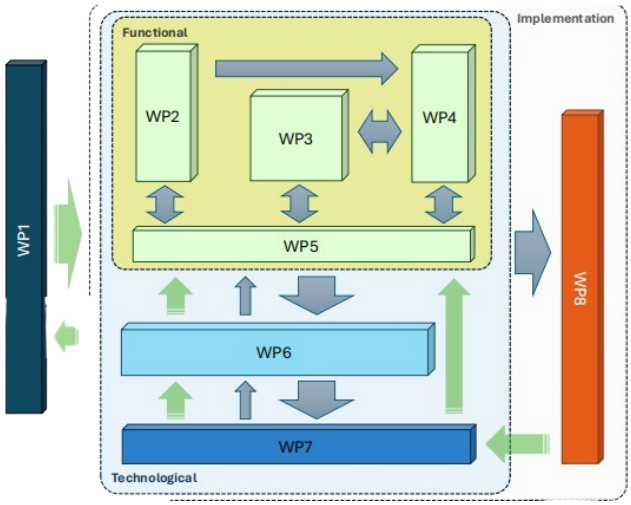
A total of **20** pilots (5 per use case) are planned within DataGEMS. Each pilot will emerge through collaboration between use case stakeholders and component developers, starting from simple examples and increasing in complexity as the project progresses. These pilots are intended to guide and structure the evaluation process by providing relevant usage analytics, quality monitoring opportunities, and direct feedback from engaged communities.

WP7 leads the coordination, implementation, and assessment of these pilots. Its goals include:

- Structuring domain-specific evaluation through representative tasks,
- Capturing performance and usability metrics,
- Feeding back into platform development and integration (WP2–WP6).

This deliverable, D7.1 – Pilot Requirements Report, represents the first step in this process. It identifies the end-user needs, target datasets, expected services, and key development tasks for each use case. These requirements will inform system design (D7.2), component integration (D7.4), and iterative testing and validation (D7.3, D7.5).

By grounding the DataGEMS platform in diverse and realistic usage scenarios, WP7 ensures the platform evolves in a user-driven, FAIR-aligned, and broadly impactful manner.




WP7 connects upstream to WP2, WP3, WP4 and WP5, taking input from data profiling, discovery, and exploration components, and downstream to WP6, guiding the integration and final evaluation of the platform. It also provides feedback for requirement refinement.

*Figure 1. Scheme of the interconnection among WPs in the frame of DataGEMS.*

## 1.2 Context and Document Structure

This report constitutes **Deliverable D7.1 – Pilot Requirements Report** of the DataGEMS project, prepared as part of **Work Package 7 (WP7) – Pilots and Co-Creation Activities**. WP7 plays a pivotal role in evaluating the DataGEMS platform in diverse, real-world settings through the implementation of domain-specific pilots.

D7.1 is the first of five deliverables within WP7 and lays the foundation for a progressive design and evaluation workflow that includes:

- D7.2: Pilot Specifications and Design Report,
- D7.3: Usage Analytics and Quality Monitoring Report,
- D7.4: Pilot Software Components,
- D7.5: Pilot Implementation Evaluation Report.

This report captures the initial functional, non-functional, and data requirements for each of the four DataGEMS use cases:

- UC1 – HIGHER EDUCATION (MathE),
- UC2 – LIFELONG LEARNING,
- UC3 – WEATHER,
- UC4 – LANGUAGE DATA INFRASTRUCTURE.

Each use case section includes:

- A profile of the target user communities and their goals,
- An overview of the datasets and products to be used,
- A description of the use case scenarios
- A structured list of user stories and requirement priorities.

To support implementation planning and ensure traceability, under each Use Case a relevant subsection introduces a task-based abstraction layer, mapping user requirements to actionable development tasks. These tasks serve as a bridge between domain-specific needs and the technical specifications and components to be developed in WP2–WP6 and later evaluated within WP7.

The deliverable is intended to serve both as a reference for platform developers and as a starting point for co-creation with pilot stakeholders. It will evolve iteratively, with refinements based on feedback gathered during pilot deployment and user engagement phases.

All pilot frontends will be initially developed in English, ensuring consistency and facilitating early-stage evaluation. During the implementation phase, the feasibility and necessity of supporting additional local languages will be assessed in collaboration with the pilot stakeholders. This decision will depend on technical constraints, user feedback, and domain-specific requirements identified throughout the ongoing co-creation and testing process.

# 2 User Requirements

This section establishes the user-centered foundation for each of the four DataGEMS use cases. For each use case, we outline: the target end-user communities, the relevant datasets and data services, the planned pilot implementations, and a structured list of user requirements, informed by early co-design activities and domain knowledge.

To ensure coherence and traceability between user needs and the technical development roadmap of the DataGEMS platform, this section introduces a structured mapping between user requirements and domain-specific tasks for each use case.

Tasks serve as an intermediate abstraction layer between the high-level user requirements and the platform features and components designed to meet those needs. Each task captures a recurring data-related activity or service expectation that reflects what needs to be done (the "what"), bridging the gap between user goals ("why") and technical implementation ("how").

This mapping supports:

- The transition from D7.1 (requirements) to D7.2 (specifications and design),
- The design and evaluation of pilots (D7.3–D7.5), and
- A modular, interpretable development process rooted in real-world usage scenarios.

The tables below present this mapping for each use case. Each user requirement is grounded in one or more concrete, reusable tasks, which in turn, guide the development and evaluation of platform features. This ensures that the DataGEMS infrastructure delivers relevant, replicable, and explainable pilot services across domains.

## 2.1 HIGHER EDUCATION (MathE) Use Case (UC1)

The HIGHER EDUCATION use case leverages the MathE platform to enhance teaching and learning mathematics in higher education level. It focuses on improving access to educational resources, strengthening assessment and feedback mechanisms, and supporting personalized learning experiences for students and educators. The user requirements for the HIGHER EDUCATION use case have been identified through structured engagement with the MathE community, including students, teachers, and platform administrators. This process drew on previous platform usage analytics, feedback from pilot participants, and expert insights from educational researchers. Discussions focused on enhancing learning resources, improving assessment tools, and supporting personalized feedback. The resulting requirements reflect real classroom needs and the operational context of higher education mathematics teaching.

### 2.1.1 End-User Communities

| End-user Community | Description |
|---|---|
| **Students** | • Benefit from a vast and diverse collection of exercises to study mathematics, as well as diversified pedagogical resources. |
| **Teachers** | • Help teachers generate multiple-choice questions (MCQuestions) |

| | |
|---|---|
| | about a certain math topic and their related pedagogical resources.<br>• Support the teachers with data insights on topic contents, for example, where the students have more difficulties, topics more used in MathE platform, and common errors, among others. |
| **MathE Administrators** | • Enrich the mathE databases with MCQuestions and pedagogical resources.<br>• Personalize the student's path of learning. |

*Table 1. Description of End-User Communities for HIGHER EDUCATION Use Case*

### 2.1.2 Datasets and Products

| Dataset | Language | Format | Access & Licensing | Size |
|---|---|---|---|---|
| **Dataset for Assessing Mathematics Learning in Higher Education** | English | CSV | CC0 1.0 | 1.1 MB |
| **MathE Database** | English | SQL | N/A | 10.7 MB |
| **Integration/Integration Techniques Questions MathE Database** | English | CSV and PDF | CC0 1.0 | 127 MB |
| **Pedagogic Resources MathE Database** | English | CSV and PDF | CC0 1.0 | 127 MB |

*Table 2 Description of Datasets for HIGHER EDUCATION Use Case.*

**Dataset for Assessing Mathematics Learning in Higher Education:** This dataset was extracted from the MathE platform. The database consists of 546 records, each representing a student's individual answer to a math question. The information recorded includes student ID, country of origin, question ID, type of answer (correct or incorrect), question level (classified as basic or advanced, according to the assessment of a collaborating teacher), mathematical topic (general category), specific subtopic, and keywords associated with the question. The data covers the period from February 2019 to December 2023.

