SMEs – Raising Awareness and Learning on Digital data, data analysis and artificial intelligence

Datathon Report

Prepared by

National Research Council of Italy

Institute for Educational Technology

Co-Funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project Number: 2023-1-IT01-KA220-VET-000151990

The SMERALD Datathon

A report on the Smerald Datathon

Table of Contents

Executive summary

This report outlines the objectives, activities, and outcomes of the SMERALD Datathon, organized as part of the SMERALD (SMEs – Raising Awareness and Learning on Digital data, data analysis and artificial intelligence) project.

The Datathon's primary aim was to promote the practical use of Digital Data, Data Analysis, and Artificial Intelligence among SMEs, nonprofits, and public sector organisations. Defined as a time-bound event combining the creativity of a hackathon with the analytical rigor of data science, the Datathon focused on hands-on problem solving using real-world datasets.

The event was held in a hybrid format, both in-person at the CNR Research Area in Palermo and online, between June 13th and 16th 2025. Participation included SMEs representatives, nonprofits, public sector representatives, and educational professionals. A keynote speech titled "Allenare il futuro: come i Data Hackathon sviluppano talenti e idee vincenti" was also organized for the participants in presence. The Datathon provide also participants with the opportunity of cooperating in open working tables to discuss data quality and the challenges and potential of utilizing generative AI.

Core Challenges

Participants addressed four main challenges, aiming to produce documented use cases and content integration proposals for the SMERALD platform:

- Challenge 1: Leveraging AI and Data for Organisational Improvement
    - *How SMEs apply AI and data analytics to improve internal processes, enhance customer service, and address real organisational challenges*
- Challenge 2: Designing Effective and Engaging Data and AI Training
    - *What data skills are most essential for SMEs, and how can we turn real-world problems into meaningful, tailored learning experiences for staff?*
- Challenge 3: Data Quality Assessment and Data Enhancement
    - *Evaluate and propose improvements to the quality of data available identifying gaps in existing datasets that are relevant to SMEs and local ecosystems.*
- Challenge 4: Open Challenge: Bring Your Own Idea
    - *Participants are encouraged to propose custom challenges that align with the SMERALD vision of using digital data, data analysis, and AI to support innovation and learning in SMEs and small organisations.*

Aim and Objectives

The SMERALD Datathon was designed as a central moment of the project to drive awareness and practical application of data literacy.

In SMERALD the aim of the Datathon was to promote the practical use of Digital Data, Data Analysis, and Artificial Intelligence among SMEs, nonprofits, and public sector organisations.

The event sought to accomplish the following specific objectives:

- Engage stakeholders in co-design sessions based on real needs.
- Leverage SMERALD project use cases and educational content.
- Explore and prototype use cases tailored to daily organisational challenges.
- Identify needs and co-create customised plans.
- Create replicable event formats and outputs to feed the project ecosystem.

Timeline

This section details the operational structure of the SMERALD Datathon, covering its logistics, timeline, the profile of participants, and the content delivered during the key sessions. The SMERALD Datathon was structured using a hybrid format, facilitating both in-person participation in Palermo, Italy, and online engagement.

The event spanned several days, starting with the challenge launch on Friday, June 13th. The Friday session included a brief introduction to the SMERALD project and the Datathon activities, launching the challenges that participants would run online. Remote participants were permitted to work asynchronously; constant online presence was not required.

The main component of the Datathon for those attending in person was held on Saturday, June 14th, 2025. The in-person activities took place at the Area della Ricerca del Consiglio Nazionale delle Ricerche (CNR), located in Via Ugo La Malfa, 153 in Palermo. In-person participants presented their Datathon results at the end of the day on Saturday.

Final presentations including participants working remotely on the challenges were organized on **Monday, June 16th, at 11:00 CET**.

Participants

The Datathon successfully engaged a diverse group of stakeholders, reflecting the project's goal of raising awareness among various organizational types. The participants included:

- **Local/regional SMEs**.
- **Nonprofits**.
- **Public sector representatives**.
- **Educational professionals and stakeholders**.

