AIML4OS

Artificial Intelligence and Machine Learning for Official Statistics

Newsletter 3

Welcome to our third newsletter for ESSnet AIML4OS! These updates share the project's progress, results, events, and other key news

PROJECT OVERVIEW

The main objectives of AIML4OS are to explore the use of Artificial Intelligence/Machine Learning (AI/ML) for the production of official statistics and to implement innovative solutions for statistical products and processes. This four-year project started in April 2024, with activities structured in the following work packages

OVERARCHING WORK PACKAGES

WP1
Project management and coordination

WP2
Communication and community engagement

WP3
ESS AI/ML lab: Technical infrastructure and organisational setup

WP4
AI/ML state-of-play and ecosystem monitoring

WP5
Standards, methodological and implementation frameworks

WP6
Knowledge repository and training material

USE CASES

WP7
AI/ML on earth observation data, satellite imagery

WP8
Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on editing

WP9
Imputation focus - Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on imputation

WP10
From text to code - Experiences and potential of the use of AI/ML for classifying and coding

WP11
Applying ML for estimating firm-level supply chain networks

WP12
Large language models

WP13
Generation of synthetic data in official statistics: techniques and applications

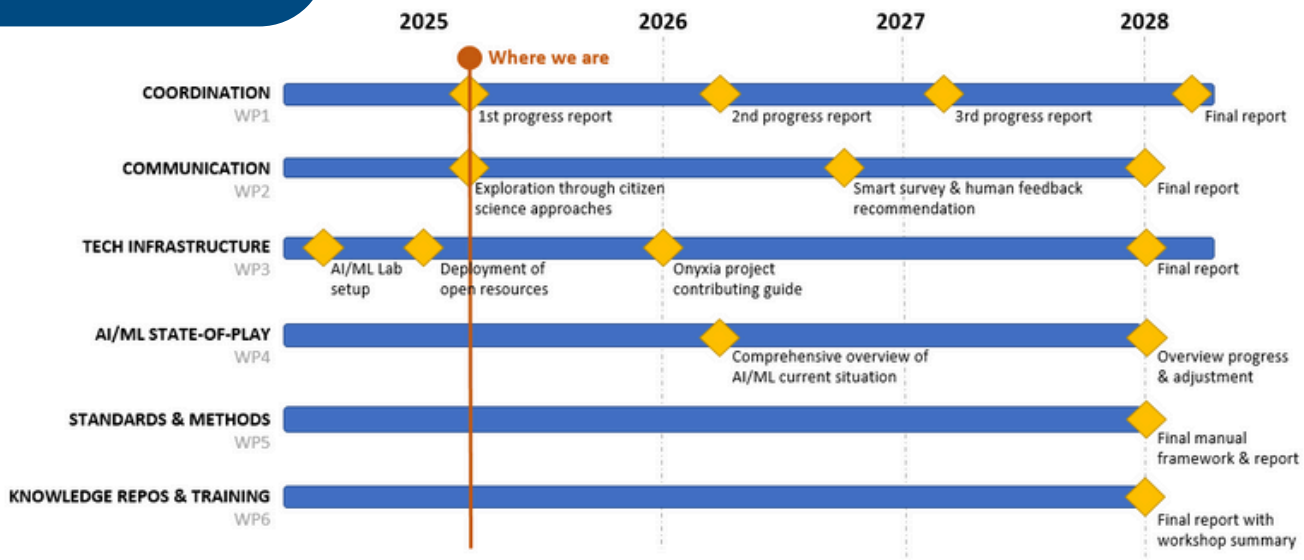
In this issue deciding the future roadmap, One Year On: looking back, moving forward & special session @ NTTS 2025

READ MORE...