**MathE Database:** The MathE database is a structured relational set derived from the MathE educational platform. This database integrates information distributed across several dimensions, including user profiles, student performance, question characteristics, teaching materials, and educational metadata, such as question topics and subtopics. Support materials are available in PDF format and cover a variety of content, such as theoretical explanations, sample solutions, and tips on the topics covered. The questions, in turn, are formatted in LaTeX style, ensuring clarity in mathematical presentation, and include additional information such as the level of difficulty assigned by experts and keywords that facilitate their categorization and search.

**Integration/Integration Techniques Questions MathE Database**: The dataset associated with the *Integration* topic and the *Integration Techniques* subtopic has 61 questions, the correct answer, and three incorrect answers with common errors made by students, keywords associated, and the related material. The database is organized as follows: Question ID, topic name; subtopic name; question; Correct answer, Incorrect answer 1, incorrect answer 2, keywords, Videos, Material. Also, in the same CSV file, on a different sheet, it has video information as follows: ID, link, and associated keywords. The teaching material for this subtopic is also provided in this dataset.

**Pedagogic Resources MathE Database**: This dataset is based on 410 PDF documents organized by keywords in a CSV file. This CSV describes documents that cover all the topics and subtopics in MathE platform, exploring the theoretical fundamental pillars, solved examples and others relevant information to each subtopic. This database is the result of contributions from the MathE network, and its content has been analyzed and validated by professionals in the field.

### 2.1.3 Use Case Scenarios

The MathE ecosystem involves three communities: students, teachers, and MathE administrators, and the following use case scenarios are perceived, structured as follows:

**Scenario 1: Student Interaction.** This scenario deploys and evaluates the same service: an interactive framework where students can ask real-time questions in the DataGEMS chatbot (embedded in MathE platform). Through natural language dialogue, students can ask real-time questions, answer multiple-choice quizzes, and explore generative pedagogical resources given by the Moma model. This approach allows us to evaluate the coherence, scalability, adaptability and users' satisfaction across students of different countries, with different backgrounds and knowledge.

**Scenario 2: Teacher Interaction.** This aims to make possible to analyze insights of MathE data users, as for example the global student´s performance can be analyzed, or by country, in order to identify trends on topic content. For example, students' difficulties with a given content, questions suggestions, material suggestions, to improve the teaching and learning processes.

**Scenario 3: Administrator Interaction.** This is designed for MathE administrators who are teachers themselves, it focuses on content quality control.

### 2.1.4 User Requirements

This subsection details the specific user requirements identified for the HIGHER EDUCATION use case, expressed as user stories, and prioritized according to their relevance and urgency. The requirements capture concrete functional needs, such as enhanced access to learning materials, improved assessment tools, and personalized feedback mechanisms, ensuring that the MathE platform developments directly address the priorities of its academic user base.

#### *2.1.4.1 User Stories*

To identify user requirements, a focus group consisting of 20 students and 5 teachers was conducted. Their feedback is captured in the questions presented in the following table.

| ID | User group | Description |
|---|---|---|
| **RQ1** | Students | As a user, I want to interact with the Chatbot DataGEMS through the MathE platform and during the quiz to have some tips about a given question. |
| **RQ2** | Students | As a user, I want the system to guide me with some hints along each question from one topic/subtopic quiz. |
| **RQ3** | Students | As a user, I want the generated questions for a given sub-topic to be adapted to me in order to have more engage-ment in my learning path. |
| **RQ4** | Students | As a user, during the conversation with the Chatbot Dat-aGEMS, I want newly generated pedagogical resources if the ones given by MathE are not enough to elucidate me. |
| **RQ5** | Teachers | As a user, I want to identify distinctive patterns by topic/subtopic/level/keywords/students' performance and countries (not limited to these ones). |
| **RQ6** | Teachers | As a user, I want to be able to request multiple-choice questions about a specific subtopic with a certain level and keywords. |
| **RQ7** | Teachers | As a user, I want to request some pedagogical resources (videos, images and pdfs) for the specific questions gener-ated in RQ6. |
| **RQ8** | Teachers | As a user, I want to identify what is the topic content where students have more difficulty and what kind of material, I should use to support them better. |
| **RQ9** | MathE administrators | For the MathE system, the MCQuestions should be gener-ated for the desired topic/subtopic, with the difficulty level associated, keywords, four answers, one correct and three incorrect, and the associated pedagogical materials, to be incorporated in the MathE database. |
| **RQ10** | MathE administrators | For the MathE system, the generated questions must con-sider the student's real-time history, in particular, adapting the difficulty level of the questions to each student during the quizzes. |

*Table 3. Description of User Stories for HIGHER EDUCATION Use Case.*

#### *2.1.4.2 Requirements Priority Table*

| ID | Priority | Value | Description / Justification |
|---|---|---|---|
| **RQ1** | High | High | Verified through user testing of queries yielding meaningful and accurate results. |
| **RQ2** | Medium | High | Confirmed via interface showing clearly the step-by-step process |
| **RQ3** | High | High | Evaluated through user feedback, performance, time in touch with the chat and number of interactions |
| **RQ4** | Medium | High | Evaluated through user feedback |
| **RQ5** | Low | Medium | Evaluated through user feedback on the usefulness of |

| | | | |
|---|---|---|---|
| | | | summaries/reports for problems/needs identification. |
| **RQ6** | High | High | Evaluated through user feedback about semantic clarity math formulation. |
| **RQ7** | Medium | High | Evaluated through user feedback about its quality. |
| **RQ8** | High | High | Evaluated through user feedback on the quality, time consuming and alignment to the subtopic. |
| **RQ9** | High | High | Validated through the latency time, suitability to the subtopic, and adaptability to the right level of difficulty for each student during the quiz. |
| **RQ10** | High | High | Verified through latency time between MathE and DataGEMS and user's feedback. |

*Table 4. Description of Requirements for HIGHER EDUCATION Use Case.*

During the project, it is expected to have access to a DataGEMS API that connects with the MathE platform.

