DataGEMS

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



Report

D7.7

D7.5a Pilots Specifications & Design Report

PU

10/7/2026

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	D7.7
Title	D7.5a Pilots Specifications & Design Report
Dissemination level	PU
Due Date	30/06/2026
Actual Delivery Date	10/7/2026
Lead Beneficiary	CNRS
Other beneficiaries	IPB, DHBW, IDS, ARC, NOA, UNIVR, ZHAW, IG
Editor	Sarab Boussouar
Authors	Ana I. Pereira, Beatriz Flamia (IPB), Antonis Bezes, Vassiliki Kotroni (NOA), Thora Hagen (IDS), Stavroula Eleftheraki, George Katsogiannis (ARC)
Reviewer	Mike Xydas, Stavroula Eleftheraki, Georgia Koutrika (ARC)
Contributors	Kurt Stockinger
Work Packages	WP7
Keywords	Pilots, Evaluation, Implementation
Version	1.0
Version information	Number of pages: 41 Last file save: 10-Jul-26 4:07:00 PM File revision: 303 File name: D7.7_M18_D7.5a Pilot Implementation Evaluation Report

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	10/07/2026	Submitted to EU	Sarab Boussouar

Document Tables

Table of Tables

Table 1 Three-Dimensional Evaluation Framework 11
Table 2 Status Classification 11
Table 3 UC1P2 Summary 14
Table 4 UC1P2 Evaluation Results 17
Table 5 UC1P2 Success Criteria Assessment 18
Table 6 UC1P3 Summary 19
Table 7 UC1P3 Evaluation Results 21
Table 8 UC1P3 Success Criteria Assessment 22
Table 9 UC1P5 Summary 24
Table 10 UC1P5 Evaluation Results 27
Table 11 UC1P5 Success Criteria Assessment 29
Table 12 UC3P1 Summary 30
Table 13 UC3P1 Evaluation Results 34
Table 14 UC3P1 Success Criteria Assessment 35
Table 15 UC4P1 Summary 36
Table 16 UC4P1 Evaluation Results 38
Table 17 UC4P1 Success Criteria Assessment 39

Table of Contents

Document Tables 5
1 General Overview 9
1.1 Introduction 9
1.2 Context and Scope 10
1.3 Structure of this document 10
2 Evaluation Framework 11
2.1 Three-Dimensional Evaluation Framework 11
2.2 Status Classification 11
2.3 Data Collection Methods 12
3 UC1P2 Math MCQs Recommendation 13
3.1 Pilot Summary 13
3.2 Implementation Status 14
3.3 Evaluation Results 15
3.4 Success Criteria Assessment 17
3.5 Observations and Recommendations 18
4 UC1P3 Math Educational Material Recommendation 19
4.1 Pilot Summary 19
4.2 Implementation Status 19
4.3 Evaluation Results 20
4.4 Success Criteria Assessment 21
4.5 Observations and Recommendations 22
5 UC1P5 Math Administrator Validation for MathE Admins 24
5.1 Pilot Summary 24
5.2 Implementation Status 24
5.3 Evaluation Results 25
5.4 Success Criteria Assessment 28
5.5 Observations and Recommendations 29
6 UC3P1 Weather Data Query Service (WDQS) 30
6.1 Pilot Summary 30
6.2 Implementation Status 31
6.3 Evaluation Results 31
6.4 Success Criteria Assessment 34
6.5 Observations and Recommendations 35
7 UC4P1 Language Feature-based Corpus View 36

7.1 Pilot Summary 36
7.2 Implementation Status 37
7.3 Evaluation Results 37
7.4 Success Criteria Assessment 39
7.5 Observations and Recommendations 39
8 Conclusions 41

Abstract

This deliverable (D7.5a Pilot Implementation and Evaluation Report) presents the initial implementation evaluation results of the first pilots deployed within the WP7 pilot phase of the DataGEMS project. The report focuses on five pilots selected for early deployment due to their technical and operational maturity, active stakeholder involvement, and coverage of diverse domains.

The pilots include three from the Higher Education MathE use case — UC1P2 Math MCQs Recommendation, UC1P3 Math Educational Material Recommendation, and UC1P5 Math Administrator Validation for MathE Admins — as well as UC3P1 Weather Data Query Service from the Weather use case and UC4P1 Feature-based Corpus View from the Language use case. Together, they provide initial evidence on the performance, usability, and adaptability of DataGEMS services in realistic operational settings.

The deployment follows a phased evaluation strategy, with pilots introduced as they reach sufficient maturity and can engage relevant users. The Lifelong Learning use case is not yet included, as its user community is still being prepared and its recommendation-related capabilities are first being validated in the more mature MathE context. The findings from this first cycle will guide service refinement, evaluation improvements, and the planning of future pilot deployments.

The report presents the evaluation findings along three complementary dimensions: system performance, data quality, and user experience. For each pilot, it documents the implementation status, the evaluation results against the success criteria defined in D7.2, and a set of observations and recommendations to guide subsequent deployment cycles.

1 General Overview

1.1 Introduction

Pilot deployment and evaluation represent a key stage in the DataGEMS project, allowing the consortium to examine how the platform and its services perform in realistic operational settings. The WP7 pilot phase goes beyond technical validation: it explores the extent to which DataGEMS components address concrete user needs, support data-driven interactions, and generate actionable feedback for further improvement.

The deliverable presents the implementation status and initial evaluation results of five pilots that were selected for early deployment because they had reached a sufficient level of technical and operational maturity, involved active stakeholder communities, and represented a diverse range of application domains.

The pilots covered in this first evaluation cycle are the following:

- For the Higher Education MathE use case:
  - UC1P2 – Math MCQs Recommendation
  - UC1P3 – Math Educational Material Recommendation
  - UC1P5 – Math Administrator Validation for MathE Admins
- For the Weather use case:
  - UC3P1 – Weather Data Query Service
- For the Language use case:
  - UC4P1 – Feature-based Corpus View

The early deployment of these pilots follows a phased evaluation strategy, according to which pilots are introduced into the evaluation process once they are sufficiently mature and can engage relevant users or stakeholders. The selected pilots cover higher education, weather-related data, and language data infrastructure, thereby enabling an initial assessment of the adaptability of DataGEMS services across different domains and usage contexts. The evidence collected through this first deployment cycle will inform the refinement of the services, the evaluation methodology, and the planning of subsequent pilot activities.

The Lifelong Learning use case is not yet represented by a deployed pilot because its user community is still being prepared and because the relevant recommendation capabilities are first being validated in the more mature Higher Education MathE context. This staged approach allows the project to reuse early findings from MathE before extending and adapting the services to Lifelong Learning scenarios.

Future planning will build directly on the outcomes of the first WP7 deployment cycle. The pilots evaluated in D7.5a will continue to be monitored and refined, with further user feedback collected to assess their usability, effectiveness, and added value. At the same time, the lessons learned from these early deployments will inform the preparation of the next evaluation cycle, including the adaptation of recommendation capabilities from the Higher Education MathE context to the Lifelong Learning use case. As the Lifelong Learning user community is further developed and the relevant services mature, this use case will be introduced in a subsequent pilot deployment. In this way, the project ensures that future evaluation activities are evidence-

based, incremental, and aligned with the technical readiness and stakeholder availability of each use case.

1.2 Context and Scope

The Deliverable D7.5a is the first of two evaluation reports foreseen in the WP7 timeline. It focuses specifically on the initial deployment cycle of the five pilots under review and serves as an input to the consolidated evaluation report, D7.5b, which will cover the full set of WP7 pilots at M36.