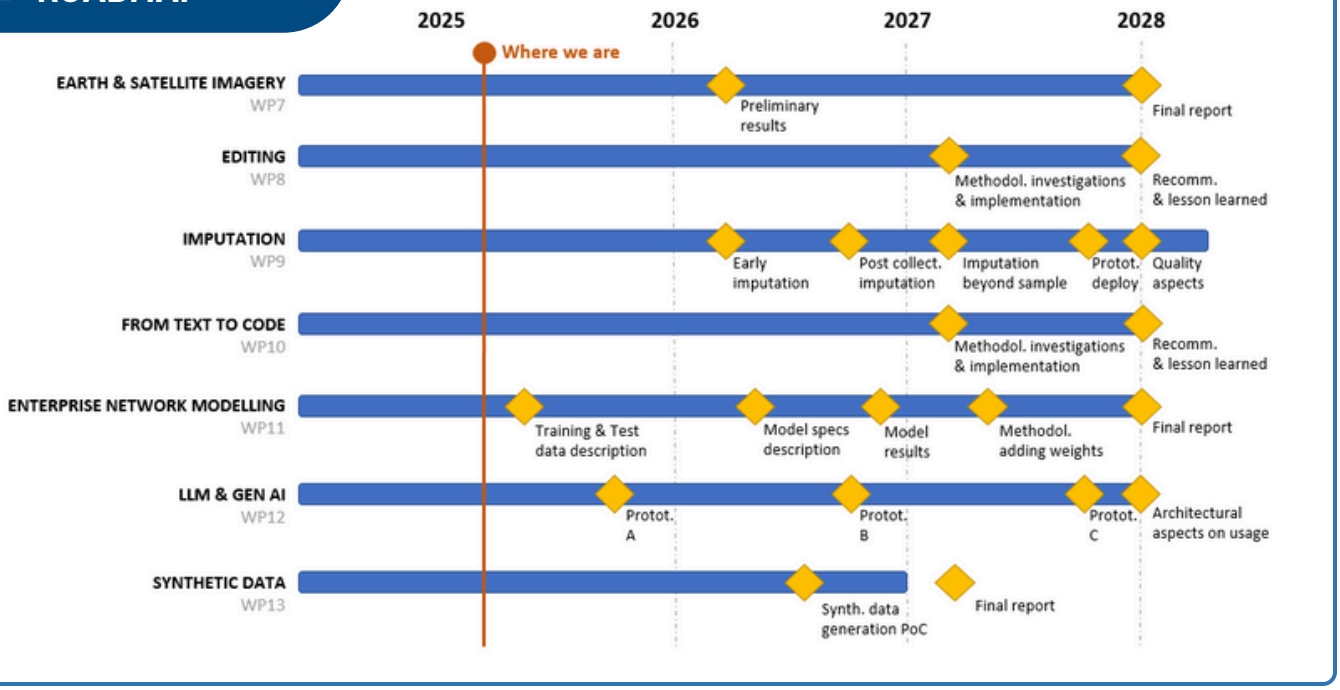
PROJECT OVERVIEW

For each WP involved, the project is divided into several phases. Below are descriptions of what will be achieved and when.

1 OVERARCHING WPS ROADMAP



2 USE CASES ROADMAP



DECIDING THE FUTURE ROADMAP

NTTS Meeting Highlights Common Approaches for Use Cases

Brussels, March 10, 2025 –
Representatives from various working groups gathered at NTTS for a pivotal face-to-face (F2F) meeting to establish a unified strategy for advancing common approaches between different Use Cases (UCs).

The discussions focused on three key areas: platform knowledge sharing, use case knowledge sharing and repository & code standards.

Platform Knowledge Sharing: The discussion focused on effective knowledge sharing within the platform. **Two approaches were explored: datasets creation** – considered for tasks like data imputation and editing (but faced challenges in cross-team usability) and **data generation via code** – deemed more practical and broadly useful. Each UC should specify whether they produce code, code with examples, or full documentation for easier navigation. Since privacy concerns is a main topic in official staistics, **the meeting emphasized synthetic datasets for prototypes**, proposing: simple illustrative datasets, metadata to enable independent dataset generation and **agents to generate datasets** from metadata.

Use Case Knowledge Sharing: The discussion focused on strengthening knowledge-sharing among UCs. **Key proposals** included: creating **datasets for training materials**, identifying **common features** to develop useful datasets and **evaluating real vs. synthetic** data, using predictive modeling or deterministic rules. **A synthetic data agent was proposed**, with an example of simulating firms for administrative and survey data. The meeting also stressed integrating **Funathon activities** with WP6, suggesting an early launch with a basic setup to drive post-event material development. Synthetic data was highlighted for its known distributions and privacy benefits.

Repository & Code Standards: Ensuring standardized and accessible code was another central theme of the discussion. The group discussed on: the **basis for code standards** (including documentation, executable code, and sample data), the necessity of runnable code and **the provision of ready-to-use environments** and deciding whether standards should be set at the WP level or collectively.
The **use of synthetic or open data** was proposed as a viable solution for

privacy concerns. To enhance collaboration, **SSP Cloud** was recommended as a **common platform**. A WP3 **hackathon** was suggested to **showcase activities**, familiarize teams with the platform, and **promote a unified approach**.