### 2.1.5 Tasks

The tables below map the tasks in the MathE Use Case to the related requirements presented in the previous section, followed by a mapping of these tasks to the corresponding scenarios.

| Task ID | Task Name | Description | Related RQs |
|---|---|---|---|
| **UC1T1** | Chatbot-guided quiz support | Students interact with the chatbot during quizzes to receive hints and clarifications. | RQ1, RQ2, RQ10 |
| **UC1T2** | Student-level adaptive Quiz recommendation | Automatically recommend MCQs adapted to the student's performance history and target subtopics. | RQ3, RQ9, RQ10 |
| **UC1T3** | Pedagogical resource creation/recommendation | Provide additional pedagogical resources (videos, PDFs) when MathE resources are insufficient. | RQ4, RQ7 |
| **UC1T4** | Teacher dashboard and pattern analysis | Allow teachers to analyze performance data across topics, levels, and countries to detect learning needs. | RQ5 |
| **UC1T5** | Teacher-driven Quiz composition | Teachers can request custom MCQs for specific subtopics, difficulty levels, and keywords. | RQ6 |
| **UC1T6** | Metadata-rich MCQ generation | Extend the existing MCQs pool in the MathE database with full metadata (level, keywords). | RQ8, RQ9 |

*Table 5. Mapping of Tasks to Requirements for HIGHER EDUCATION Use Case.*

| Scenarios | Included Tasks | Justification |
|---|---|---|
| **Student Interaction** | UC1T1, UC1T2, UC1T3 | Students engage with the chatbot during quizzes (UC1T1), receive personalized MCQ recommendations based on performance (UC1T2), and are supported with extra pedagogical materials when needed (UC1T3). |
| **Teacher Interaction** | UC1T4, UC1T5 | Focuses on teachers using dashboards to analyze student performance and trends (UC1T4), and compose quizzes tailored to pedagogical needs (UC1T5). |
| **Administrator Validation** | UC1T6, UC1T3 | Targets the validation of automatically generated MCQs with full metadata (UC1T6) and checks the coherence and value of associated pedagogical resources (UC1T3). |

*Table 6. Mapping of Tasks to Scenarios for HIGHER EDUCATION Use Case.*

## 2.2 LIFELONG LEARNING Use Case (UC2)

The LIFELONG LEARNING use case focuses on supporting workforce upskilling and reskilling through the integration of skills assessment and training opportunities. It aims to help both individuals and organizations identify skill gaps, match them with relevant learning resources, and track progress over time. The user requirements for this use case were identified through previous interactions and discussions with partner organizations of DHBW and close cooperation with stakeholders. The resulting requirements address practical needs for improving employability, organizational adaptability, and cross-language accessibility. A methodological verification is planned as part of the project.

### 2.2.1 End-User Communities

| End-user Community | Description |
|---|---|
| **Skill researchers** | Researchers looking into the changes of skill needs (historically and future oriented). |
| **Human Resource Executives** | Strategic decision makers of organizations regarding the structure of the workforce of an organization. |
| **Managers** | People of organizations who are responsible for decisions regarding the need for certain skills in a unit of the organization. Skills can be brought in by building the skills in the existing workforce or by hiring additional personnel with the skills needed. |
| **Workers** | People of a certain role or function who are interested in determining whether their skills are sufficient. |
| **Trainers** | People providing skill-building offerings for a certain function or role. |

*Table 7. Description of End-User Communities for LIFELONG LEARNING Use Case*

### 2.2.2 Datasets and Products

The specific datasets to be used is not yet determined for the LIFELONG LEARNING Use Case. However, five basic types of datasets can be identified as relevant within its context:

| Type | Description |
|---|---|
| **Competency Taxonomies** | Taxonomies that provide a defined means for the description of competencies, skills, qualifications and occupations as, for example, the ESCO taxonomy. |
| **Skill Models** | Models that describe the occupations, functions, skills and competencies of an organization, defined using management software like, for example, MuchSkills. |
| **Offerings** | Datasets describing an educational offering identifying the competencies which will be conveyed when attending the offering. |
| **Skill Demands** | Datasets that describe the anticipated development of skills in a certain context, as for example the “Skill survey” published by the CEDEFOP. |
| **Sector Taxonomies** | Taxonomies that define which skills are necessary in a certain organizational context. The skill forecast of the CEDEFOP is an example of such a dataset. |

*Table 8. Description of Datasets for LIFELONG LEARNING Use Case.*

### 2.2.3 Use case Scenarios

In the context of the LIFELONG LEARNING Use Case, we anticipate pilots to implement two major use case scenarios:

**Scenario 1 – Worker-Focused Skill Gap Analysis (English, German, French).** This case targets workers in three language groups and evaluates their ability to assess personal skill gaps using the DataGEMS platform. It will test the extraction of a personal skill model, comparison with target roles, and retrieval of relevant training opportunities. The evaluation will focus on usability, semantic accuracy, and multilingual support.

**Scenario 2 – Manager-Focused Organizational Skill Analysis (English, German, French).** This case targets managers responsible for workforce planning and reskilling. It evaluates the system's ability to process organizational unit skill models, to identify gaps against future needs, and to suggest training programs.

### 2.2.4 User Requirements

This subsection presents the detailed user requirements for the LIFELONG LEARNING use case, formulated as user stories and organized by priority. They capture essential capabilities such as personalized skill gap analysis, organizational-level skills reporting, targeted training recommendations, and multilingual access. These requirements provide a clear foundation for guiding platform development and ensuring that the implemented services deliver measurable value to both individual learners and organizational stakeholders.

#### *2.2.4.1 User Stories*

| ID | User Group | User Story |
|---|---|---|
| **RQ1** | Managers | Providing the purpose of my organizational unit and a skill model of it, I want to retrieve an analysis which skills are missing or outdated in my organization. |
| **RQ2** | Workers | Providing an identification of an occupation and a current skill model of myself, I want to retrieve an analysis of skills that are missing or skills that are outdated, |
| **RQ3** | Managers<br>Workers | Providing an analysis of skill gaps, I want to retrieve a list of educational offerings to close the gaps. |

*Table 9. Description of User Stories for LIFELONG LEARNING Use Case.*

#### 2.2.4.2 Requirements Priority Table

| Requirement | Priority | Value | Description / Justification |
|---|---|---|---|
| **RQ1: Organizational skill gap analysis** | High | High | Core functionality to support managers and organizations in analyzing workforce skill models, detecting gaps between current and future competencies, and recommending suitable training programs, while also validating the system's performance across different languages and organizational contexts. |
| **RQ2: Individual skill gap analysis** | High | High | Core functionality to enable individual workers to assess their personal skill profiles, identify gaps relative to desired roles, and retrieve relevant training opportunities through multilingual, user-friendly interaction with the DataGEMS platform |
| **RQ3: Match of educational offerings** | High | High | Core functionality to showcase how the DataGEMS platform supports both individuals and organizations in identifying skill gaps and matching them with relevant training opportunities |

*Table 10. Description of Requirements for LIFELONG LEARNING Use Case.*

The user requirements described in this section are based on the needs articulated by the target end-user communities and are directly aligned with the objectives of the LIFELONG LEARNING Use Case and its associated pilots. Given the iterative nature of the DataGEMS project, these requirements may be refined or extended during the implementation phase. Any updates will be made in close consultation with stakeholders and pilot participants to ensure that emerging needs are addressed and that the platform continues to provide contextually relevant support.