The evaluation builds on the pilot descriptions, evaluation methodology, and success criteria defined in D7.2. In line with the three-dimensional framework established in that deliverable, the assessment considers system, data, and human evaluation aspects. For each pilot, the report examines the extent to which the relevant success criteria have been achieved, records deviations or unexpected findings, and identifies recommendations for the next deployment phase.

The evidence used in this evaluation combines multiple sources, including:

- System interaction logs generated by DataGEMS services during pilot sessions.
- Quantitative metrics computed automatically by pilot components.
- Structured user feedback collected through questionnaires and expert review sessions.
- Qualitative observations provided by pilot coordinators and use case partners.

Since the scope of this report is limited to UC1P2, UC1P3, UC1P5, UC3P1, and UC4P1, the observations and conclusions should be interpreted as preliminary findings. They will be revisited, complemented, and consolidated in D7.5b once additional pilots have been deployed and a more comprehensive body of evaluation data is available.

1.3 Structure of this document

The remainder of this document is structured as follows:

- Section 2 (Evaluation Framework) provides a brief recap of the evaluation framework defined in D7.2 and its application in this report.
- Sections 3-5 present the evaluation findings for the Higher Education MathE use case, covering pilot UC1P2 (Math MCQs Recommendation), pilot UC1P3 (Math Educational Material Recommendation), and pilot UC1P5 (Math Administrator Validation for MathE Admins).
- Section 6 presents the evaluation findings for pilot UC3P1 (Weather Data Query Service), corresponding to the Weather use case.
- Section 7 presents the evaluation findings for pilot UC4P1 (Feature-based Corpus View), corresponding to the Language use case.
- Section 8 concludes the document and outlines the next steps.

2 Evaluation Framework

2.1 Three-Dimensional Evaluation Framework

The evaluation of the DataGEMS pilots follows the three-dimensional framework established in D7.2, which is designed to capture both technical performance and real-world usability across all pilot contexts. The three dimensions are applied consistently across all pilots evaluated in this report.

Dimension	Description and Key Indicators
System Dimension	Technical performance and reliability of DataGEMS services. Key indicators include system response time, output stability across repeated executions, scalability with increasing data volumes.
Data Dimension	Correctness, completeness, and relevance of data processed by the system. Indicators include retrieval accuracy, consistency of computed outputs, correctness of derived statistics, and proper identification and integration of relevant datasets.
Human Dimension	User perspective and effectiveness of the system in supporting real tasks. Quantitative indicators include task completion rates, time-on-task, and interaction frequency. Qualitative indicators include perceived usefulness, interpretability of results, interface usability, and efficiency gains.

Table 1 Three-Dimensional Evaluation Framework

2.2 Status Classification

Each success criterion is assessed against one of four status labels, based on the evidence collected during the deployment cycle:

Status	Definition
Met	The criterion has been fulfilled based on the collected evidence.
Partially Met	The criterion has been partially fulfilled: some aspects require further work or data.
Pending	The criterion cannot yet be assessed: more deployment cycles or data are needed.
Not Met	The criterion has not been fulfilled in this evaluation cycle.

Table 2 Status Classification

2.3 Data Collection Methods

For this initial evaluation cycle, data collection relied on the following methods across pilots:

1. Automated system logs capturing query response times, execution success rates, and error frequencies.
2. Expert review sessions conducted by use case partners to assess output quality and correctness.
3. Comparative analysis of system outputs against reference datasets or independently computed results.
4. Questionnaires distributed to pilot participants and end users to collect qualitative feedback on system usefulness, user satisfaction, perceived ease of use, relevance of outputs, and overall experience with the platform.

The number of pilot users involved in this evaluation phase ranges from 5 to ~ 60 participants per pilot, reflecting the early stage of deployment. User samples include internal project partners and domain experts and selected external stakeholders. Statistical generalization is not claimed at this stage; the evaluation aims to identify key performance patterns and usability issues to guide the next deployment cycle.

3 UC1P2 Math MCQs Recommendation

3.1 Pilot Summary

UC1P2 focuses on personalized multiple-choice questions (MCQs) recommendation during mathematical tests. The system tailors recommendations to the student's experience by analyzing previous answers and performance history. The pilot evaluates whether the recommender can adaptively suggest useful questions that support learning progression, confidence building, and sustained engagement.

At this stage, two pilot evaluation iterations were conducted:

- 1st UC1P2 Iteration: done on February 27th, 2026 with 59 international students of mainly engineering courses.
- 2nd UC1P2 Iteration: done on June 5th, 2026 with 30 students from Polytechnic Institute of Bragança (IPB) under the Pre-Calculus unit course.

The two UC1P2 pilot evaluation iterations had complementary objectives and were not intended for direct comparison at this stage. The 1st iteration evaluated the DataGEMS recommendation system use the MathE MCQ database. The main objective of this iteration is to obtain data related to students behavior.

The 2nd iteration evaluated the MCQs generated and validated in UC1P5 (Administrator Validation for MathE Admins) using the actual MathE recommendation system. This iteration also establishes a baseline for evaluating the improved DataGEMS recommendation system in a future UC1P2 (Math MCQs Recommendation) pilot iteration with a larger group of students.

Field	Description
Pilot Code	UC1P2
Full Name	MCQs Recommendation
Use Case	UC1 Higher Education (MathE)
Target Users	Students who want to test a topic and subtopic on the MathE platform
DataGEMS Components	Query recommendation
Lead Partners	CNRS, IPB
Point of Contact	Sarab Boussouar (CNRS), Beatriz Flamia & Ana Pereira (IPB)
Implementation Period	M14- M18
Evaluation date	June 2026

Number of Pilot Participants	1st UC1P2 Iteration: 59 International students (including IPB students) 2nd UC1P2 Iteration: 30 IPB Students

Table 3 UC1P2 Summary

3.2 Implementation Status

1st UC1P2 Iteration:

As a first evaluation of the UC1P2 pilot, an online pilot activity named the **MathGEMS Datathon** was organized on February 27th, 2026 with the primary objective of collecting feedback data on the recommendation system and assessing its suitability for personalized mathematics learning. The event involved 59 international students, including students from IPB, Portugal and Greece. To facilitate rapid testing and data collection, the recommendation system was deployed through a dedicated Streamlit-based interface rather than being integrated directly into the MathE platform.

For this experiment, 50 questions from the topic Fundamentals of Mathematics (Algebraic Expressions and Elementary Geometry) were selected to compose the dataset used by the system. The activity began with a classification test of 12 questions to assess each student's initial mathematical proficiency. Students obtaining a score above 0.6 (i.e., answering at least half of the questions correctly) continued to the one-hour upskilling session, during which they answered a sequence of up to 50 questions. The questions were selected and ordered dynamically by the recommender, allowing the system to personalize the learning path according to each student's performance and estimated knowledge level.

At the end of the session, students were asked to complete a feedback questionnaire based on an adapted version of the **NASA Task Load Index (NASA-TLX)**. The purpose of the questionnaire was to collect user perceptions regarding both the effectiveness of the recommendations and the workload associated with the learning activity. Responses were recorded using a 5-point Likert scale across six dimensions:

- **Feeling of Accomplishment (Q1):** How much did the recommended tests help improve your understanding?
- **Effort (Q2):** What is your perceived effort in solving the recommended tests?
- **Mental Demand (Q3):** How mentally demanding was it to solve the tests?
- **Perceived controllability (Q4):** How well did the difficulty levels of the tests match your level?
- **Temporal Demand (Q5):** How time-pressured did you feel during the upskilling session?
- **Frustration (Q6):** How frustrated did you feel while taking the recommended tests?
- **Satisfaction (Q7):** How satisfied are you with the recommended tests?

The data collected during the MathGEMS Datathon provided the first user-centered evaluation of the UC1P2 recommendation system, offering valuable insights into its usability, perceived effectiveness, workload characteristics, and overall acceptance by learners.

