AIML4OS

Artificial Intelligence and Machine Learning for Official Statistics

Newsletter 2

Welcome to our second newsletter for ESSnet AIML4OS! These updates will share the project's progress, results, events, and other key news highlighting the progress, the results achieved, the event and all the news about this project

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

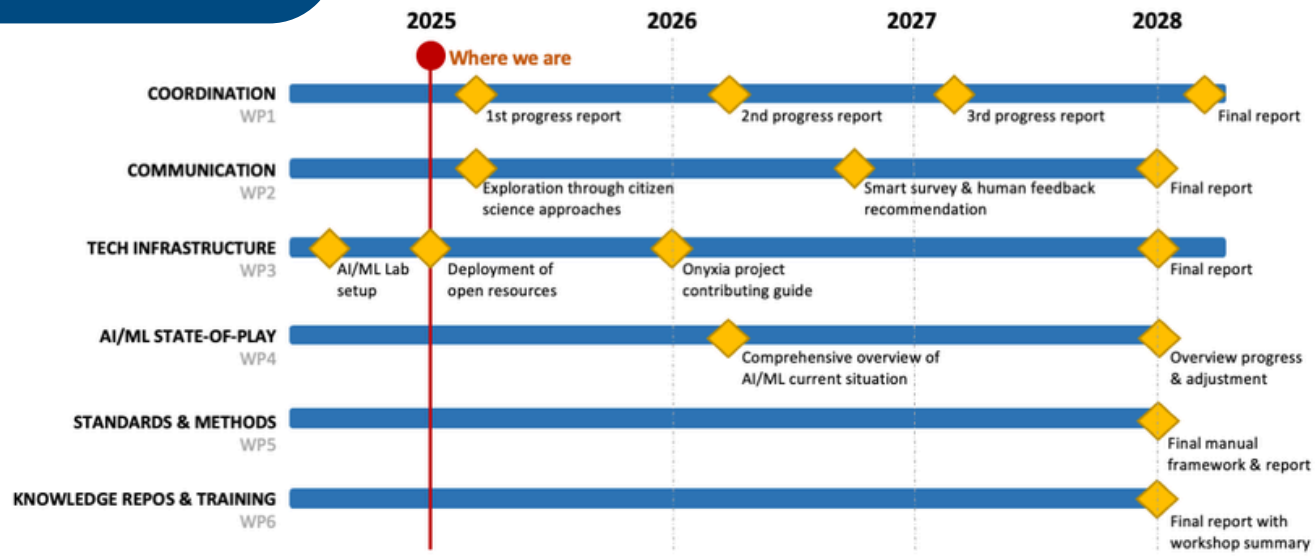
In this issue the state of play of different Work Packages, Datalab webinar, NTTS official programme

READ MORE...