### 2.2.5 Tasks

The tables below map the tasks in the LIFELONG LEARNING Use Case to the user requirements presented in the previous section, followed by a mapping of these tasks to the corresponding scenarios.

| Task ID | Task Name | Description | Related RQs |
|---|---|---|---|
| **UC2T1** | Organizational Skill Gap identification | Analyze an organization's current skill set against future skill needs to identify gaps or outdated competencies. | RQ1 |
| **UC2T2** | Unit Mapping | Map organizational structures (e.g., departments, units) to skills and functions to contextualize gap analysis. | RQ1 |

| | | | |
|---|---|---|---|
| **UC2T3** | Individual Skill Gap Analysis | Compare a worker's current skill model to the expected skills for a given role to determine what is missing or outdated. | RQ2 |
| **UC2T4** | Training Offer Recommendation | Suggest relevant educational offerings that align with identified skill gaps for individuals or organizations. | RQ3 |
| **UC2T5** | Integrating Sector Taxonomies | Integrate standard taxonomies (e.g., ESCO, CEDEFOP) to enable semantic alignment between roles, skills, and learning offers. | All RQs |

*Table 11. Mapping of Tasks to Requirements for LIFELONG LEARNING Use Case.*

| Scenario | Task IDs | Justification |
|---|---|---|
| **Worker-focused** | UC2T3, UC2T4, UC2T5 | Focuses on individual skill gap analysis (UC2T3), recommending relevant educational offerings to close gaps (UC2T4), and semantic alignment with taxonomies (UC2T5). |
| **Manager-focused** | UC2T1, UC2T2, UC2T4, UC2T5 | Targets organizational skill gap identification (UC2T1), mapping units to skills (UC2T2), recommending trainings (UC2T4), and integrating sector taxonomies for consistency (UC2T5). |

*Table 12. Mapping of Tasks to Scenarios for LIFELONG LEARNING Use Case.*

## 2.3 WEATHER Use Case (UC3)

The WEATHER use case focuses on improving access to and analysis of meteorological and climate data for diverse user communities. It demonstrates how DataGEMS can integrate real-time and historical datasets, apply advanced analytics, and deliver results through intuitive, user-friendly services. The WEATHER use case requirements were defined in close collaboration with climate and data scientists from NOA's METEO unit, drawing on extensive experience in delivering operational weather and climate services. The process incorporated insights from interactions with civil protection authorities, urban planners, sectoral decision-makers, and educators to capture both real-time operational needs and long-term analytical requirements. This ensures that the identified requirements are grounded in real-world applications and diverse user contexts.

### 2.3.1 End-User Communities

The WEATHER Use Case serves a diverge range of end-user communities, each with distinct needs related to climate data access, interpretation, and application:

| End-user Community | Description |
|---|---|
| **Climate Scientists** | Use climate data and derived indices for regional analysis, trend assessment, and climate change research. |
| **Civil Protection Authorities** | Require real-time and historical climate metrics to evaluate risks (e.g., heatwaves) and trigger response protocols. |
| **Urban and Regional Planners** | Need spatially contextualized climate indicators to inform long-term resilience strategies for infrastructure and zoning. |
| **Sectoral Decision Makers** | From sectors such as tourism, energy, and construction who need climate insight for planning and operational purposes. These stakeholders focus on planning rather than emergency response. |
| **Educators and Journalists** | Translate complex climate data into understandable insights for communication, awareness raising, and public education. |

*Table 13. Description of End-User Communities for WEATHER Use Case.*

These communities face common barriers: lack of direct access to curated climate indicators, difficulty relating them to geospatial variables (population, land use), and limited technical capacity to process raw datasets or extract actionable information. The WEATHER Use Case aims to address these barriers by offering intuitive access, natural language querying, and geo-aware interpretation of weather/climate data products.

### 2.3.2 Datasets and Products

The following datasets have been identified as foundational for the WEATHER Use Case:

| Dataset/Product | Provider | Format | Access Type | Usage |
|---|---|---|---|---|
| **ERA5-Land Atmospheric Reanalysis (1950–Present)** | Copernicus / ECMWF | NetCDF / GRIB | Open (via service) | Long-term trend analysis; index derivation |
| **NOA Historical Weather Stations (53 sites)** | METEO unit @ NOA | CSV | Public | Daily/historical climate parameters |
| **NOA Real-Time Station Network (550 stations)** | METEO unit @ NOA | JSON | Private (via service) | Daily monitoring |
| **Digital Elevation Model (DEM, 30m)** | NASA | GeoTIFF | Open (via service) | Topographic context for impact analysis |
| **CORINE Land Cover 5.1.1** | Copernicus | GeoTIFF / Vectors | Open (via service) | Zonal statistics and affected area assessments |
| **EU Population Density Grid (1km resolution)** | European Environment Agency | GeoTIFF | Public | Human impact quantification |

*Table 14. Description of Datasets for WEATHER Use Case.*

These datasets enable the computation (among others) of WMO/ETCCDI climate indices (e.g., warm days, heavy precipitation days). The WMO/ETCCDI refers to the World Meteorological Organization (WMO) and its Expert Team on Climate Change Detection and Indices (ETCCDI). This expert team has defined a set of 27 internationally recognized climate indices, designed to characterize climate extremes in terms of temperature and precipitation. These indices include measures such as the number of consecutive dry days, heavy precipitation events, warm nights, and are widely used in climate change monitoring and reporting. Based on the computation of these indices, the developed service will support impact-aware analysis (e.g., overlays with population data) and provide the basis for smart data querying in an exploratory user interface.

A more detailed analysis of the datasets follows.

**ERA5-Land Atmospheric Reanalysis (1950–Present).** The ERA5-Land dataset is a high-resolution global reanalysis of land variables developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) under the Copernicus Climate Change Service. It provides hourly data starting from 1950, using consistent modeling techniques to represent past climate conditions. With a spatial resolution of 0.1° x 0.1°, the dataset includes key variables like soil moisture, temperature, and precipitation. It is widely used for climate change monitoring, model validation, and sectoral impact studies. This dataset is openly accessible through the Copernicus Climate Data Store (CDS) in NetCDF and GRIB formats. For the purposes of the project, we will

focus on dataset **4 Countries:** Greece, Switzerland, Germany, and Portugal. The selected variables for analysis are temperature, wind speed, precipitation, and solar radiation. These data will be stored in the project dataspace.

**NOA Historical Weather Stations (53 sites).** Operated by the METEO unit at the National Observatory of Athens (NOA), this dataset consists of historical observations from 53 meteorological stations across Greece. These stations represent long-term climatic monitoring sites, offering daily measurements of temperature, wind speed, humidity, and precipitation. The data, available in CSV format, supports climate trend analysis and extreme weather studies, particularly in the context of analysing climatic conditions in Greece. These data will be stored in the project dataspace or retrieved from the Climpact Project dataspace in real-time.

**NOA Real-Time Station Network (550 stations).** This high-frequency observational network comprises over 550 automated weather stations deployed throughout Greece and maintained by NOA's METEO team. These stations provide updated measurements on various meteorological parameters such as temperature, rainfall, wind speed, and atmospheric pressure. The data are stored in JSON format and made available via NOA's public portal (www.meteo.gr). Designed to support real-time applications such as civil protection, early warning systems, and public weather awareness, this dataset enables operational monitoring of weather conditions. These data are real-time updated, will not be stored (they are not open-data) but they will be accessible for the open-service queries.

**Digital Elevation Model (DEM, 30m).** The DEM used in the METEO use case is sourced from NASA's Shuttle Radar Topography Mission (SRTM). It provides elevation data at a spatial resolution of 30 meters, enabling accurate topographical representation across the globe. This data supports climate impact analysis by allowing researchers to link meteorological variables to terrain features such as slope, and elevation. The dataset is available in GeoTIFF format and can be accessed freely through NASA's Earthdata portal. For the needs of DataGems this dataset will be stored in the project dataspace.

**CORINE Land Cover 5.1.1.** The CORINE Land Cover (CLC) dataset is a pan-European land use and land cover mapping product coordinated by the European Environment Agency. Version 5.1.1 provides detailed classification (44 thematic categories) of natural and human-made surfaces across Europe. Available in both raster (GeoTIFF) and vector formats, CORINE data are crucial for spatial overlay analyses in climate studies. For example, overlays of precipitation indices with urban or agricultural zones help quantify potential climate impacts and inform adaptation planning.