2nd UC1P2 Iteration:

The second iteration of the UC1P2 pilot was conducted on June 5th, 2026, involving 30 Portuguese students enrolled in a Pre-Calculus course at IPB. Unlike the first pilot, which focused on evaluating the newly developed DataGEMS recommendation system, this experience was carried out on MathE-DataGEMS (mathedatagems.ipb.pt): a twin platform of the original MathE environment (mathe.ipb.pt). The main objective of this pilot was to test the

generated MCQs validated on UC1P5 (Administrator Validation for MathE Admins) with the students with the recommendation system of MathE, to establish a baseline for future comparisons. The results obtained from this iteration will be a reference point for assessing the improvements introduced by the new DataGEMS recommendation approaches in subsequent pilot evaluations.

During the activity, students completed two sets of recommended questions. The first set consisted of 10 questions from the topic Fundamentals of Mathematics (Algebraic Expressions), while the second set included 10 questions from the topic Differentiation (Derivatives). The questions were selected using the MathE recommendation system, allowing students to interact with the system under conditions representative of the current platform deployment.

The two pilot iterations provided valuable insights into both the technical performance of the recommendation approaches and the students' perceptions of their learning experience. The feedback collected during the **MathGEMS Datathon** and the subsequent evaluation on the MathE-DataGEMS platform highlighted several areas for improvement. Taking these observations and participant suggestions into consideration, the project team plans to conduct a third iteration of the UC1P2 pilot in July/September with students from the Polytechnic Institute of Bragança (IPB), and fourth and fifth iterations in France and Poland.

This upcoming evaluation will represent a more mature stage of the pilot, incorporating the lessons learned from the previous experiments and deploying the improved recommendation functionalities directly within the MathE-DataGEMS mirror platform.

3.3 Evaluation Results

- **1st UC1P2 Iteration:**

A total of 59 participants enrolled through the registration form of the datathon and were invited to interact with the DataGEMS adaptive recommendation system. Of these, 44 participants successfully registered on the testing platform, while 37 initiated a learning session by starting the pre-test phase, and 9 participants completed the entire workflow, including the pre-test, adaptive learning phase, and post-session survey. The remaining participants either abandoned the study at different stages or did not proceed beyond registration. Despite the relatively low completion rate, the collected data provide valuable insights into the system's technical performance, the effectiveness of the adaptive recommendation process, and participants' perceptions of usability and learning experience.

- **2nd UC1P2 Iteration:**

The second iteration evaluates the original MathE recommendation system under the same experimental conditions, serving as a baseline for future comparison with the adaptive recommendation approach.

Pilot version	Dimension	Measure	Target	Observation / Result
1st UC1P2 Iteration	**System**	Participation rate	Users successfully access and interact with the system	44 users registered, 37 initiated a session (84.1% participation rate).

		End-to-end completion rate	All users complete the full workflow	9 of 37 initiated sessions were completed end-to-end (24.3% completion rate).
		System latency	Acceptable response time	Mean system latency: 12.8 seconds.
	Data	Correct response rate	Higher correct answer scores over time	Mean learning phase score: 66.5%
		Mastery progression	Positive learning gain	Average mastery increment per learning step: +0.043.
	Human	Feeling of accomplishment	Positive learning experience (≥ 4/5)	Mean score 3.67/5, indicating a generally positive sense of accomplishment.
		Effort required	Moderate effort required (3-4/5)	Mean score 3.89/5, indicating a moderate level of effort.
		Mental demand	Moderate cognitive demand (3-4/5)	Mean score 3.89/5, indicating a moderate cognitive demand.
		Perceived controllability	Difficulty levels appropriately match learner proficiency (≥ 4/5)	Mean score 3.89/5, suggesting that the assigned difficulty generally matched students' proficiency.
		Temporal demand	Low to moderate time pressure (≤ 3/5)	Mean score 3.67/5, indicating moderate perceived time pressure.
		Frustration	Low frustration (≤ 2/5)	Mean score 2.67/5, indicating relatively **low** frustration.
		Satisfaction	High satisfaction in the adaptive	Mean score 3.67/5, indicating generally positive satisfaction

			recommendations (≥ 4/5)	with the adaptive recommendations.
2nd UC1P2 Iteration	System	Participation rate	Users successfully access and interact with the system	100% participation rate (30 logged in mathedatagems mirror).
		End-to-end completion rate	All users complete the full workflow	100% finished 4 math tests (each with 5 questions).
	Data	Correct response rate	Higher correct answer scores over time	58% correct answer for the derivatives topic. 72% of correct answers for fundamental mathematics topics.
		Mastery progression	Positive learning gain (1-5 level). All students begin in level 1.	In the derivative topic, the students were stable on level 3 (in average). In Fundamental Mathematics, the students were stable on 4 level (in average).
	Human	Trust in recommendations for official assessment	Majority ≥4/5	Mean score 3.73/5, 57% rated ≥4.
		Relevance of recommended MCQs	Positive & relevant recommendation (≥ 4/5)	80% of the students indicate more than or equal 4.

Table 4 UC1P2 Evaluation Results

3.4 Success Criteria Assessment

The success criteria assessment presented below refers exclusively to the first UC1P2 (Math MCQs Recommendation) pilot iteration, which evaluated the DataGEMS recommendation system using the existing MathE MCQ pool. The second UC1P2 pilot iteration had a different objective, namely the evaluation of the MCQs generated and validated in UC1P5 (Administrator Validation for MathE Admins) using the original MathE recommendation system. As such, its results are considered a baseline for future comparisons with the improved DataGEMS

recommender. For example, the second iteration achieved full student participation, providing a valuable reference for identifying and evaluating improvements that will be introduced and assessed in future UC1P2 iterations.

Success Criterion (from D7.2)	Status	Evidence
Adaptive question recommendations support learning progression.	**Met**	Average mastery increment per learning step: +0.043.
Adaptive question recommendations maintain student engagement	**Partially Met**	Completion rate: 24.3% (9/37 sessions completed). Abandonment was observed in 21 sessions.
The system functions reliably under real usage conditions.	**Met**	44 users registered, 37 initiated sessions, and the mean system latency was 12.8 s.
Recommendation behavior demonstrates adaptation to learner proficiency.	**Met**	Difficulty adaptation confirmed through action distribution and correct answer patterns.

Table 5 UC1P2 Success Criteria Assessment

3.5 Observations and Recommendations

Feedback from IPB students was valuable and has already led to improvements in the recommendation system.

The main issue identified was the number of MCQs proposed during the one-hour session. Students could receive up to 50 questions (5 questions per step, and it was a 10-step assessment), which is excessive for a testing phase and likely contributed to the observed abandonment rate. Future evaluations will reduce the number of questions.

Another issue was question repetition. The current recommendation logic may repeatedly suggest a question that a student answered incorrectly, which can become frustrating. Based on comparisons with the baseline recommender and student feedback, the recommender is being improved to take previous repetitions into account and avoid recommending the same incorrectly answered question more than twice.

4 UC1P3 Math Educational Material Recommendation

4.1 Pilot Summary

UC1P3 aims to evaluate recommendations that supports students using the MathE platform by suggesting relevant study materials and videos. When a student works on a math problem, the system identifies and recommends the most relevant resources available on the platform to help the student solve the problem at hand. The pilot evaluates whether the recommender can identify useful supporting materials from the MathE repository by combining curricular metadata with semantic similarity between questions and available materials.

