AIML4OS

Artificial Intelligence and Machine Learning for Official Statistics

Newsletter 5

Welcome to our Newsletter #5 for ESSnet AIML4OS! Here you will find project updates highlighting progress, achievements, events and all 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: citizen-in-the-loop deliverable, transformer-based text classification, text-to-code collaboration, and network reconstruction progress.

READ MORE...

AIML4OS

NEW REPORT: "CITIZEN-IN-THE-LOOP IN AI/ML" NOW AVAILABLE

Citizen feedback to boost AI accuracy and trust

A new deliverable titled "**Citizen-in-the-Loop in AI/ML**" has just been published and is now available.

Developed as part of **Communication and Community Engagement** group, this report explores how involving citizens directly in the design and development of **AI/ML systems** can enhance model accuracy and reliability.

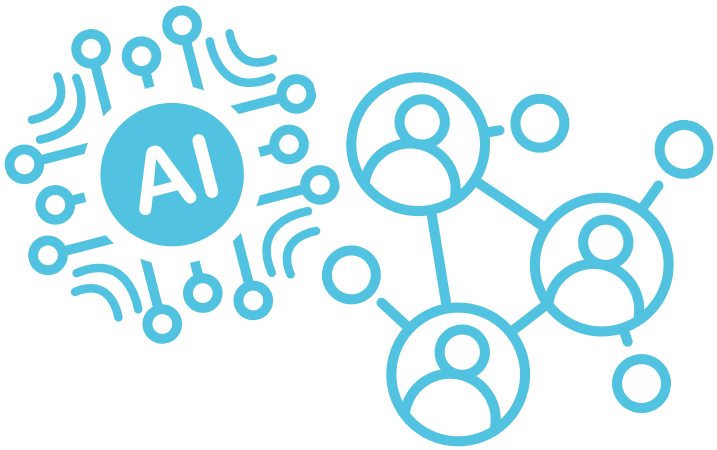
The deliverable investigates the potential of recruiting "citizens" or survey participants—acting as humans-in-the-loop—to contribute to key stages of AI/ML workflows. It identifies strategies for sampling and engaging citizens to support three main tasks:

- **Feature selection**
- **Creation of training data**
- **Updating of pre-trained models.**

From both statistical and methodological perspectives, the report examines the challenges of **concept and feature drift**, **annotator clustering**, **feature sufficiency**, as well as **annotator bias** and drop-out. It also outlines a plan for simulations and small-scale qualitative studies across three case studies to further explore these dynamics.

This deliverable represents an important step toward **citizen participation in AI**, promoting transparency, inclusivity, and collective intelligence in the next generation of data-driven systems.

Read the full report



AIML4OS

EXPLORING HOW TRANSFORMERS CAN REDEFINE TEXT CLASSIFICATION IN OFFICIAL STATISTICS

A workshop on transformer models, data augmentation, and RAG pipelines

At the recent **STATEC Methodology Workshop**, we took a deep dive into one of the most transformative areas of modern Artificial Intelligence: **transformer models** and their potential to revolutionize **text classification in official statistics**.
The session explored how advanced AI methods can enhance the way statistical institutions process, categorize, and interpret large volumes of textual information. From the early stages of setting up a **modern**, **cloud-based workspace** within the **SSP Cloud**, to leveraging **Synthetic Data** for improving model robustness, participants engaged in hands-on experiments that connected cutting-edge technology with real-world statistical challenges.

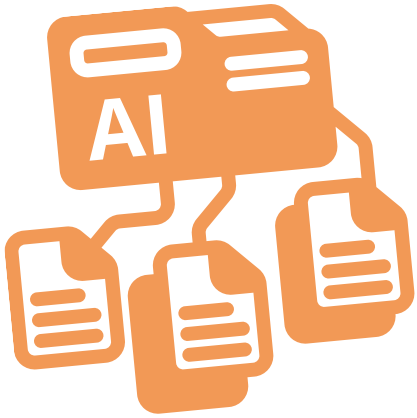
Key highlights included:

- **Data augmentation** techniques to expand and enrich training datasets.
- **Fine-tuning transformer models** using the Hugging Face framework for more precise classification.
- Testing a **Retrieval-Augmented Generation (RAG)** pipeline to enhance model understanding and improve classification accuracy.

But beyond the technical achievements, the workshop underscored a broader point: AI isn't just about technology—it's about empowering the production of official statistics with tools that are smarter, faster, and more transparent.

To support further experimentation, all materials are openly available:

- **Presentation & slides**: available here
- **Ready-to-use notebooks** on SSP Data Lab:
  - Data Augumentation
  - Fine-Tuning a Transformers Model,
  - RAG Classification
- **Source code**: Github link



AIML4OS

IMPROVING STANDARDISED TEXT CLASSIFICATION

European experts collaborate on “Text-to-code” - Advancing AI for Official Statistics

Experts from **8 National Statistical Institutes** (NSIs) gathered for the **first workshop of “Text to Code” project**, led by **Statistisches Bundesamt** (Destatis).
This initiative focuses on one of the central challenges in official statistics: how to **classify free-text information**—such as job titles or household expenses—into **standardised codes like ISCO, COICOP, or NACE**. Given the diversity of languages, formats, and data structures across Europe, there is no single solution that fits all contexts. This makes **cross-country collaboration** essential.
To tackle this, the team has organised its work into **5 thematic clusters**, each exploring different approaches. Among them, several are **experimenting with LLMs to generate synthetic training data** and enhance the accuracy and scalability of text classification.
During the workshop, Germany presented an overview of ongoing activities across participating NSIs, while the clusters shared **literature reviews**, methodological insights, and progress updates. A remark came from our colleagues at the **Instituto Nacional de Estadística** (INE), who showcased a first draft of their LLM-based pipeline for generating additional training data — a promising step toward more automated and efficient coding systems.
Once again, this workshop demonstrated the **power of European cooperation** in advancing innovation and trust in official statistics through AI and machine learning.