Additionally, the event featured contributions from the academic world, with a keynote talk from prof. Antonella Longo designed to stimulate discussion, present high-level policy, and showcase innovation.

The event began with a significant keynote speech delivered by Prof.ssa Antonella Longo of the Università del Salento. Her intervention was titled "Allenare il futuro: come i Data Hackathon sviluppano talenti e idee vincenti" (Training the Future: How Data Hackathons Develop Talents and Winning Ideas). This talk aimed to offer valuable input to the dialogue among the training, research, and business communities, inspiring participants regarding the potential of Datathons for innovation and skills development.

A central feature of the in-person event involved two open working tables. These technical discussion groups concentrated on critical thematic areas:

1. Data quality
2. The challenges and potential of utilizing generative artificial intelligence

Datathon Challenges and Outputs

This chapter details the specific challenges proposed to participants during the SMERALD Datathon and outlines the concrete deliverables expected from their collaborative data science and innovation efforts. The challenges were designed to be hands-on and relevant to the contexts of SMEs, non-profits, and public sector organizations, aligning with the project's overall goal of fostering practical data and AI usage.

Challenge Overview

The Datathon featured four distinct thematic challenges, allowing participants to choose areas that best matched their expertise or organizational needs.

Challenge 1: Leveraging AI and Data for Organisational Improvement

This challenge was focused on the practical application of advanced technologies within organizations. Participants were asked to explore how SMEs apply AI and data analytics to improve internal processes, enhance customer service, and address real organisational challenges. This area aimed to generate innovative solutions for efficiency and service delivery using data as a core asset.

Challenge 2: Designing Effective and Engaging Data and AI Training

This training-focused challenge addressed the critical gap in data literacy and skills within SMEs. The core questions posed to participants were: What data skills are most essential for SMEs, and how can we turn real-world problems into meaningful, tailored learning experiences for staff?. This challenge was crucial for developing content integration proposals for the SMERALD platform, ensuring the educational material is practical and engaging.

Challenge 3: Data Quality Assessment and Data Enhancement

Recognizing that the success of AI depends heavily on data quality, this challenge placed a strong emphasis on data foundational issues. The task was to evaluate and propose improvements to the quality of data available, identifying gaps in existing datasets that are relevant to SMEs and local ecosystems.

This challenge directly links to the technical discussions held during the Datathon regarding:

- The essential quality requirements for data (including accuracy, consistency, completeness, timeliness, and representativeness).
- The critical problem of timeliness (e.g., water quality data often being delayed by up to a year due to required laboratory validation).
- The need for technical mechanisms, such as the Data Quality Vocabulary (DQV) (a W3C standard), to specify data quality levels (e.g., indicating if data is non-validated or incomplete) to maintain transparency even when publishing imperfect data.
- The legislative backing provided by the Italian Transparency Decree (Decreto Trasparenza, Art. 6), which stipulates that the necessity of ensuring adequate quality cannot justify omitting or delaying the publication of data.

Challenge 4: Open Challenge: Bring Your Own Idea

This challenge provided maximum flexibility, encouraging participants to propose and pursue their own data-related projects. Participants were encouraged to propose custom challenges that align with the SMERALD vision of using digital data, data analysis, and AI to support innovation and learning in SMEs and small organisations.

Concrete Outputs

The Datathon was specifically organized to produce tangible results that could be integrated back into the SMERALD project framework and shared with a broader community.

The expected outputs were:

- **Documented use cases in real-world contexts**. These documented solutions serve as practical examples of data application.
- **Content integration proposals for the SMERALD platform**. These proposals ensure that the project's educational platform is enriched with content based on practical needs and Datathon findings.
- **Case studies for replication by third parties in different contexts**. These documents help achieve the objective of creating replicable event formats and outputs to feed the project ecosystem.

Furthermore, the insights gained, particularly regarding the use of advanced tools like **Large Language Models (LLMs)**for data harmonization (e.g., extracting CSV from unstructured PDFs) and the adoption of Linked Data standards (like the Semantic Sensor Network - SSN), were intended to form the basis for future technical development within the project.

Results

The results and concepts developed during the Datathon. The main results presented here come from the concepts developed and derived from the different teams contributing to the Datathon.