Field	Description
Pilot Code	UC1P3
Full Name	Educational Material Recommendation
Use Case	UC1 Higher Education (MathE)
Target Users	Students using the platform for self-study. Teachers and lecturers curating materials for instruction.
DataGEMS Components	MathE material recommender
Lead Partners	ARC, IPB
Point of Contact	Antonis Mandamadiotis, Stavroula Eleftheraki (ARC), Beatriz Flamia & Ana Pereira (IPB)
Implementation Period	M17 - M18
Evaluation date	June 2026 (M18)
Number of Pilot Participants	10 higher education mathematics professors

Table 6 UC1P3 Summary

4.2 Implementation Status

By M18, UC1P3 has reached a functional implementation state.

For a given math question from the MathE platform's question catalog, the recommender first identifies the corresponding topic/subtopic pool and retrieves eligible teaching materials from it, in line with the recommendations scope defined by the use case providers. It then ranks these materials using a combination of metadata-based and semantic similarity signals between the question and the available materials.

The current production recommender is restricted to document-based teaching materials, i.e., PDF, DOCX, and PPTX resources. Video materials are not yet processed by this service and will be supported in a later system release.

4.3 Evaluation Results

The evaluation of UC1P3 includes two complementary parts: an offline evaluation of the recommendation service and a human/expert evaluation of the recommended materials.

Offline Evaluation

The offline evaluation was designed to test the recommender on realistic MathE learning situations, using questions that students had attempted in the platform. For each available topic-subtopic combination, one representative benchmark question was selected from the MathE assessment history. The selection prioritized questions with substantial student interaction and higher difficulty, using indicators such as number of distinct students, total attempts, and wrong-answer rate. In this way, the benchmark focuses on questions where students are more likely to need supporting educational material.

The evaluation reference sets were collected directly from the use case providers. For each benchmark question, the providers were asked to identify the document materials that were useful and related to the question content, and that could support a student in answering it. The document material IDs provided by the use case providers were then used as the ground truth for the evaluation.

The resulting ranked recommendation lists were evaluated using standard metrics that capture retrieval success and ordering quality. Hit@5 and Hit@10 were used to measure whether at least one provider-identified relevant document appeared within the first five or ten recommendations, respectively. This choice reflects the sparsity of the provider-defined reference sets, which contained approximately two relevant documents per question on average. Finally, nDCG@5 and nDCG@10 were used to assess ranking quality, rewarding cases where relevant documents appear higher in the recommendation list. In addition to these ranking metrics, we also measured the time needed by the system to fetch the recommendations, reported in seconds.

Online Evaluation

An online evaluation was conducted with an expert team of 10 mathematics teachers in higher education. Unlike the offline benchmark, this evaluation considered a broader set of MathE questions from three selected topic-subtopic pools: *Fundamentals of Mathematics (Algebraic Expressions, Equations, and Inequalities); Differentiation (Derivatives); and Integration (Triple Integrals)*. Across these pools, the experts reviewed 295 MathE questions and validated or corrected their associations with the available educational materials.

Dimension	Measure	Target	Observation / Result
System	Response time for generating recommendations	The recommender should return results within a time range acceptable for pilot use	The recommender achieved a mean response time of 1.13

			seconds and a median response time of 0.54 sec.
Data	Relevance of recommended materials	Recommended materials should match the provider-identified useful documents for each math question	Hit@5 = 0.824, Hit@10 = 1.000, nDCG@5 = 0.540, nDCG@10 = 0.635
Human	Usefulness for teaching support	Teachers should perceive recommendations as useful for supporting explanation, instruction, or material curation	Overall, there is a correct association with the educational material.
	Usefulness for learning support	Students should perceive recommendations as useful for self-study	Not yet formally evaluated with students in this cycle
	Transparency	Users should understand why the recommended materials were shown	Not yet formally evaluated with end users in this cycle

Table 7 UC1P3 Evaluation Results

4.4 Success Criteria Assessment

Success Criterion (from D7.2)	Status	Evidence
Relevant educational materials are successfully retrieved and ranked for representative math questions.	**Met**	Offline evaluation showed strong retrieval coverage: Hit@5 = 0.824 and Hit@10 = 1.000. Ranking quality was more moderate, with nDCG@5 = 0.540 and nDCG@10 = 0.635.
Teachers perceive the recommended materials as relevant and supportive of their learning needs.	**Met**	The share of correctly associated materials differed across the evaluated pools: ~70% for *Fundamentals of Mathematics – Algebraic Expressions, Equations, and Inequalities*; ~23% for *Differentiation – Derivatives*; and 100% for *Integration – Triple Integrals*.

Students perceive the recommended materials as relevant and supportive of their learning needs.	**Pending**	This criterion will be assessed in the next UC1P3 evaluation step through questionnaire feedback from MathE students.
Users perceive the recommendation process and results as transparent and understandable.	**Pending**	This criterion will be assessed in the next UC1P3 evaluation step through questionnaire items on whether MathE platform users understand the recommendation rationale.

Table 8 UC1P3 Success Criteria Assessment

4.5 Observations and Recommendations

The offline evaluation shows that the recommender was able to retrieve provider-validated materials for the evaluation questions. The recommender retrieved at least one useful document within the top 10 recommendations for every evaluable question (Hit@10 = 1.000), and within the top 5 recommendations for most evaluable questions (Hit@5 = 0.824). The ranking-based results were more moderate (nDCG@5 = 0.540; nDCG@10 = 0.635), indicating that relevant materials were generally retrieved, but not always placed among the highest-ranked recommendations. The response time was also suitable for pilot use, with recommendations fetched in 1.13 seconds on average, showing that the service can return results within a practical time frame for integration in the MathE platform.

These results suggest that the ordering of materials within each topic-subtopic pool can be further improved. Since all candidates already belong to the same pool as the question, the main challenge is to rank higher the materials that directly support the specific question, rather than those that are only broadly related to the topic. Future refinements may include enriching material metadata, improving keyword and concept extraction from questions and documents, testing math-specific embedding models, and adjusting the balance between metadata-based and semantic similarity signals.

The online evaluation results showed that the material recommendation system correctly associated more than half of the evaluated questions with appropriate educational materials. To validate the proposed approach, MathE expert professors evaluated a sample of questions from each of the three selected topic-subtopic pools. The sample size was defined to ensure an 80% confidence level and a 10% margin of error.

The results varied across the evaluated pools. For *Fundamentals of Mathematics – Algebraic Expressions, Equations, and Inequalities*, the MathE platform includes 22 learning materials and 122 questions. In this case, the three materials with the highest relevance scores were almost always correctly associated with the questions, while lower-ranked materials were less relevant and often considered inappropriate. Overall, approximately 70% of the evaluated associations were correct.

For *Differentiation – Derivatives*, only 9 learning materials are available for 85 questions. This limited material coverage affected the quality of the recommendations, as the available resources do not sufficiently cover the range of concepts addressed by the questions. As a result, only approximately 23% of the evaluated associations were considered correct.

For *Integration – Triple Integrals*, 73 learning materials are available for 88 questions. The broader availability of materials provided better content coverage, allowing the recommender to identify appropriate resources more effectively. For all evaluated questions in this pool, the recommended materials were considered appropriate by the experts.

Overall, considering this initial iteration, the system performs as expected. However, further improvements are necessary in the near future to enhance the quality of the recommendations and increase the system's effectiveness across different topics.

5 UC1P5 Math Administrator Validation for MathE Admins

5.1 Pilot Summary

UC1P5 focuses on supporting MathE administrators in validating educational content generated by DataGEMS, including multiple-choice questions (MCQs) and associated metadata such as keywords and difficulty levels. The pilot evaluates the quality and relevance of generated content before its integration into the MathE database and recommender system.

At this stage, validation is performed manually by MathE experts from IPB. Generated MCQs and metadata are exported in PDF, JSON, and CSV formats and reviewed over time by administrators. This evaluation process was preferred by MathE partners instead of an immediate interface-based validation workflow, as expert assessment requires detailed pedagogical review. A dedicated validation dashboard integrated into the MathE platform is planned for a later stage of deployment.