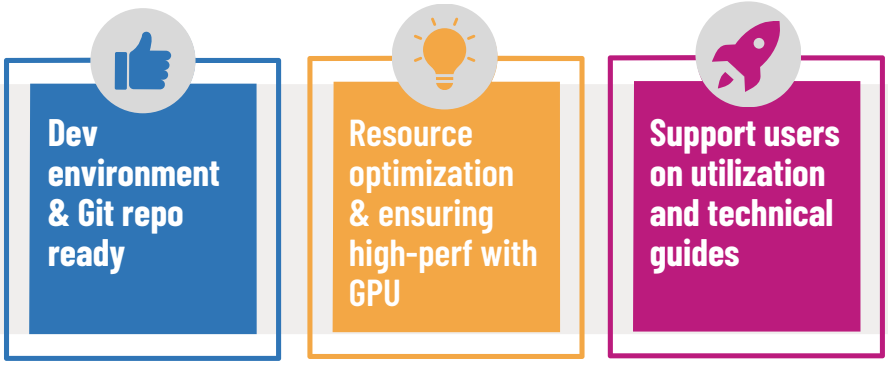
Conclusion and Future Activities: **Several key directions** emerged from the discussions, highlighting the next steps in the collaborative process. A major focus will be on **transforming these ideas into practical applications**, ensuring that the proposed strategies move from theory to implementation. The **hackathon** was identified as a **crucial initiative** for grounding these concepts, providing a space to experiment, refine, and integrate different approaches in a hands-on environment. Another important aspect will be **defining maturity levels for product releases**. Establishing clear criteria for these levels will help standardize outputs and align expectations across teams. Additionally, significant emphasis was placed on **dataset generation and the agent-based approach**, both of which are essential for creating a solid foundation for future developments.
By advancing these structured efforts, the F2F meeting aims to **foster a more cohesive and effective framework for cross-UC collaboration**, ultimately driving innovation and enhancing knowledge sharing within the community.

FACE-TO-FACE MEETING @NTTS

Watch the reply

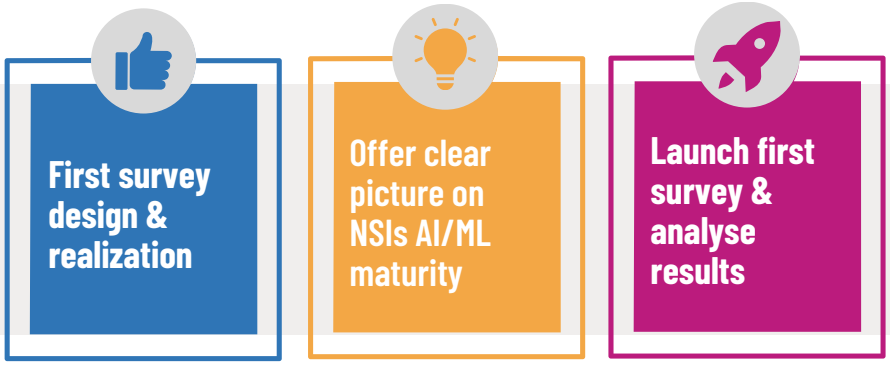
ONE YEAR ON: LOOKING BACK, MOVING FORWARD

WP3 - ESS AI/ML LAB AND FUTURE PERSPECTIVES: ADVANCES IN SSP CLOUD



WP3 continues to advance artificial intelligence, machine learning, and cloud computing. This year, two key deliverables achieved.
The first launching the ESS AI/ML Lab through the **SSP Cloud**. A comprehensive report and **documentation** guide users on leveraging the platform and deploying open resources.
Beyond deliverables, WP3 supports other WPs with technical guidance. WP6 benefits from learning material solutions, while a Git repository facilitates code storage and deployment across WPs. Efforts to optimize the SSP Cloud include resource management rules and on-demand high-performance computing access, such as GPUs, ensuring smooth operations.
Looking ahead, WP3 plans three additional deliverables: a contributing guide to Onyxia by year-end, a technical guide for similar sandpits, and a final report by 2027.