**EU Population Density Grid (1km resolution).** Developed by the European Environment Agency (EEA), this dataset provides harmonized estimates of population distribution across Europe at a 1 km² grid resolution. It integrates data from national census offices and remote sensing sources such as LUCAS (Land Use/Cover Area frame Survey). Available in GeoTIFF format, this dataset is instrumental in assessing the human dimension of climate risks. When overlaid with extreme weather indicators, it supports the identification of vulnerable populations and enables socioeconomic impact estimation for urban planning and emergency preparedness.

### 2.3.3 Use Case Scenarios

In the frame of Use Case 3 "WEATHER", the following scenarios will be implemented in pilots:

**Scenario 1: Replication of the Climate Indicators Service:** This approach has been chosen for coherency and supports the context of evaluating the scalability, generalizability, and FAIRness of the same DataGEMS-enabled WEATHER service across diverse regional contexts. The core service is an interactive service for climate indices exploration with geospatial overlay. The evaluation goal is to measure accuracy, scalability, completeness, FAIRness and user satisfaction across varying geographies and sectoral needs.

**Scenario 2: Real-Time Weather conditions in Greece.** This scenario introduces a distinct yet complementary service, which focuses on **operational, real-time measurements**, titled: "Live Weather Watch: Real-Time Weather Data". This service will be designed for real time weather station data provided by NOA (as non-open dataset), and it will be designed as a modular service for future adoption in other countries. The key features of this pilot will be: (a) Integration of real-time meteorological data (e.g., from 550+ Greek stations via METEO.GR) (b) An interface for natural language queries such as "How much did it rain today in Thessaloniki?" (c) Options for short term forecast integration to complement real-time observations.

The evaluation focus can relate to the system characteristics (latency, reliability, spatial coverage), to the data (completeness and update frequency), or to human factors (ease-of-use, perceived value during live situations).

### 2.3.4 User Requirements

This subsection details the specific user requirements for the WEATHER use case, expressed as user stories and prioritized according to relevance and urgency. These requirements address functionalities such as natural language querying, climate index computation, geospatial overlays, anomaly detection, and export capabilities. Together, they form a clear roadmap for developing weather and climate services within DataGEMS that are both scientifically robust and operationally useful.

#### *2.3.4.1 User Stories*

| ID | User Group | User Story |
|---|---|---|
| **RQ1** | All | I want to query climate indices by time and region in natural language, so I can quickly assess recent extremes. |
| **RQ2** | Scientists, Authorities | I want to calculate climate indices (such as mean number of days with T>37C) over a region and visualise them on a downloadable map (such as geotiff), so I can assess long-term changes and support reporting. |
| **RQ3** | Authorities, Planners | I want to correlate climate indices with population density and land use, so I can assess risks and vulnerable areas. |
| **RQ4** | Stakehold-ers, Public Users | I want a simplified interface with predefined query templates, so I can access and visualize data without needing technical skills. |

| | | |
|---|---|---|
| **RQ5** | Civil Protection, Journalists | I want a visualization (map/plot) of custom threshold exceedances so I can communicate abnormal events to the public and media. |
| **RQ6** | All | I want downloadable results in common formats (eg. CSV), as well as easy to understand plots and maps (eg. Geotiff), so I can reuse and share them. |
| **RQ7** | All | I want to monitor real-time weather conditions so I can initiate timely response actions. |
| **RQ8** | Civil Protection, All | I want to ask real-time questions like "How much did it rain in Thessaloniki today?" and get summaries, so I can assess risks and inform the public. |

*Table 15. Description of User Stories for WEATHER Use Case.*

#### *2.3.4.2 Requirements Priority Table*

| Requirement | Priority | Value | Description / Justification |
|---|---|---|---|
| **RQ1: Natural language query support** | High | High | Simplifies access for non-technical users; boosts usability and engagement |
| **RQ2: ETCCDI index computation** | High | High | Enables standardized scientific reporting; based on established climate extremes methodology |
| **RQ3: Geospatial data integration (land/population)** | Medium | High | Supports decision-making with contextual insights |
| **RQ4: Predefined query templates** | Medium | High | Crucial for stakeholder usability across economic sectors |
| **RQ5: Visual analytics and anomaly summarization** | Medium | Medium | Aids in awareness and communication during extreme events |
| **RQ6: Export in structured formats** | Medium | Medium | Facilitates reuse and transparency of the results |
| **RQ7: Real-time weather monitoring** | High | High | Supports emergency response with timely situational awareness. |
| **RQ8: Semantic real-time querying** | Medium | High | Provides intuitive access to live risk insights for non-technical users; aligns with DataGEMS' NLP goals. |

*Table 16. Description of Requirements for WEATHER Use Case.*

The user requirements outlined in this section are grounded in the needs articulated by the target end-user communities and are directly aligned with the objectives of the WEATHER Use Case and and the overall project design. However, given the iterative nature of the DataGEMS project, these requirements may be refined or extended during the implementation phase. Any such updates will be based on consultation with stakeholders and pilot participants to ensure that emerging needs are captured, and that the platform continues to provide context-relevant support.

### 2.3.5 Tasks

The following table maps the tasks defined for the WEATHER Use Case to the corresponding requirements outlined in the previous section. The outputs of these tasks will be delivered in different forms depending on user needs and contexts: results may be returned as natural language summaries, as structured datasets in widely used formats (e.g., CSV), or as visual products such as maps, or plots (e.g., GeoTIFF). This variety of presentation modes ensures that both expert and non-expert users can access, interpret, and reuse the information in ways that best support their decision-making and communication workflows. Finally, we provide a mapping of the tasks to the corresponding scenarios.

| Task ID | Task Name | Description | Related RQs |
|---|---|---|---|
| **UC3T1** | NL Querying Interpretation | Users provide queries on weather/climate conditions that should be interpreted to understand the data needed to explore in order to provide a response to the query. | RQ1, RQ3, RQ4 |
| **UC3T2** | Data Digestion | Based on the interpreted user queries, define and digest the data bases (weather/climate/, geospatial, etc.) which should be used to extract the output related to the user query. | RQ2, RQ3, RQ4, RQ5, RQ7, RQ8 |
| **UC3T3** | Data Processing | This task is important because it defines the transformation logic between raw inputs and meaningful outputs. Once the relevant datasets are selected based on the interpreted query (e.g., temperature time series, land use maps), this task performs the necessary computations to generate the outputs. This includes statistical aggregation, spatial analysis, and the derivation of climate indices (e.g., WMO/ETCCDI indicators) or thresholds for anomaly detection. | RQ2, RQ3, RQ5, RQ8 |
| **UC3T4** | Predefined Templates for Non Experts Data Access | Offer query presets for simplified access to key statistics. for non-experts to access climate data. | RQ4 |

| | | | |
|---|---|---|---|
| **UC3T5** | Result Export & Reusability | Enable users to download outputs (indices, overlays, summaries) in structured formats like CSV, GeoJSON, NetCDF. Provide also possibilities for Data Visualization | RQ5, RQ6 |
| **UC3T6** | Restricted Access Data use for open service development | In the frame of this use case, there is a challenge to provide a response by querying also restricted access data sets (real-time NOA stations real time data). Development of a service is needed, with the limitation that external users will not have access to these data. The data will be ingested, interpreted and evaluated to extract semantic information for the current conditions. | |