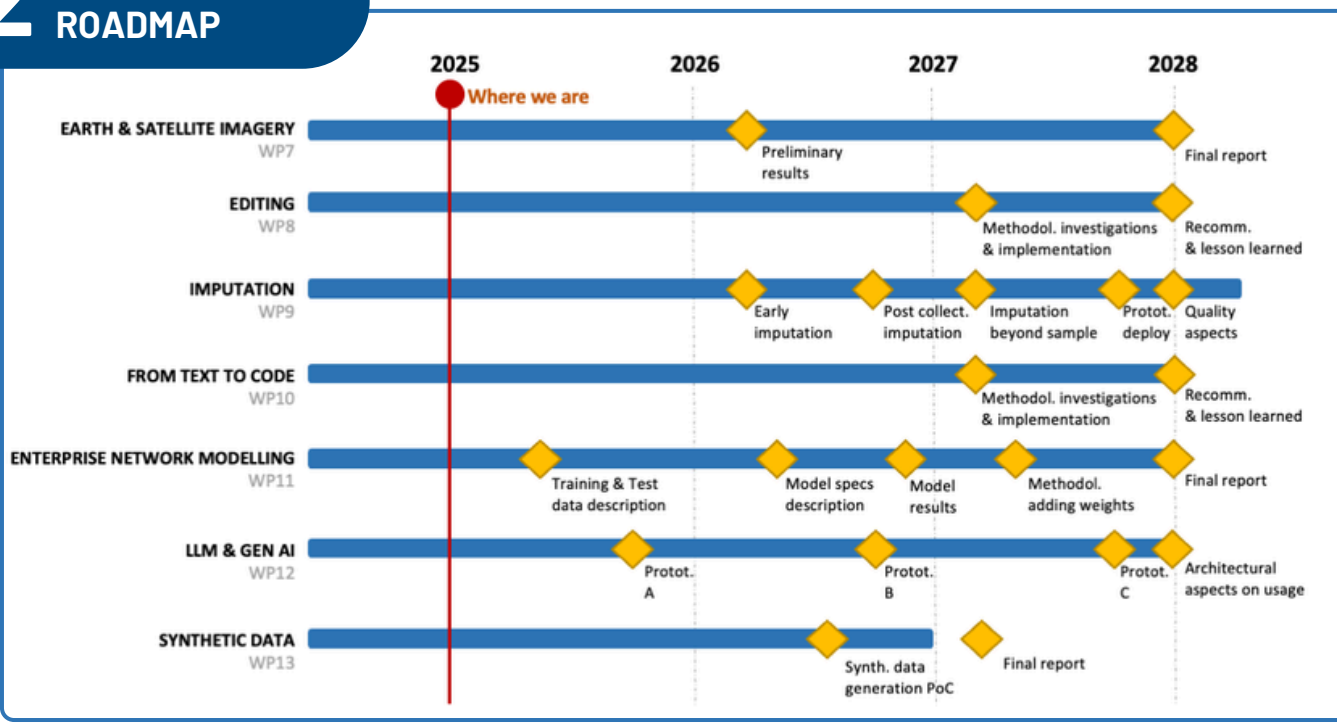
PROJECT OVERVIEW

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



AIML4OS STATE OF PLAY

Curious about our progress?
Dive into the latest updates from WPs!

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

Members of WP3 gathered in Oslo, hosted by Statistics Norway. There were productive discussions about Onyxia, the software powering the SSP Cloud, a computing solution for the ESS-Net.

The agenda covered **data processing solutions** across participating National Statistical Institutes and included a demonstration of the SSP Cloud's real-world applications. In collaboration with Eurostat's Shared Tools Expert Group, participants explored Onyxia's agnostic and flexible design. SSB showcased Dapla, a Kubernetes-based platform using Onyxia for secure statistical data processing, while Mercator Ocean presented its use of Onyxia and Kubernetes for global ocean modeling.

The event also featured a hackathon, where developers from SSB, Insee, and Statistics Austria worked on Onyxia improvements, such as **catalog filtering and local installations**. This in-person meeting deepened WP3 collaboration and understanding of Onyxia's potential, relevant for official statistics and beyond.
WP3 will continue monthly virtual meetings, with the next face-to-face meeting in Vienna planned for 2025.

Team Members
WP3 - AIML4OS

WP7 - AI/ML on earth observation data, satellite imagery

In November, WP7 had their first and successful physical meeting in Paris. The main goal was to select which models use and to make a high-level plan on how to proceed. Everybody was very involved, enthusiastic and the atmosphere was very good. The WP Leader was very happy with the team and confident that will accomplish their tasks.

The research question is: can existing AI/ML models based on earth observation be generalised over space (countries) and time (periods) and under what conditions? During the inventory phase (task 7.1), **we gathered the earth observation models** within our group (25 models were identified) and selected 6 models we considered feasible for our research. For these models, more information was collected and templates were filled in.

During the meeting, all 6 models were presented and discussed, and it has been selected 2 models to proceed with. The **first model is about land cover** and is developed by IGN, France. The **second model is about crop type**, developed by GUS, Poland. On the second day, a high-level plan on how to proceed until 2026 was defined. Furthermore, it was defined homework to gather necessary information to use these models and questions. It was planned to use the Copernicus Data Space Ecosystem (CDSE), in combination with the Onyxia platform of INSEE, France where necessary.

Team Members
WP7 - AIML4OS

WP8/9 - Editing focus & Imputation focus

The WP8/WP9 meeting in Vienna highlighted the transformative potential of machine learning (ML) in statistical editing while addressing key implementation challenges. As is well known, ML offers automation, accuracy improvements, and **reduced manual workloads** but faces significant hurdles in methodological, computational, and organizational domains.

Methodological discussions focused on model retraining schedules, data integration, and evaluation metrics that blend traditional and ML approaches. The adaptation of ML to specific statistical needs, such as univariate or multivariate editing, along with effective feature engineering, was also emphasized. Clear links between ML predictions and statistical products, coupled with uncertainty measurement, remain critical for success.