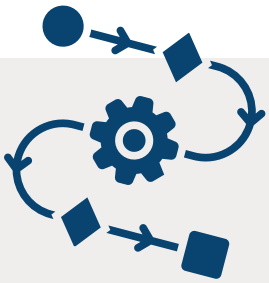
WP4 - SURVEY ON AI/ML USE ON NSI: READY TO LAUNCH



WP4 has initiated an important survey **to gather insights** from National Statistical Institutes (**NSIs**) within the European Statistical System (ESS). The survey's primary objective is to obtain a comprehensive understanding of **the current state of AI/ML** adoption. The survey is designed to highlight evidence on the use of AI/ML, horizon scanning to identify trends and developments that could benefit the field

of official statistics. Additionally, it aims to identify the specific needs of NSIs concerning these technologies, offering a clearer picture of how these institutions are integrating advanced technologies into their statistical processes.
Survey 1 will focus on the institutional level, gathering high-level data from management-level respondents within each NSI 1 by June 2025. Survey 2 will dive deeper into specific projects within the NSIs **to foster networking and collaboration across the ESS.**

WP5 - SHAPING THE STANDARDS OF AI IN OFFICIAL STATISTICS



WP5 is spearheading an initiative to establish robust **standards, methodologies, and implementation frameworks**. It seeks to bridge the gap between theoretical AI/ML advancements and their practical applications in official statistical processes.
By collecting and comparing existing approaches, the team aims to develop guidelines and standardize AI/ML applications. Special emphasis is placed on the **Generic Statistical Business Process Model** (GSBPM), which cover data processing, analysis, and dissemination.
WP5 is also dedicated to refining AI/ML methodologies. The **Total Machine Learning Error** (TMLE) model serves as the foundational framework for evaluating model accuracy and performance. This model takes into account various error types, including estimation, measurement, and sampling errors. The goal is to create methodological guidelines ensuring AI/ML techniques are scientifically sound and reliable.
WP5 is not just about theory—it is also about practice. A comprehensive implementation framework will provide insights into the technical components needed for **production-ready AI/ML solutions** (such as ML-ops), as well as the skills required to support these initiatives.
While it actively develops standards, methodologies, and implementation frameworks, it also **seeks input from experts and stakeholders**. The team invites contributions on existing guidelines, challenges faced in AI/ML adoption, and potential solutions.

WP6 - GATHERING TRAINING MATERIALS & KNOWLEDGE SHARING

Collection of contributions on ML cocepts & workflows

Improve standardised knowledge sharing

Feed repos & organize training sessions

WP6 is advancing towards creating **a comprehensive knowledge repository** and an extensive range of training materials. These resources aim to support stakeholders across various sectors, providing valuable educational content and standardized materials for the application of AI/ML techniques in different fields. The initial task involves researching topics to include in the knowledge repository: general ML and AI topics, task-driven use cases, best practices, and specialized techniques related to data handling and big data management. Additionally, SSP Cloud will host various "cards" for each topic, **making the information easily accessible**. The second task is dedicated to compiling an inventory of ready-to-use training materials. Early results show a collection including about 20 files (slides, PDFs, and scripts), 11 online courses hosted by NSIs, and 13 courses and books hosted by other institutions. These materials primarily **focus on introducing data science and ML concepts**, Python and R programming, and best practices for ML workflows.
Looking ahead, other tasks involves listing additional training materials, gradual transfer of data to the repository and training sessions.

WP7 - EARTH OBSERVATION MODELS: TIME TO EXECUTE

Listed foundation models to use

Models refinement for practical application

Define requirements & strengthen collab.

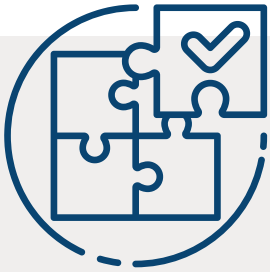
Among the team's achievements, an **inventory of Earth observation models** within the consortium was successfully compiled, laying a strong foundation for future developments. A particularly productive in-person meeting in Paris saw team

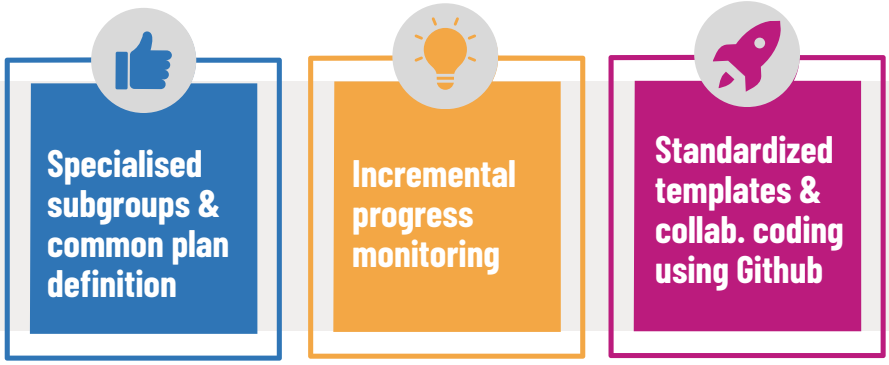
members actively engaged in discussions, sharing insights and refining goals for the project. The team has selected two models to continue research efforts: the **land cover model from IGN**, France, and the **crop type model from GUS**, Poland. Both models hold significant promise for advancing Earth observation capabilities, and the team is now focused on refining them for use in future applications.
Future activities will be based on challenges regarding, the **preparation phases, model execution plan and define requirements** for storage, processing, and performance for Copernicus to fulfil.
Among other challenges is that of strengthening collaboration with other working groups, in order to improve the expected results and in sharing with the whole consortium.