Eurotraining project visualization and classification tool

This result emerged from an activity combining the use of AI and data analytics. The project aimed to organize and visualize Eurotraining projects. Initially, there was no existing classification system detailing the categories, start, and end dates of the projects.

The first step involved gathering all project information, including key activities, from the dedicated section of the Eurotraining website. Subsequently, AI was utilized to help classify the projects and define categories, which did not exist before.

The second phase involved data cleaning processes. This was necessary because the AI sometimes duplicated projects, assigning them to multiple categories (e.g., transition/circular economy and digital activities) if the project content was relevant to both.

The final result is a chart visualizing several categories, including transition circular economy, vocational education and entrepreneurship, social innovation, culture, creativity, art, youth inclusion, and refugees. The interface created within the Datathon is interactive, clicking on a section the related projects are shown. Specific filters (e.g., for higher education or academic programs) are also available.

This solution was deemed useful for both organizational needs and internal departments. It is considered a preliminary solution that can be improved. Technically, the visualization is easy to embed on a website using an iframe. This tool is especially important for networking, helping potential consortium members or partners quickly understand Euro Training's projects and fields of activity, providing an immediate idea of potential collaboration. It allows external users to decide where to stop and where to engage further.

"AI Ready" SME Toolkit Concept

Catro presented a conceptual outcome developed by the team during the Datathin. The goal was to demonstrate that AI and data can be accessible, practical, and valuable even for smaller organizations.

The concept addresses the main issues faced by Small and Medium Enterprises (SMEs) today: limited time and budget, insufficient digital skills, internal processes that are still manual, and collected data that is not sufficiently utilized. For many SMEs, AI feels intimidating, irrelevant, or they are simply unaware of how to use it.

Focusing on Human Resources (HR), common SME challenges include insufficient internal workflows, time-consuming recruitment processes, inconsistent customer experience, poor decision-making due to lack of insights, and lower employee engagement.

AI and data solutions proposed to mitigate these challenges include:

- Using CV screening and chatbots for manual HR tasks.
- Employing AI-driven tools that offer personalized learning recommendations to fill training gaps.
- Applying sentiment analysis to reviews or emails for customer feedback.
- Utilizing business dashboards to improve data-driven decision-making.
- Conducting AI-driven anonymous engagement surveys.

The central concept is the AI Ready SME Toolkit, designed to be low-cost, solution-oriented, practical, and modular. Key proposed elements are:

1. A self-assessment tool for SMEs to estimate their AI readiness.
2. A process mapping canvas to assess where automation can assist internal processes.
3. An AI-powered HR assistant for recruitment, feedback, and engagement surveys.
4. Dashboard templates to visualize key business trends with real-time data.

5. A training module covering AI basics, ethical issues, and GDPR for non-technical employees.

This toolkit is designed to work because it uses low-cost, "no-code" tools, requires no in-house AI expertise, builds on data that SMEs already collect, solves real business pain points, and promotes ethical and inclusive use of artificial intelligence. The ultimate goal is to bring AI closer to SMEs, focusing on human capital, time, and quality, thus empowering them to grow sustainably and smartly.

Datathon Results and Proposals from Italy

CNR-ITD reported on the in-person Datathon held in Italy. Two groups addressed different challenges:

Group 1: Water Data Quality and Transparency

This group focused on problems related to water data, drawing inspiration from the Open Data Sicilia community, which successfully transformed regional government data published in non-machine-readable PDF files into structured, usable formats, even creating a Telegram bot to provide real-time information.

Problems Identified:

1. Lack of Real-Time Data: Published data is often delayed (e.g., after a week), making it ineffective for immediate monitoring of water levels or supply issues.
2. Reusable Data Deficiency: While much data is produced (quantity data), most cannot be reused directly and requires extensive cleaning (quality data).
3. Delay due to Validation: Institutions often delay publication, citing the need to validate data, a time-consuming process better suited for historical records than for current monitoring.