*Table 17. Mapping of Tasks to Requirements for WEATHER Use Case.*

| Scenarios | Task ID | Justification |
|---|---|---|
| **Climate Indicators** | UC3T1 | Enables natural language querying of historical weather/climate data, improving accessibility across diverse user types in all replication countries. |
| | UC3T2 | Define and digest the data bases (weather/climate/, geospatial, etc.) which should be used to extract the output related to the user query |
| | UC3T3 | this task performs the necessary computations to generate the outputs. This includes statistical aggregation, spatial analysis, and the derivation of climate indices (e.g., WMO/ETCCDI indicators) or thresholds for anomaly detection. Overlays climate indices with geospatial data (e.g., land use, population), supporting vulnerability assessment across regions. |
| | UC3T4 | Offers template-based queries for non-expert users, critical for usability and adoption in diverse contexts. |
| | UC3T5 | Exports and visualizes climate information, helping users to present and reuse the data extracted. Facilitates export of results in FAIR formats (CSV, GeoJSON, NetCDF), supporting data reuse and interoperability evaluation. |
| **Real-Time Early Warning Dashboard for GR** | UC3T1 | Extends natural language querying to real-time weather context (e.g., “What is the current max wind?”). |
| | UC3T3 | Enables semantic linking of weather conditions to land use and population zones to assess immediate local impact. |

| | | |
|---|---|---|
| | UC3T4 | Enhance non-expert access for real-time contexts |
| | UC3T5 | Summarizes and visualizes real-time weather data boosting awareness and decision-making. |
| | UC3T6 | Core component of the real-time pilot; provides live weather monitoring and threshold-based alerts via NOA's sensor network. Ensures that the system can access the restricted data with an API while at the same time external users will not have access to these data |

*Table 18. Mapping of Tasks to scenarios for WEATHER Use Case.*

## 2.4 LANGUAGE DATA INFRASTRUCTURE Use Case (UC4)

The Language Use Case in DataGEMS explores how advanced language technologies can support the comparison and interpretation of large-scale text corpora. It is important to note that the scope of the Language Use Case has evolved from the original plan described in the project proposal. While the initial focus was on contemporary, spoken, multilingual language data, particularly parliamentary speech corpora, the current direction centers on historic encyclopedic texts. This adjustment shifts the emphasis towards diachronic linguistic analysis and a broader topical and cultural spectrum, while still preserving the multilingual dimension. Despite the absence of spoken language, the encyclopedic corpora offer a topically rich and, more importantly, semi-structured data source for investigating comparative semantics and conceptual history, thus serving as a representative foundation for the Language Use Case in DataGEMS.

### 2.4.1 End-User Communities

| End-User Community | Description |
|---|---|
| **Computational Linguists** | Benefit from a vast and diverse collection of textual data to study language phenomena, e.g. diachronic semantic shift, or multilingual concept developments. |
| **Digital Humanists** | Comparatively explore cultural discourse by grouping texts into thematic or temporal clusters and generating interpretative visualizations, alongside transparent analytic steps. Scholars can switch from visual overviews to the textual layer at any time to integrate quantitative with qualitative analysis. |
| **Lexicographers and Terminologists** | Query individual lexical items to obtain precise word-in-context examples and co-occurrence networks across domains or time periods. |
| **Educators and Students** | Browse and explore a diverse text collection and provide summaries and visualizations of complex material to aid with teaching and learning. |
| **Domain Experts** | Retrieve highly specialized subject knowledge (e.g., political history, philosophy) and answer multi-layered questions with evidence from the retrieved texts. |

*Table 19. Description of End-User Communities for LANGUAGE DATA INFRASTRUCTURE Use Case.*

Despite their disciplinary differences, these communities face a set of overlapping challenges when engaging with large textual datasets: identifying relevant material, navigating multilingual or domain-specific terminology, and translating complex research questions into operational workflows. Existing tools often lack the necessary flexibility, interpretability, or scalability to handle large corpora and nuanced vocabularies effectively. By supporting natural language queries, integrating interpretable metrics, and linking textual evidence directly with analytical outputs, the DataGEMS platform is designed to empower users to investigate texts on their own terms, while maintaining full transparency and reproducibility throughout the process.

### 2.4.2 Datasets and Products

| Dataset | Language | Format | Access & Licensing | Size |
|---|---|---|---|---|
| **EncycNet** | German | TEI XML, PNG (image set) | Zenodo (CC BY 4.0) | 757 MB text + 7 GB images |
| **Britannica** | English | XML, Plain Text | GitHub (CC BY 4.0) | 988 MB |
| **Diderot** | French | JSON, Plain Text | Wikisource, GitHub (CC BY 4.0) | 117 MB |
| **Wikipedia** | German, English, French | JSON, Plain Text | Multiple data dumps available (CC BY 4.0) | 2 GB |

*Table 20. Description of Datasets for LANGUAGE DATA INFRASTRUCTURE Use Case.*

**EncycNet:** EncycNet is a digital corpus of historical German encyclopedias, encompassing 22 works from the early 18th to early 20th century. Some of these are considered comprehensive standard works in the German speaking sphere, e.g. *Meyers Konversations-Lexikon* or *Brockhaus Conversations-Lexikon*, while other encyclopedias in the corpus cover specialized domain knowledge, such as philosophy or the railway system. In total the corpus contains on the order of 49.3 million word tokens. Each encyclopedia has been uniformly encoded in TEI-XML to capture their individual structures in a machine-readable way. The text is accompanied by a set of digitized images that are linked within the XML files. The dataset additionally includes a derived knowledge-graph version: as of version 0.3 it contains RDF/Turtle triples that link concepts across the encyclopedias and Wikidata Q-nodes. It also explicitly encodes lexical relations where possible using the Wikidata property schema. This linked data graph currently contains 4 of the comprehensive encyclopedias in EncycNet. EncycNet is released under a CC BY 4.0 open license and is available via Zenodo and GitHub. By providing cross-linked, diachronic encyclopedia content, EncycNet supports historical language research: for example, German vocabulary evolution and concept drift can be studied across the century. Its semantic encoding and links to Wikidata already support the DataGEMS vision of enabling FAIR data discovery and analysis.

**Britannica:** The "19th-Century Knowledge Project" led by Peter Logan aims to provide machine-readable editions of historical *Encyclopædia Britannica* volumes. These are digital, full-text transcriptions intended for computational analysis. Each entry of the printed encyclopedias has been extracted and encoded both as plain text and as TEI-XML, allowing structural elements (titles, paragraphs, citations, footnotes) to be machine-interpretable. The collection will span major 19th-century editions (3rd ed., 1788–97; 7th ed., 1830–42; 9th ed., 1875–89; 11th ed., 1910–11); however, as of now, only the 7th and the 9th editions are available. By recreating these seminal sources of knowledge in digital form, the project builds a large, open dataset of historical English-language content. Additionally, the TEI editions also include automatically generated metadata: index terms are stored in the header of each entry, enabling topic groupings and potentially linkage to other authority files. All releases are CC BY 4.0 licensed and are published on GitHub/Zenodo in segmented files. By covering several Britannica editions, this dataset also allows us to study the evolution of 19th century knowledge and language. The time span also aligns with EncycNet, making cross-cultural analyses possible.

**Diderot's Encyclopédie:** Diderot and D'Alembert's *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers* (1751-1772) is a foundational 18th century French encyclopedia,

comprising 17 text volumes and 11 plate volumes. It covers a wide range of knowledge from natural history to philosophy, theology, and social thought, providing valuable insights into Enlightenment-era perspectives. As a culturally distinct reference for historical encyclopedic knowledge, it serves as an important complement to the Language Use Case, particularly for diachronic and cross-cultural analyses. For the use case, the 17 text volumes were scraped from Wikisource, where it is freely available. It spans approximately 45.000 entries. In parallel, a dataset compiled by Pierre Nugues focuses on geography-related articles, of which about 10.000 entries have been manually linked to Wikidata. This linked subset, similar to the first two Language datasets, facilitates the integration of the *Encyclopédie* with contemporary knowledge. During the course of the project, both the full text corpus and the annotated subset can be consolidated to enrich conceptual and entity-level analyses in DataGEMS.