WP8 - DATA EDITING: A CONTINUOUS COLLABORATION & FUTURE DIRECTIONS

WP8 initiative continues its primary mission to harness AI/ML for **automation, efficiency, and quality improvements in data editing**. Over the past year, several key milestones have been achieved, including **a comprehensive literature review**, conducted in collaboration with WP9. The initiative has also fostered **a strong knowledge-sharing** culture through monthly use-case presentations, where different countries present their experiences and findings. These sessions have provided a platform for exchanging best practices and challenges, enriching the overall understanding of ML applications in data editing.
Another notable achievement is the progress made in developing **a collaborative research paper**.
Looking ahead, several priorities have been identified for the coming months. The team **aims to develop a standardized template** for collecting detailed information about use-cases, ensuring a structured approach to documenting experiences.
The WP8 team remains committed to advancing the role of ML in data editing, leveraging innovative solutions to enhance efficiency and data quality in statistical processes.

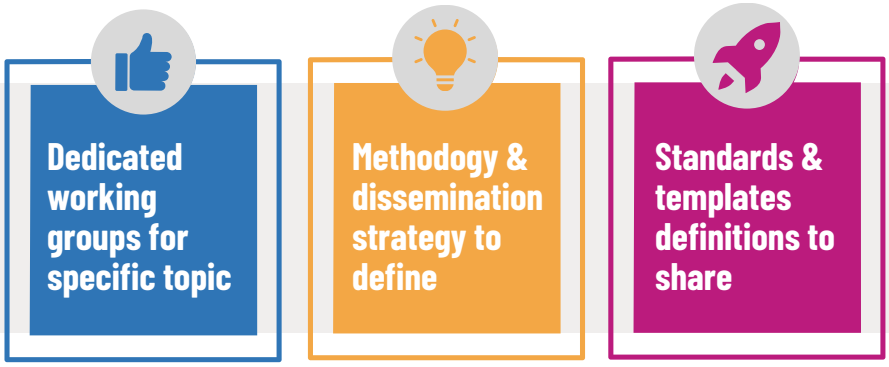
WP9 - PROGRESS AND CHALLENGES IN IMPUTATION RESEARCH





WP9 has successfully completed its first year **focused on imputation research**. With 9 participant countries and 3 observer nations, the project has effectively launched a structured routine of meetings, ensuring steady progress in its research agenda. A major achievement has been the formation of **3 specialized subgroups**, each dedicated to a specific area of imputation research. These include **early imputation, post-collection imputation and imputation beyond the sample.** In addition, the project has outlined a common plan of work, focusing on 3 key stages: **Exploratory Data Analysis** (EDA), the development and **evaluation of machine learning models**, and a **robust quality assessment framework** covering both statistical products and production processes.
Looking ahead, WP9 has identified 3 challenges: creation of **a standardized template** describing project updates, develop an efficient method to gather information from various projects and **launch a GitHub team** reinforcing the importance of collaborative coding.
In addition to these challenges, WP9 is working on methods for assessing input data quality and strategies for determining model retraining.

WP10 - TEXT CLASSIFICATION: STRATEGIES TO COVER USE CASES



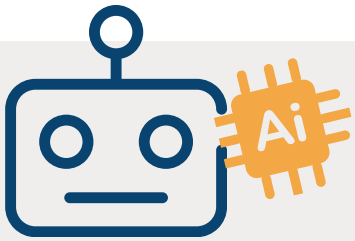
WP10 aims to enhance **text classification techniques** using advanced machine learning methods and address challenges associated with training data and model deployment. The project is structured into **3 key tasks**: a literature review and

project overview, methodological investigations and implementation, and the final dissemination of the results. The literature review provides the foundation for future developments. To manage the various points of view of the methodological investigations in text classification, it was chosen to organize the team into **5 different dedicated clusters**: Cluster 1, tackling data gaps by **generating synthetic training material** in multiple languages; Cluster 2, enhancing natural language understanding using **RAG LLM and Transformer** models; Cluster 3, incorporating **hierarchical structures into classification models**; Cluster 4, ensuring robust deployment and maintenance of classification models through **quality assurance measures**; Cluster 5, adapting to **evolving classification system revisions**, particularly in the NACE framework.
Key deliverables include a working classification codebase in Python and/or R, as well as a comprehensive report documenting methodologies and recommendations.