Field	Description
Pilot Code	UC1P5
Full Name	Administrator Validation for MathE Admins
Use Case	UC1 Higher Education (MathE)
Target Users	MathE administrators (Math Teachers) at IPB, Portugal
DataGEMS Components	Unstructured dataset profiling, MCQ Generation
Lead Partners	CNRS, ZHAW, IPB
Point of Contact	Sarab Boussouar (CNRS), Kurt Stockinger (ZHAW), Beatriz Flamia & Ana Pereira (IPB)
Implementation Period	M12 - M18
Evaluation date	M14 - M18
Number of Pilot Participants	4 MathE administrators + 6 Math Teachers from Higher education

Table 9 UC1P5 Summary

5.2 Implementation Status

By June 2026, UC1P5 has reached a functional deployment state. The MCQ generation service is operational and has been used in continuous testing cycles with MathE experts since January 2026. For this pilot, two MathE topics were considered: Fundamental Mathematics and Derivatives. The keyword validation functionality developed by ZHAW is integrated and operational as well in offline mode.

Key implementation milestones achieved in the M12-M18 period include:

- Deliver the MCQ output structure, including question statement, answer options, correct answer identification, and metadata (topic, subtopic, difficulty level, keywords, related resources).
- Associate keywords with the education material and question sentences.
- Validation cycles conducted by MathE Admin and teachers enrolled in the MathE platform.

The end-to-end administrator validation interface within the MathE platform for the topics Fundamental Mathematics and Derivatives were deployed at MathE-DataGEMS (mathedatagems.ipb.pt).

The generated MCQs, in a total of 873, were delivered directly to MathE expert teachers as PDF documents for offline manual review, as explicitly requested by IPB partners at this stage given the level of pedagogical examination required. This does not affect the validity or completeness of the evaluation results, as the core objective of this pilot (assessing the quality, correctness, and pedagogical relevance of generated MCQs and keywords) has been fully achieved through this manual review process. After that, the suggestions of the admin and the expert team were incorporated into the JSON file and deployed in MathE.

This pilot can therefore be considered fully deployed for its current validation scope. A subsequent pilot iteration is planned to evaluate the usability of the integrated validation dashboard and to incorporate the generation improvements identified during this first cycle.

5.3 Evaluation Results

The evaluation results presented below cover the three dimensions defined for this pilot (System, Data, and Human) and correspond to the core objective of UC1P5 (Math Administrator Validation for MathE Admins) as defined in D7.2: MathE expert teachers (admins) reviewing and validating the MCQs, difficulty levels, and keywords generated by DataGEMS before their integration into the MathE platform.

Dimension	Measure	Target	Observation / Result
System	MCQ generation coverage	All defined MathE topics	MCQs have been generated, validated and deployed for two topics: Fundamental Mathematics and Derivatives, in a total of 210 MCQs. MCQs have been generated for the other 14 topics, which are in the validation process. To June 2026, IPB teachers have validated **489** from a total of **873** questions generated among the topics Differentiation (Derivatives, Implicity and Chain Rule, Partial Differentiation), Fundamental Mathematics (Algebraic Expression and Geometry), Graph Theory, Integration Techniques (Definite Integration and Double integration), Linear

			Optimization, and Non-linear Optimization. Validation of the remaining topics is ongoing.
	Export formats availability	PDF, JSON, CSV	Confirmed: all three formats used for expert review and validation workflow.
	LaTeX formatting quality	Questions render correctly	Formatting errors (LaTeX notation issues) identified across all topics but described as **minor** and **easily correctable**.
	Keywords association to educational material and MCQs	CSV	Validation through CSV files by math expert team.
Data	MCQ acceptance rate	Majority of generated MCQs accepted	489 generated questions reviewed: approval rates by experts per topic: Derivatives 70/82 (85%), Algebraic Expressions: 109/122 (89%), Geometry 48/68 (71%), Graph Theory 54/70 (77%), Linear Optimization 31/45 (69%), Non-linear Optimization. 27/32 (84%).
	Repetitive question rate	Low redundancy	Most removed MCQs were excluded due to repetition, meaning substantial similarity in structure or underlying idea rather than exact duplication. This accounted for 12 removals in Derivatives, 12 in Algebraic Expressions, 19 in Geometry, 16 in Graph Theory, and 5-6 in Non-linear Optimization.
	Correct answer accuracy	High accuracy on correct answer generation	Near-perfect correct answer generation for most topics (e.g. 82/82 Derivatives, 121/122 Algebraic Expr., 67/68 Geometry, 69/70 Graph Theory, 32/32 Non-linear Opt.). Notable exception: Linear Optimization where only 9/45 correct answers are generated.
	Distractor quality	Plausible distractors that capture common errors	Expert review confirms distractors are **highly similar** to correct answers, requiring full problem resolution (e.g. Derivatives Q3, Q37. Algebraic Expr.

			Q1, Q44). Assessed positively as pedagogically challenging. No low-quality distractor complaints.
	Topic/Subtopic Alignment	Consistent alignment with MathE taxonomy	All generated MCQs are correctly assigned to the intended topic and subtopic.
	Keyword Association Accuracy	Correct association	49% correct association of keywords in the education material associated with the Derivatives topic. 78% correct association of keywords in the education material associated with the Algebraic Expression topic. 100% correct association of keywords in the education material associated with the Linear Optimization topic. 91% correct association of keywords in the education material associated with the Statistics topic.
Human	Expert alignment with LLM evaluations	Agreement between automated and human assessments	Expert validation generally confirmed LLM-reported correctness for most topics, though notable misalignment occurred for difficulty-level assignment (15-72% of questions reassigned per topic) and for Linear Optimization, where only 9 of 45 correct answers were confirmed despite high LLM-reported accuracy. For the newly generated topics (LLM solver accuracy ranging from 80-100%), expert validation is still pending, so alignment cannot yet be confirmed.
	Perceived usability of MCQs generation service	Positive overall satisfaction	Experts reported overall **satisfaction** with the generated MCQs and associated metadata. Validated MCQs for Fundamentals of Mathematics and Derivatives have already been integrated into the MathE database for future use.
	Perceived Satisfaction of Keyword Association	Positive overall association	Experts reported overall **satisfaction** with the association of keywords.

Table 10 UC1P5 Evaluation Results

In addition to the administrator validation foreseen in the original pilot design, the UC1P2 (Math MCQs Recommendation) testing session held on June 5th, 2026 with IPB students was used to understand students' perspectives related to the generated MCQs.

Besides collecting baseline feedback on the original MathE recommendation system, the session provided an opportunity to replace the MathE questions database with generated MCQs and receive students feedback related to their usefulness, correctness, clarity, and perceived difficulty levels. This complementary evaluation offers valuable insights into the quality of the generated content from an end-user perspective.

After completing the UC1P2 (Math MCQs Recommendation) tests, students filled in a satisfaction questionnaire combining closed and open-ended questions, designed to capture their perceptions of question quality, difficulty level, and overall satisfaction. Responses to the closed-ended questions were collected using a 5-point Likert scale (1 = strongly disagree/very poor, 5 = strongly agree/excellent).

Students' perceptions of the generated MCQs indicated an overall high level of quality across all assessed criteria. The relevance of the distractors received a mean score of 4.00/5.00, where 77% of learners assigned a score of 4 or higher, suggesting that the incorrect answer options were appropriate. The clarity of the answer choices was also evaluated positively, achieving a mean score of 3.87/5.00, with 67% of students rating this criterion as 4 or 5. The clarity of the question wording received a mean score of 3.67/5, where 60% of students assigned ratings of 4 or higher, indicating that the majority of questions were considered clearly formulated. Regarding the assigned difficulty levels, students moderately agreed that the assigned difficulty levels matched their own perception of question difficulty, with an average rating of 3.50/5.