Solutions and Tools Proposed:

- Legislative Leverage: Utilizing the Italian "transparency decree" which permits providers to publish data even if unvalidated.
- Technical Publication Standards: Publishers should indicate the quality level and validation status of the data upon publication, allowing users to proceed with the data while being informed.
- LLM Automation: Using LLMs to automate the workflow management for publication, including enriching metadata and automatically classifying data.

Group 2: Data Quality and Metadata Management

This group focused on metadata management and data quality, analyzing key dimensions like completeness and accuracy (both syntactic and semantic), in line with ISO standards and legal compliance.

Problems Identified:

- Lack of Interoperability: Institutions rely on large software providers whose solutions are often non-interoperable, making it difficult to harmonize published data.

Solutions Proposed:

- Mandatory Requirements for Software Vendors: Imposing contractual or legislative obligations on software suppliers to support data quality and interoperability aspects.
- Pipeline Design: Designing a specific pipeline for managing data quality.

Competence Framework Video Project

Blinc presented a separate initiative aimed at making the Smerald competence framework easier to understand and communicate.

The main goal was to effectively explain the concept of **competence-oriented learning** and the Smerald competence framework, which is often unclear to partners.

The methodology involved utilizing **Perplexity** as an interview partner. The AI asked targeted questions (e.g., why use a competence framework, how it helps identify gaps), forcing the presenter to provide detailed explanations.

The result is a set of 11 video recordings, ranging from 6 to 15 minutes, intended to be cut down into shorter, more usable clips. This interactive approach is considered more engaging than standard text or video resources.

A technical challenge arose in recording the interaction with Perplexity from the browser. Blinc had to modify the recording software (OBS Studio) with custom code to correctly capture the audio and display the screen, as standard recording tools are not designed for this. This approach aims to demonstrate how AI can be an effective tool for both learning and communication in the future.

In essence, the presented outcomes of the SMERALD Datathon demonstrated practical uses of AI and data: from creating interactive project overviews and conceptualizing solutions for SMEs (AI Ready Toolkit), to analyzing the technical and legal difficulties of public data quality

and innovative methods for producing engaging training material using AI (Competence Framework Video Project).

Conclusion

The discussions of the results of the Datathon highlighted that the success of AI development in Europe largely depends on a successful European data strategy. Achieving this requires adequate data governance and management practices focusing primarily on data availability and quality.

Data Governance and Quality

- Legislative Compliance: Relevant policies include the European Data Strategy, the AI Act, and the Open Data Directive. Discussions emphasized the need to ensure adequate data quality cannot constitute a reason for the omission or delayed publication of data.
- Quality Issues (Timeliness): A critical issue identified was data timeliness. For example, water quality data is often available only with a significant delay (up to a year or more), not in *real time*, due to the lengthy laboratory validation process.
- Quality Standards: Data must be accurate, consistent, complete, timely, and representative. For AI systems, traceability is also a crucial requirement. Tools like the Data Quality Vocabulary (DQV) (W3C standard) can be used to technically specify the quality level of a dataset (e.g., if it is non-validated or incomplete) to provide users with transparency.
- Geospatial Data: high quality geospatial data are vital for producing geographical High-Value Datasets (HVDs). A required geometric precision is needed to achieve conformity with the INSPIRE Directive.

Interoperability and AI Tools

- Modeling and Linked Data: Technical solutions proposed included the use of existing models, such as the water quality monitoring ontology based on the Semantic Sensor Network (SSN) standard (W3C), which can be adapted to various regional contexts. Tools like RML (RDF Mapping Language) can be used to define pipelines for transforming structured data (e.g., CSV) into RDF (Linked Data).
- Automation: open-source toolkits were referenced as a software solution that can implement an automated pipeline for data extraction, transformation, publication, and validation using a workflow management system.

- AI Integration: The potential use of Large Language Models (LLMs) was discussed, particularly for data harmonization and for extracting structured data (CSV) from "horrendous PDFs" that are designed for human reading. LLMs could also assist in the automatic preparation and metadation of data according to standards like DCAT-AP.

In conclusion, the Datathon provided practical insights into the challenges faced by organizations in adopting data analytics and AI, strongly emphasizing that foundational issues related to data quality, governance, and the adoption of European standards must be addressed to unlock the full potential of AI and Data innovation.