WP11 – SUPPLY CHAIN NETWORK: DATASETS & PIPELINE THE FUTURE WORK

WP11, focused on **reconstructing supply chain networks** (SCN) using advanced modeling techniques, achieved important successes in both the **creation of essential datasets** and the development of **a comprehensive pipeline** for training and validating models, laying a strong foundation for future work.
One of the important dataset is based on data from Portuguese companies that serves as a critical step towards building more accurate and detailed representations of SCN. Furthermore, the team has designed a robust data structure and modeling approach, including **a dedicated environment on SSP-cloud and GitHub**.
The next project phase will focus on **build the pipeline, training multiple models, and validating their accuracy**. The team is also preparing to apply the trained models to other National Statistical Institutes (NSIs), allowing for a broader application of the developed methodology.
In June 2025, **the team will present their findings** at the Supply Chain Satellite during the **Global Network Science Society** conference.

WP12 - GENERATIVE AI: DOMINATE THE CONSTANT CHANGE



- Defined the use cases to work with Generative AI
- Counteract the speed of evolution of the topic
- Cross collab. with NSIs to guarantee knowledge & skills

WP12 has successfully mapped out common issues and areas of **focus in Generative AI**, ensuring that efforts align with the most pressing needs in the field. **Examples from ongoing work** within National Statistical Institutes (NSIs) have been instrumental in highlighting the core themes that the team must concentrate on moving forward. One of the key point is the realization that can harness the rapid advancements in Generative AI to its advantage. Additionally, the team has developed **a conceptual framework** outlining how various deliverables will interconnect, laying a solid foundation for future work.

Among the main challenges facing **the group are keeping up with the knowledge and skills needed** to address such a constantly evolving topic, and mitigating issues related to technical constraints within the AI lab, as some constraints may require specific directions for **prototype development**.

To overcome these hurdles, **strong collaboration with the various NSIs** is necessary, ensuring participation in the initiative.

WP13 - SYNTHETIC DATA: STRENGTHEN THE USEFULNESS

- Framawork to categorize synthetic data purposes
- Finalize methodologies & proof its utility
- Running experiments in statistical domains

WP13 made significant progress in **defining the role of synthetic data in official statistics, selecting datasets** for experimental use, and **exploring data generation methods**. The group presented its work at key forums, such as the User Group and Expert Group on Statistical Disclosure Control (UG and EG SDC), establishing the relevance of synthetic data in the statistical community.

A framework was developed to categorize synthetic data purposes, including: structural datasets for software testing, public-use files for educational purposes, scientific use files for research, realistic "twins" for analysis and ML and secure microdata sharing aligned with privacy expectations. **Several methods explored** about synthetic data generation, including statistical, rule-based, and machine learning methods like Generative Adversarial Networks (GANs). It also **assessed privacy-preserving frameworks** such as Differential Privacy and methods like Synthpop. The group also **began categorizing privacy risk assessment methods**, including attack models and risk metrics.
In the next months, WP13 **plans to finalize synthetic data methodologies**, map statistical domains to methods, and run proof-of-concept experiments, ensuring research based on data's privacy without compromising its utility.

AIML4OS @ NTTS 2025 CONFERENCE

Special session (Bruxelles 11-13 March 2025)

The **Special Session** chaired by **Mauro Bruno** brought together experts to discuss key advancements of AIML4OS. **Eimear Crowley** (Coordinator of the project) presented Year One achievements outlining the project's organization, goals and future strategies. A key part of the presentation was a case study on the use of generative AI presented by **Jakob Engdahl** (Gen. AI WP Leader), showing the direction that AI must take in order to be integrated into official statistical contexts. Communication also plays a pivotal role in the project: **Orietta Luzi** (Istat's Methodology Director) presented how is possible to interact between several NSIs sharing the diverse activities carried out by the team.

WATCH THE REPLY
(time start: 14:33:25)

Subscribe Newsletter

To stay informed about the latest developments of the project, please subscribe to the newsletter

AIML4OS WEBSITE

FOLLOW US ON LINKEDIN