5.4 Success Criteria Assessment

Success Criterion (from D7.2)	Status	Evidence
Generated MCQs meet pedagogical standards	**Met**	Overall acceptance rate 81% across validated topics (Derivatives 85%, Algebraic Expressions 89%, Geometry 71%, Graph Theory 77%, Linear Optimization 69%, Non-linear Optimization 84%).
Human and LLM evaluations align	**Partially Met**	Alignment is generally confirmed for correct-answer generation, but difficulty-level assignment required reassignment for 15-72% of questions per topic.
Administrators perceive reduction in authoring effort	**Met**	No quantitative measure of authoring-effort reduction is currently available. However, qualitative feedback from the expert validators indicates that reviewing and validating automatically generated MCQs requires significantly less time and effort than creating new questions from scratch, particularly when existing question banks have already exhausted

		many of the most obvious ideas. Experts generally agreed that the generated MCQs provide a useful starting point that reduces the cognitive workload associated with question authoring. Furthermore, validated MCQs (Fundamentals of Mathematics and Derivatives) have already been integrated into MathE.
Distractor quality meets plausibility threshold	**Met**	Student questionnaire (n=30): mean 4.00/5 (77% rated ≥4) for distractor relevance/plausibility. Expert review independently confirmed distractors closely mirror correct answers, assessed as pedagogically challenging with no quality complaints.
Keyword relevance meets threshold	**Met**	Keyword extraction threshold successfully passed for tested datasets. With domain-specific fine tuning, even higher accuracies could be achieved in the future.

Table 11 UC1P5 Success Criteria Assessment

5.5 Observations and Recommendations

Several concrete areas for improvement have already been identified and are being incorporated into the next stages of the generation pipeline.

Difficulty-level alignment: Prompting techniques are being refined including calibration examples per difficulty level and requiring the model to justify its difficulty assignment based on reasoning steps to better align generated difficulty with expert and student perception.

Reducing repetition: Generation prompts will be adjusted to explicitly encourage variation in question structure, phrasing, and underlying concepts/values, reducing redundancy and the manual filtering burden on experts.

Formatting: A lightweight post-generation formatting check is recommended to catch LaTeX/notation issues.

Validation dashboard: A dedicated dashboard integrated into MathE is under development, allowing both administrators and the technical team to review, flag (accept/reject/revise), and track generated MCQs more systematically feeding identified issues back into prompt and pipeline improvements.

6 UC3P1 Weather Data Query Service (WDQS)

6.1 Pilot Summary

UC3P1 implements the Weather Data Query Service (WDQS), the first Weather use case pilot of DataGEMS. The pilot enables users to query meteorological and climate datasets through natural language, with an initial focus on ERA5-Land reanalysis data for Greece and Switzerland for the periods of 1990 to 2020. The objective is to reduce the technical barrier currently faced by users who need climate indicators but do not necessarily have expertise in handling large-scale gridded meteorological datasets, specialised formats such as NetCDF/GRIB, SQL querying, or geospatial data processing.

For the M18 evaluation cycle, the pilot focused on the first operational WDQS scenario using ERA5-Land data. The evaluation combined technical testing of the underlying Text-to-SQL developed framework (named STRATOS) on the ZHAW infrastructure with user-perspective testing through the DataGEMS Weather frontend.

The evaluation covered representative weather and climate queries of increasing complexity, including monthly temperature summaries, climatological monthly averages, annual threshold exceedance counts, frost-day counts, tropical-night counts, yearly regional temperature extremes etc.

Field	Description
Pilot Code	UC3P1
Full Name	Weather Data Query Service (WDQS)
Use Case	UC3 Weather
Target Users	Climate scientists, civil protection authorities, urban planners, sectoral decision-makers, educators and journalists
DataGEMS Components	Natural language interpretation service (text-to-geo, text-to-SQL), query planning, climate data access layer, SQL execution, visualization service
Lead Partners	ZHAW, NOA
Point of Contact	Zhang Yi (ZHAW), Vassiliki Kotroni & Antonis Bezes (NOA)
Data Domain	ERA5-Land reanalysis data Switzerland and Greece
Implementation Period	M6 – M18
Evaluation date	10-25/6/2026
Number of Pilot Participants	15 pilot testers and contributors: 5 from ZHAW and 10 from NOA

Table 12 UC3P1 Summary

6.2 Implementation Status

By M18, the first implementation of the Weather Data Query Service (WDQS) had reached a functional end-to-end stage. The system can process a natural language weather or climate query, identify the requested meteorological variable, location, time period and aggregation logic, generate the corresponding SQL query, execute it against the ERA5-Land database, and return the result to the user.

The main technical implementation and benchmark-oriented evaluation were carried out on the ZHAW infrastructure, which provided a stable and controlled environment for developing and testing the spatio-temporal Text-to-SQL framework. This evaluation focused on the core technical components of the system, including natural language interpretation, spatial entity resolution, meteorological semantic translation, SQL generation and query optimisation. The detailed technical results are reported in a scientific paper prepared for submission to the VLDB 2027 conference.

Following this technical validation, the system was evaluated from the user perspective through the DataGEMS Weather frontend. In this step, 5 from ZHAW and 10 NOA users tested representative natural language queries of increasing complexity, assessing whether the implemented workflow could support realistic interaction with ERA5-Land data through the DataGEMS platform interface.

For this first implementation cycle, WDQS focuses primarily on temperature-related ERA5-Land queries for Greece and Switzerland. This scope allows testing a broad set of representative meteorological operations, including unit conversion, monthly and yearly aggregation, climatological summaries, threshold-based indices, annual counts and regional extrema.

The WDQS pipeline currently includes:

- Natural language interpretation of the user query.
- Identification of the requested variable, location, time period and aggregation type.
- Geospatial resolution of cities and administrative regions.
- Translation of meteorological concepts, such as frost days, tropical nights and threshold exceedance events, into explicit computational definitions.
- Text-to-SQL generation.
- Query optimisation for efficient execution over large spatio-temporal datasets.
- Delivery of tabular and/or graphical results through the Weather frontend.

The implementation is considered sufficiently mature for controlled expert testing, while further work is still required to expand the set of supported variables, indices, locations and user groups.

6.3 Evaluation Results

The WDQS evaluation was conducted in two complementary stages. The first stage was the technical and benchmark-oriented evaluation on the ZHAW infrastructure. It assessed the ability of the framework to translate natural language meteorological questions into executable SQL queries, resolve symbolic locations into spatial filters, map weather and climate concepts into formal definitions, and optimise computationally demanding spatial queries.

The second stage was the user-perspective evaluation through the DataGEMS Weather frontend. Ten NOA users tested representative queries.

The tested queries were selected to cover progressively increasing complexity as the examples show:

- What is the mean monthly temperature in Athens for July 2000?
- What is the mean monthly temperature in Athens for each month of the year during the period 1991–2020? Report by month.
- Find how many days per year during summer temperature exceeds 35 degrees Celsius in Athens during the period 1991–2020. Report per year.
- Find the number of frost days in Florina per year during the period 1991–2020. Report per year.
- Find the number of tropical nights in Athens per year during the period 2010–2020. Report per year.
- Show lowest temperature of Thessaly Region for each year from 2010 to 2020. Report per year.

These examples tested basic spatial, temporal and variable interpretation, climatological aggregation over a 30-year period, seasonal threshold exceedance logic, climate-index definitions, and regional annual extrema. The initial user-perspective evaluation confirmed that the WDQS frontend can support realistic weather and climate data exploration tasks. Detailed structured user feedback is still being collected and will be reported in the next evaluation cycle.