**Wikipedia:** Wikipedia is a large multilingual encyclopedic dataset and remains one of the most visited websites worldwide, which shows the continued relevance of encyclopedic content. Its wide availability has contributed to the decline of traditional encyclopedias; as an example, *Brockhaus* stopped publishing in 2014. This shift highlights how important this type of reference material has been and still is for many. For the DataGEMS Language Use Case, a selected subset of Wikipedia articles is used. This subset overlaps with entries found in some of the other encyclopedias included in the project. It was created using the MediaWiki API, and, where possible, each entry was collected in three languages: German, English, and French. These multilingual versions and historical equivalents are already linked to one another, which supports comparative analysis within the DataGEMS system. Additional Wikipedia articles and the Wikidata knowledge graph may also be used as the project develops. All Wikipedia content is freely available under the CC BY-SA 4.0 license.

Although the Grant Agreement made mention of parliamentary and legal corpora, Deliverable D7.1 strategically pivots to historical and encyclopedic texts due to their structured, factual nature, and broader accessibility and impact. These corpora offer a scalable foundation for downstream tasks in law and parliamentary domains, while also supporting wider adoption across research, education, and civic tech communities. This direction aligns with concrete user requirements and use case-specific needs, guiding both the pilots and corpus development. It strengthens the project's long-term impact but also ensures relevance and usability across diverse communities, including legal, civic, and academic stakeholders.

### 2.4.3 Use Case Scenarios

Each scenario aims to activate a distinct portion of the overall pipeline while grounding these operations in concrete analytical needs of different user communities.

**Scenario 1: Diachronic Change:** It targets expert users specifically with a linguistic background. It leverages the full pipeline (with a focus on final reasoning) to explore how a lexical item has changed over a specific time period (groupings by time are only supported via metadata). For example, the English word *awful* shifted from meaning “awe-inspiring” in the 18$^{th}$ century to its predominantly negative sense today. Another example would be *computer*, which moved from describing a female-dominated profession to denoting a machine, where the German *Rechner* (calculator) also showed a similar trend at the same time. The system must extract term usages at the retrieval stage and group contexts by meaning using extracted features. At the synthesis stage, the platform identifies whether and to what extent change has occurred using

representative examples from the groups, features given, and LLM reasoning. Evaluation focuses on the plausibility of grouped results (which may be broken down into the word-in-context task, where annotated data is already available) and transparency of reasoning.

**Scenario 2: In-Domain Analysis:** It is aimed for domain experts working in fields such as philosophy or political science who are primarily interested in thematic exploration of subject-specific content. It mainly tests how well the platform can retrieve thematically consistent material across datasets that fit the domain as well as the specific question at hand. This is also where the image-to-text conversion may be most relevant. Take, for example, for the domains of architecture elements such as arches or buttresses, or for biology detailed plant drawings or anatomical illustrations. Engineering often includes diagrams of machinery, to give another example. The main goal is to understand whether users can navigate and synthesize knowledge from their area of expertise. Evaluation criteria can include quality of sub-questions by the system, coherence of grouped content, image conversion quality, and output/summary plausibility.

**Scenario 3: Structured Feature Extraction:** It serves computational researchers who rely on DataGEMS for large-scale feature extraction for external analyses. It assesses the functionality and output of the pipeline in terms of the structured output generation (e.g. lexical distributions, co-occurrences, named entities, etc.). The features could be on the level of terms or whole texts. The aim is to validate whether the pipeline provides reproducible and exportable features that fit the user question and whether the outputs are compatible with standard statistical or machine learning workflows, for example a time-series analysis using scikit-learn or fine-tuning using the transformers library. For example, if a user simply asks: “How often does the word *empire* appear in English texts across the 19$^{th}$ and 20$^{th}$ centuries?”, the pipeline could output a series of normalized word frequencies, which could then be used in change-point detection or regression models to see when usage spikes or declines significantly. If the user is interested in semantic shifts like in pilot 1, it could be possible that they want use DataGEMS for training data creation, for example the sentences clustered by meaning to then fine-tune a BERT model on the word sense disambiguation task.

**Scenario 4: Educational Accessibility:** It specifically targets less technical users and highlights minimal user intervention. It is therefore aimed at educators and students who benefit from simplified access to complex materials. It mainly tests the extent to which the platform can create simplified summaries and visualizations that aid in understanding without manual tuning. The main objective is to evaluate whether the system lowers entry barriers while preserving the original meaning of the texts. Evaluation criteria can include ease of use, clarity of generated outputs, summary coverage, and translation quality/multilingual accessibility.

**Scenario 5: Human-in-the-Loop for Advanced Research:** It targets expert users who wish to intervene directly in the system pipeline at multiple points. It tests the extent to which the system supports customization and transparency and explores how researchers can iteratively refine results by adjusting intermediate components. As such, its unique contribution lies in the validation of the flexibility and modularity of the platform and serves as a stress test for programmability and responsiveness to intervention. It should measure how adjustments in one stage propagate through the rest of the pipeline and how the analytical results ultimately change. Evaluation criteria are the perceived granularity of control, usefulness of intermediate outputs, and reproducibility after manual intervention.

### 2.4.4 User Requirements

This subsection outlines the specific user requirements for the LANGUAGE DATA INFRASTRUCTURE use case. User requirements and user stories were gathered through collaborative discussions at IDS, with active participation from representatives of the targeted user communities to align the system design with their specific needs. These requirements focus on capabilities such as semantic search, advanced corpus exploration, multilingual text processing, and the integration of specialized linguistic resources. They are designed to support both expert researchers in computational linguistics and digital humanities, as well as non-technical users seeking intuitive access to complex language data.

#### *2.4.4.1 User Stories*

| ID | User Group | User Story |
|---|---|---|
| **RQ1** | All | As a user, I want to interact with the system using natural language queries, so that I can easily pose complex research questions. |
| **RQ2** | All | As a user, I want the system to transparently show each analytical step, so that I can understand, verify, and reproduce the analysis. |
| **RQ3** | All | As a user, I want to group texts by theme or time period, so that I can focus my analysis on coherent subsets of the data. |
| **RQ4** | All | As a user, I want to analyze texts in different languages simultaneously, so I can conduct multilingual research. |
| **RQ5** | Educators and Students, Domain Experts | As a user, I want to generate automatic summaries of long documents or the systems' analyses, so I can quickly grasp and share key points. |
| **RQ6** | Digital Humanists, Domain Experts | As a user, I want to upload images and extract their text, so I can include them in the text-based analysis. |
| **RQ7** | Educators and Students, Computational Linguists | As a user, I want to assess the complexity of texts, so that I can filter or compare documents based on linguistic difficulty. |
| **RQ8** | Computational Linguists, Lexicographers and Terminologists, Domain Experts | As a user, I want to compare how a concept changes over time, so that I can trace semantic or discursive developments across centuries or domain changes. |
| **RQ9** | Computational Linguists, Lexicographers and Terminologists, Digital Humanists | As a user, I want to identify distinctive words and concepts of text groupings, so that I can quickly grasp dominant terms and conceptual differences. |
| **RQ10** | Computational Linguists, Lexicographers and Terminologists | As a user, I want to examine co-occurrence patterns and lexical networks, so that I can explore semantic relationships in different text collections. |
| **RQ11** | Educators and Students, Digital Humanists, Domain Experts | As a user, I want to switch between textual and visual representations, so I can combine close and distant reading strategies. |
| **RQ12** | Computational Linguists, | As a user, I want fine-grained control over the analysis |

| | | |
|---|---|---|
| | Digital Humanists | pipeline, so I can adapt it to my research needs (e.g. adjusting intermediate questions, adjusting text groupings, customizing the MCP). |