Computational challenges center on resource needs, scalable infrastructure, and the transition from testing to production. Financial constraints, expert shortages, and concerns over open-source tools complicate adoption. Meanwhile, **organizations** face resistance to change, limited expertise, and siloed production systems, all of which hinder ML implementation and scaling.

Despite these obstacles, participants acknowledged ML's potential to standardize processes and reduce manual effort. Moving forward, they prioritized centralized documentation, modular coding practices, and fostering collaboration to ensure transparency and efficiency in statistical editing workflows.

Moreover, the team of WP9 has divided its research and development **strategy into three sub-groups**: one about **early imputation**: focused on finding ML solutions for those situations in which we shall impute values which will be known and measured after some process step has been completed (e.g. after collection or after editing during collection). This track mainly affects to timeliness. The second about **post collection imputation**: focused on finding ML solutions for those situations which are typical in both item and unit non-response in usual production conditions. It covers imputation of values that are related to units in the sample (thus, having auxiliary and target information to aid the imputation). This track mainly affects to accuracy. The last one about **imputation beyond the sample**: focused on finding ML solutions for those situations in which we shall impute population frames or datasets (e.g. administrative sources), unavailable in sample. This focuses on reconstructing information for the whole population from smaller datasets. This track mainly affects to granularity.

WP11 - Applying ML for estimating firm-level supply chain networks

WP11 uses Machine Learning to train models on good quality datasets (e.g. based on rich administrative data), in order to allow National Statistical Institutes in countries that don't have such administrative data to reconstruct supply chain network datasets with a basic quality. One of the ongoing activities is building the **supply chain network dataset from administrative data** in Portugal. That will form the basis for creating training, test and validation sets. A first complete dataset will become available at the end of the first quarter of 2025. Due to the sensitive nature of the data, it can only be accessed on-site in Portugal. **A synthetic dataset has been created** for technical preparation and software testing, and will be available for use by other WP.

Furthermore, the modeling approach has been determined: work has started on **setting up the software pipeline** for experimenting, training and validating. An important issue is to make sure that both the training pipeline as well as the resulting network reconstruction models are transferable to

datasets from other countries. First intermediate results should become available in the course of 2025, with comparisons between various trained models as well as comparisons of reconstructed networks for different countries.
The WP is a good showcase for illustrating the potential of AI/ML for creating completely new kinds of relevant output for statistics. The project was recently presented at the webinar from the **UN AI and Data Science Sprint for Economic Statistics**, (to see the **registration**) and to get it on the agenda of the December meeting of the Business Statistics Directors Group, where supply chain networks recently became an item on the Research agenda.

WP13 – Synthetic Data

As minimising the risk of disclosure is paramount in the generation of synthetic data, the leaders of WP13 (Generation of Synthetic Data) presented the **current status and next steps to the ESS Expert Group** on Statistical Disclosure Control (EG SDC). It was agreed that there will be regular communication between the EG SDC and WP13. This will be facilitated by a personal overlap and reporting back from WP13 to the group towards the end of the first quarter of 2025. This reporting is likely to take the form of a short meeting.

AIML4OS DATALAB PRESENTATION

AIML4OS Datalab Presentation

Missed the webinar? Don't worry! Unlock the insights and revisit the action with the **webinar recording**, now available for you!
The AI/ML Datalab provides a comprehensive resources designed to facilitate the development of use cases across the work packages. It offers a rich suite of services for seamless data access and processing. Built on Onyxia, an open-source project launched by Insee, the Datalab empowers organizations to establish state-of-the-art data science platforms leveraging cloud technology. The Datalab is open to all European National Statistical Institutes (NSIs) and Other National Authorities (ONAs) via the European Datalab portal.

AIML4OS AT NTTS 2025 CONFERENCE

See the official programme (Bruxelles 11-13 March 2025)

Good news! The NTTS programme is available. Immerse yourself in a world of innovation, knowledge and collaboration with the newly released NTTS programme. Explore cutting-edge sessions, innovative research, and connect with experts who are shaping the future of statistics and data science. Don't miss your chance - check out the programme now, register and start planning your NTTS experience.

Subscribe Newsletter

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

AIML4OS WEBSITE

FOLLOW US ON LINKEDIN