Dimension	Measure	Target	Observation / Result
System	End-to-end query execution	Natural language query is processed and a result is returned	Achieved for the first ERA5-Land pilot scenario. The technical workflow was evaluated on the ZHAW infrastructure and then tested from the user perspective through the DataGEMS Weather frontend
	Text-to-SQL generation	The system generates executable SQL from natural language weather/climate queries	Achieved. The developed framework demonstrated the capacity to generate SQL for meteorological queries requiring spatial, temporal and semantic interpretation.
	Query optimisation	Spatial and temporal queries should be executable within	Partially met. The ZHAW implementation includes query-optimisation mechanisms that improve the

		an operationally meaningful time	execution of spatial queries over large meteorological datasets. Response times should be measured and checked to remain within predefined limits across a broader set of query types, under DataGEMS platform conditions.
Data	Correct interpretation of meteorological variable	Requested variables are mapped to the correct ERA5-Land fields and units	Achieved for the tested queries. Unit conversion and temperature aggregation were included in the evaluated workflow.
	Correct interpretation of temporal constraints	Years, months, seasons and reporting periods are correctly interpreted	Achieved for the tested examples, including specific months, periods of time, monthly summaries, seasonal and annual counts.
	Correct interpretation of spatial constraints	City and regional locations are resolved into appropriate spatial filters	Achieved. The tested locations covered both city-level and regional queries.
	Correct handling of derived climate indices	Concepts such as frost days and tropical nights are translated into explicit thresholds	Partially met. The tested examples confirmed that the system can process such concepts, while additional expert validation is needed for a wider set of indices.
Human	Expert plausibility assessment	Domain experts assess whether the outputs are meaningful and relevant	In progress. 5 from ZHAW and 10 NOA users tested representative queries through the Weather frontend. Initial feedback confirms the relevance of the scenarios and the usefulness of the service. A more structured user-status analysis is ongoing.

	Usability and interpretability	Users should be able to formulate queries without programming expertise	Not started. The pilot demonstrates that natural language access to ERA5-Land climate information is feasible. Further interface refinement and systematic feedback collection are required.

Table 13 UC3P1 Evaluation Results

6.4 Success Criteria Assessment

Success Criterion (from D7.2)	Status	Evidence
Repeated identical queries produce consistent results	**Met**	Initial technical testing on the ZHAW infrastructure indicates stable execution. A systematic repeatability test through the DataGEMS frontend is still required.
Performance aligns with validated benchmark dataset	**Met**	The WDQS implementation is supported by the developed evaluation workload, which provides a benchmark of meteorological natural language/SQL query pairs.
Response time within acceptable operational limits	**Partially Met**	Query optimisation has been implemented and evaluated on the ZHAW infrastructure by 5 scientists from ZHAW. Further monitoring is needed under broader DataGEMS platform usage.
Users assess generated outputs as plausible and meaningful	**Partially Met**	10 NOA users and 5 from ZHAW tested representative weather queries through the DataGEMS Weather frontend. Initial feedback was positive, while detailed structured user-feedback analysis remains ongoing.
NL query is correctly interpreted (spatial, temporal, variable)	**Met** for tested queries	The pilot successfully processed queries involving city-level and regional locations in Greece and Switzerland, different temporal aggregations, meteorological variables and selected weather/climate indices. Broader

		testing with additional variables, indices and regions is planned.

Table 14 UC3P1 Success Criteria Assessment

6.5 Observations and Recommendations

The M18 evaluation confirms that WDQS has progressed to a functional end-to-end implementation for the first ERA5-Land pilot. The service addresses a demanding Weather use case, combining natural language interaction, large spatio-temporal meteorological datasets, geospatial reasoning, domain-specific climate indicators and scalable SQL execution.

The technical implementation and benchmark evaluation on the ZHAW infrastructure provided a stable basis for testing the core STRATOS workflow. In parallel, the DataGEMS Weather frontend enabled initial user-perspective evaluation by NOA experts, who tested representative queries involving monthly averages, annual summaries, threshold exceedances, frost days, tropical nights and regional extrema.

For the next implementation cycle, the following actions are recommended:

- **Continue structured user-perspective evaluation.** The current testing by 5 ZHAW+10 NOA users should be followed by a more systematic questionnaire-based evaluation, covering perceived usefulness, ease of use, trust in the results, clarity of outputs and efficiency gains.
- **Expand the tested meteorological indicators** to more complex ones.
- **Increase spatial diversity.** Test a wider range of spatial cases, including administrative boundaries and user-defined locations.
- **Validate operational response-time thresholds for WDQS.** Query optimisation should be further assessed through measurements across broader spatial and temporal query categories.
- **Improve frontend transparency** by showing the interpreted location, period, variable, threshold and aggregation logic, allowing users to verify how their query was understood.
- Further **consolidate platform-level stability** so that the robustness demonstrated on the ZHAW infrastructure is progressively transferred to the broader DataGEMS environment.

7 UC4P1 Language Feature-based Corpus View

7.1 Pilot Summary

UC4P1 focuses on providing a linguistic feature presentation of a text corpus retrieved from a user specification in natural language. The pilot evaluates the usefulness of these capabilities to the workflow of linguists and researchers in general, the perceived transparency of the underlying system, as well as the relevance of the texts that were retrieved.

At the current stage, evaluation is performed on a small group of linguists over an initial implementation that supports the “Encyclopedia Britannica” dataset and a small set of linguistic features (i.e., term frequency, word collocations, and sentiment analysis). The decision for this initial implementation was to create a minimum viable product (MVP) that implements the basic architecture and communication pipelines of the used components, while remaining easily extendable for the additional planned features and configurable to any received feedback.

Field	Description
Pilot Code	UC4P1
Full Name	Feature-based Corpus View
Use Case	UC4 Language Data Infrastructure
Target Users	Digital humanists, computational linguists, researchers with limited programming expertise
DataGEMS Components	Cross-dataset discovery, in-dataset discovery, unstructured profiling
Lead Partners	ARC, IDS
Point of Contact	George Katsogiannis (ARC), Thora Hagen & Andreas Witt (IDS)
Data Sources	EncycNet (German), 19th-Century Encyclopædia Britannica (English), Diderot's Encyclopédie (French), Wikipedia subset
Implementation Period	M16 – M18
Evaluation date	18/06/2026
Number of Pilot Participants	5 from IDS

Table 15 UC4P1 Summary

7.2 Implementation Status

By June 2026, UC4P2 has reached a functional deployment state and showcased the expected capabilities and functionality. The feature-based corpus view is operational and has been used by linguists and DataGEMS partners alike.

Key implementation milestones achieved include:

- Design and deployment of the User Interface (UI) that allows the user to submit a corpus request, select the desired linguistic features, and view the resulting features and retrieved text corpus.
- Corpus retrieval based on the user request has been implemented by Cross-Dataset Discovery in a two-stage approach. First stage comprises of coarse-grained retrieval of 100 text passages from the Britannica dataset, while the second stage filters irrelevant texts with greater precision, resulting in fine-grained results.
- Calculation of linguistic features for the retrieved text corpus has been implemented by In-Dataset Discovery, currently supporting three types of features: (i) term frequency, (ii) word collocations, and (iii) sentiment analysis. The respective endpoint has been designed to be easily extensible to the additional planned features.
- Communication of underlying services through the DataGEMS gateway has been implemented, facilitating the modular extension of each service without disrupting the defined workflow.

The UC4P1 can be considered as an MVP that showcases the workflow and functionality of the final application. A subsequent pilot iteration is planned to extend the supported features and capabilities and to be evaluated by a broader audience of experts.