*Table 21. Description of User Stories for LANGUAGE DATA INFRASTRUCTURE Use Case.*

#### 2.4.4.2 Requirements Priority Table

| ID | Priority | Value | Description / Justification |
|---|---|---|---|
| **RQ1** | High | High | Verified through user testing of natural language queries yielding meaningful and accurate results. |
| **RQ2** | High | High | Confirmed via interface walkthroughs showing clearly interpretable, stepwise analytical outputs. |
| **RQ3** | High | High | Validated by user ability to automatically organize text data by time or topic and perform focused comparative analysis. |
| **RQ4** | High | High | Acceptance demonstrated by multilingual search and analysis yielding consistent and interpretable outputs across languages. |
| **RQ5** | Medium | High | Evaluated through user feedback on the usefulness of summaries for comprehension and sharing. |
| **RQ6** | Medium | Medium | Acceptance confirmed via image uploads successfully converted to analyzable text and integrated into workflows. |
| **RQ7** | Medium | Medium | Confirmed when users can view and act on complexity scores to filter or compare texts. |
| **RQ8** | High | High | Verified through temporal analysis outputs reflecting semantic change and confirmed as useful by expert reviewers. |
| **RQ9** | Medium | Medium | Acceptance shown when users identify key terms and distinctions via word clouds and keyword extraction. |
| **RQ10** | Medium | Medium | Validated by researchers generating and interpreting co-occurrence patterns and lexical networks. |
| **RQ11** | Medium | High | Confirmed through seamless switching between textual and visual representations during exploratory analysis. |
| **RQ12** | Low | Medium | Accepted when advanced users successfully adjust the pipeline and outputs of individual steps for research needs. |

*Table 22. Description of Requirements for LANGUAGE DATA INFRASTRUCTURE Use Case.*

The user requirements described in this section are grounded in the needs articulated by the target end-user communities and are directly aligned with the objectives of the Language Use Case and the overall project design. However, given the iterative nature of the DataGEMS project, these requirements may be refined or extended during the implementation phase. Any such updates will be based on consultation with stakeholders and pilot participants to ensure that emerging needs are captured, and that the platform continues to provide context-relevant support.

### 2.4.5 Tasks

The first table below presents the mapping for the tasks defined for the Language Use Case to the corresponding requirements given in the previous section. The second table presents the mapping of these tasks to the relevant scenarios.

| Task ID | Task Name | Description | Related RQs |
|---|---|---|---|
| **UC4T1** | Knowledge Graph Integration | Link corpora to external knowledge graphs (e.g., Wikidata) and align semantic entities across datasets. The external information aids in retrieval and adds additional information for data discovery and analysis. | RQ4, RQ8 |
| **UC4T2** | Natural Language Querying | Enable users to submit research queries in natural language and receive meaningful, multi-step responses. The system generates query results in an accessible format, including textual excerpts and summaries. | RQ1 |
| **UC4T3** | Query Decomposition | Automatically decompose complex queries into structured sub-questions and reasoning chains. The user can investigate the generated queries. | RQ1, RQ2 |
| **UC4T4** | Steps Logging | Automatically log and display each step of the analysis to ensure transparency and reproducibility, allowing users to intervene. | RQ2, RQ12 |
| **UC4T5** | Semantic and Metadata-Based Retrieval | Retrieve relevant texts using full-text search, semantic embeddings, metadata, or linked data. | RQ3, RQ4 |
| **UC4T6** | Document Grouping and Classification | Automatically group texts by time, theme, domain, or user-defined criteria. Optional visualizations can include word clouds as well as key and distinctive terms per cluster. | RQ3 |
| **UC4T7** | Summary Generation | Automatically summarize documents and generate analytical outputs for quick understanding. Summaries are linked to underlying texts, allowing users to inspect the source documents. | RQ5 |
| **UC4T8** | Feature/Text Extraction | Extract text or other linguistic features like frequency, sentiment, complexity, distinctive terms and lexical networks. Includes generating text from images. Outputs include feature tables (e.g. sentiment per document, frequency per word etc.), co-occurrence matrices, embedding representations, and graphs of lexical networks. | RQ6, RQ7, RQ9, RQ10 |

| | | | |
|---|---|---|---|
| **UC4T9** | Concept Tracking | Analyze how concepts evolve over time within and across corpora. Outputs can include two-dimensional embedding plots, including cluster information. | RQ8 |

*Table 23. Mapping of Tasks to Requirements for LANGUAGE DATA INFRASTRUCTURE Use Case.*

| Scenario | Task IDs | Justification |
|---|---|---|
| **Diachronic Change** | UC4T1, UC4T5, UC4T6, UC4T9 | Focuses on tracking the evolution of concepts over time (UC4T9), retrieving semantically enriched texts (UC4T5), clustering them by time or theme (UC4T6), and aligning semantic entities via knowledge graphs (UC4T1). |
| **In-Domain Analysis** | UC4T2, UC4T3, UC4T5, UC4T6 | Supports domain-specific insights by enabling natural language queries (UC4T2), decomposing them into reasoning chains (UC4T3), and retrieving/grouping relevant documents (UC4T5, UC4T6). |
| **Structured Feature Extraction** | UC4T8, UC4T6, UC4T5 | Enables extraction of linguistic and statistical features (UC4T8), complemented by document grouping (UC4T6) and retrieval capabilities (UC4T5) to organize and analyze corpus data. |
| **Educational Accessibility** | UC4T2, UC4T7, UC4T5 | Targets less technical users by offering natural language querying (UC4T2), simplified access to summaries (UC4T7), and semantically aware document retrieval (UC4T5) with minimal user intervention. |
| **Human-in-the-Loop for Advanced Research** | UC4T4, UC4T9 | Tests expert-driven control by enabling step-by-step logging and changing (UC4T4) and concept tracking (UC4T9), allowing users to iteratively refine analyses. |

*Table 24. Mapping of Tasks to Scenarios for LANGUAGE DATA INFRASTRUCTURE Use Case.*

# 3 Conclusion

This deliverable, D7.1 – Pilot Requirements Report, has established the initial user-centered foundation for the four DataGEMS use cases: Higher Education, Lifelong Learning, Weather, and Language Data Infrastructure. For each use case, we identified target user communities, key datasets, and structured user requirements expressed through user stories and mapped to concrete tasks and use case scenarios. These mappings form a traceable bridge between high-level user needs and the technical components to be developed and tested in subsequent work packages.

The harmonized structure across use cases ensures comparability and facilitates cross-domain learning, while the explicit mapping from requirements to tasks provides clarity for the upcoming design (D7.2), implementation (D7.4), and evaluation (D7.3, D7.5) activities. This process guarantees that platform development remains grounded in real-world contexts and aligned with FAIR data principles.

As a living document, D7.1 will guide the co-creation and refinement of the use case scenarios and development of pilots throughout the project. Future updates and evaluations will incorporate feedback from user communities and technical teams, ensuring that the DataGEMS platform delivers relevant, explainable, and impactful services across domains.