7.3 Evaluation Results

The UC4P1 was evaluated both online by the aforementioned group of experts, as well as offline by the technical team to extract more quantitative metrics. The table below showcases the calculated metrics in the three dimensions defined for this pilot (i.e., System, Data, Human). To calculate retrieval scores, a small benchmark was manually created over the used dataset (“Encyclopedia Britannica”) where a linguist defined natural language questions (NLQs) along with the ground truth texts passages that should be retrieved for each question.

Additionally, the UC4P1 workshop held on June 18th with linguists from IDS was used to evaluate the user perspective of the implemented service. This evaluation offers valuable insights into the usability and potential value of the service to its users. After a brief presentation of the project and service, the linguists were able to freely use and explore the implementation of the feature-based corpus view. At the end of the workshop, participants were asked to complete a questionnaire combining closed and open-ended questions, designed to capture their perceptions of the system’s usefulness, transparency, and efficiency. Responses to the closed-ended questions were collected using a 5-point Likert scale (1 = strongly disagree/very poor, 5 = strongly agree/excellent).

Dimension	Measure	Target	Observation / Result
System	Average Response time	The system should response to user requests within a	19.63s

		reasonable time for pilot use	
	Average Retrieval time	The system should retrieve results within a reasonable time for pilot use	17.77s
Data	Document Retrieval Recall	The retrieved results should contain texts that are relevant to the user request	83.33%
Human	Intuitiveness and helpfulness of NL interface	Majority ≥4/5	Mean score 3.8/5, 80% rated ≥4.
	Perceived Relevance of Retrieved Texts	Majority ≥4/5	Mean score 4/5, 80% rated ≥4.
	Coverage of needs by provided linguistic features	Majority ≥4/5	Mean score 3.2/5, 20% rated ≥4. Majority rated 3.
	Clarity and Usability of Dashboard	Majority ≥4/5	Mean score 4.2/5, 100% rated ≥4.
	Possibility of using the system of their research	Majority ≥4/5	Mean score 3.6/5, 60% rated ≥4.
	Perceived Efficiency Gain	Majority ≥4/5	Mean score 3.6/5, 40% rated ≥4. Majority rated 3.
	Importance of export functions	Majority ≥4/5	Mean score 5/5, 100% rated 5.
	Perceived understanding of how to use the system	Majority ≥4/5	Mean score 4/5, 60% rated ≥4.
	Perceived Transparency	Majority ≥4/5	Mean score 3.6/5, 40% rated ≥4. Majority rated 3.
	Confidence in explaining results	Majority ≥4/5	Mean score 3.6/5, 60% rated ≥4.
	Perceived efficiency of individual steps	Majority ≥4/5	Mean score 4.6/5, 100% rated ≥4.

Table 16 UC4P1 Evaluation Results

7.4 Success Criteria Assessment

Success Criterion (from D7.2)	Status	Evidence
Response time scales acceptably with retrieved data volume	**Met**	Average response time is observed to be below 20s, which is adequate given the volume of data and the filtering process that guarantees high quality results.
Relevant documents rank above random baseline for thematic queries	**Met**	Both the measured recall score and the perceived relevance reported by users are substantially high.
Users perceive efficiency gain vs. manual workflows	**Partially Met**	Majority of users perceive that the service meets the time needed to complete similar tasks manually, while some users report efficiency gains.
Users perceive system transparency	**Partially Met**	Perceived transparency scores are adequate but could still be improved.
Exported data usable in downstream NLP/statistical tools	**Pending**	This feature will be implemented and evaluated in the next pilot cycle.
Cross-language retrieval functional	**Pending**	This feature will be implemented and evaluated in the next pilot cycle.

Table 17 UC4P1 Success Criteria Assessment

7.5 Observations and Recommendations

The evaluation at M18 shows that feature-based corpus view is already in a mature state and confirms that it has progressed to a functional implementation for the first Language pilot. The service provides an intuitive and easy-to-use interface that users reported could be integrated in their linguistic research workflow and even possibly accelerating it. While the results derived by this pilot are overall positive, there are still areas for improvement in the upcoming months. These improvements will not only help to enhance the pilot at hand, but also the upcoming Language pilots.

For the next implementation cycle, the following actions are recommended:

- **Additional linguistic features**. While the initial implementation with three types of linguistic features served as an MVP of the service, the following iterations aim to provide richer and more diverse options that will meet the needs expressed by the users.
- **Implementation of export feature**. As it was also highlighted by the pilot users, that the ability to download the retrieved corpus and the computed features is highly desired. This feature will allow users to better integrate our service in their workflows and fully enjoy its capabilities by moving the exported data into other applications or their own code.

- **Cross-language retrieval functionality**. The capability to retrieve texts that are available in different languages to the question would be greatly beneficial to the users. On one hand, this would allow the users to express questions in the language they feel most comfortable with. On the other hand, it would increase the relevant texts that could be retrieved for any given question.
- **Increased transparency in the UI**. While users reported that they found the interface useful and intuitive, some users said they would not be very confident in explaining exactly what is happening in the background. This could be improved by adding simple changes to the UI that explain the exact steps that are happening after the user has submitted their questions and even by showing some intermediate results to clarify how the final features are computed.
- **Multi-dataset Functionality**. The current implementation only allows user to search texts in the "Encyclopaedia Britannica" dataset, as its purpose was to provide an initial end-to-end implementation. In future releases, the users will be able to choose and use additional datasets from the DataGEMS platform.

With the addition of more datasets to facilitate cross-language functionality as well as the addition of more features to the analysis dashboard, the complexity of the workflow will naturally increase. We will continuously monitor key dimensions that are affected by these changes, such as response time and perceived efficiency, throughout multiple iterations of development and validation.

8 Conclusions

This deliverable presented the first partial implementation and evaluation results of the WP7 pilot activities within the DataGEMS project. The evaluation covered five pilots selected for the initial M18 deployment cycle: UC1P2 Math MCQs Recommendation, UC1P3 Math Educational Material Recommendation, UC1P5 Math Administrator Validation for MathE Admins, UC3P1 Weather Data Query Service, and UC4P1 Language Feature-based Corpus View. Together, these pilots provided early evidence of the applicability and adaptability of DataGEMS services across three distinct domains: higher education, weather and climate data access, and language data infrastructure.

Across the five pilots, the DataGEMS components were successfully exercised in realistic operational scenarios. These included adaptive question recommendation, educational material retrieval, AI-assisted MCQ generation and validation, natural language querying of meteorological data, and feature-based exploration of text corpora. The evaluation also confirmed the relevance of the three-dimensional assessment framework defined in D7.2, covering system, data, and human evaluation aspects.

Overall, the evaluated pilots demonstrated a satisfactory level of operational readiness and produced relevant and usable results. Feedback from end users and stakeholders was generally positive, confirming the potential value of the DataGEMS services in supporting data-driven interactions and domain-specific tasks. At the same time, the evaluation identified areas for further improvement, including workflow efficiency, data quality, transparency, usability, and the need for continued refinement based on broader user feedback.

The results reported in this deliverable should be understood as preliminary, reflecting the first step of a phased WP7 evaluation strategy. The pilots included in D7.5a were selected because they had reached sufficient technical and operational maturity by M18 and could involve relevant stakeholder communities. Other pilot activities, including those related to the Lifelong Learning use case, will be introduced in subsequent evaluation cycles as the corresponding services, user communities, and testing conditions become mature.

In conclusion, D7.5a demonstrates that the first group of WP7 pilots has successfully progressed from specification and design toward operational deployment and initial evaluation. The findings provide a sound basis for refining DataGEMS services, improving the evaluation approach, and preparing the broader pilot activities planned for the next reporting period. These preliminary conclusions will be revisited, extended, and consolidated in D7.5b at M36, when additional pilots, larger user samples, and more comprehensive evaluation data will be available.