Team Members WP12 AIML4OS

ADVANCES IN NETWORK RECONSTRUCTION MODELS

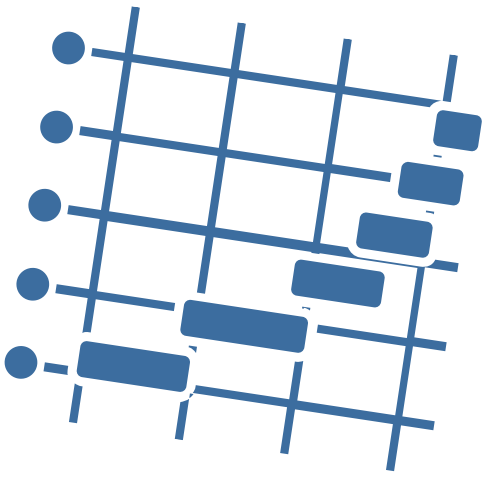
Developing more realistic, scalable network models for firm-level supply chain analysis

The **Supply Chain team is moving forward** with several upcoming **training sessions for network reconstruction models**, marking an important step in developing firm-level supply chain analytics within the project.

These models will be trained using the **Portuguese firm-level supply chain network dataset**, created earlier this year. Due to the confidential nature of the data, the training will take place **on-site in Portugal**, led by our **Polish colleagues**. The process will rely on the **software pipeline** that has been carefully developed over the past months.

In addition, the team has designed a **new approach to assign weights** to the links in reconstructed networks, enhancing the realism and analytical value of the results. In the coming months, **experiments** will begin to implement and test this approach using information from the Portuguese dataset.
Work is also underway on building a **separate software pipeline** to apply the reconstruction models to other countries, using **basic NSI data** as input. **Ireland and the Netherlands** will be the first to implement this pipeline, with results to be **validated and analyzed** to guide future model improvements.

Finally, the team has established **contact with the OECD**'s LIFT initiative (**Leveraging Inter-Firm Transactions**), creating opportunities to **exchange insights and experiences** on how firm-level network data can be used to inform **evidence-based policy**.

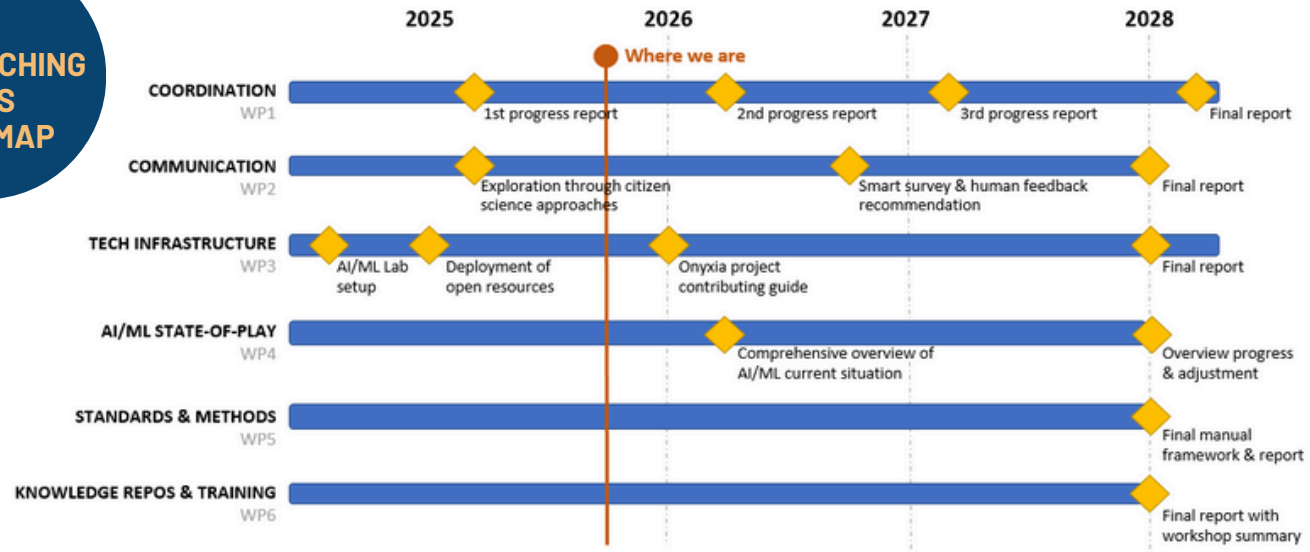


AIML4OS

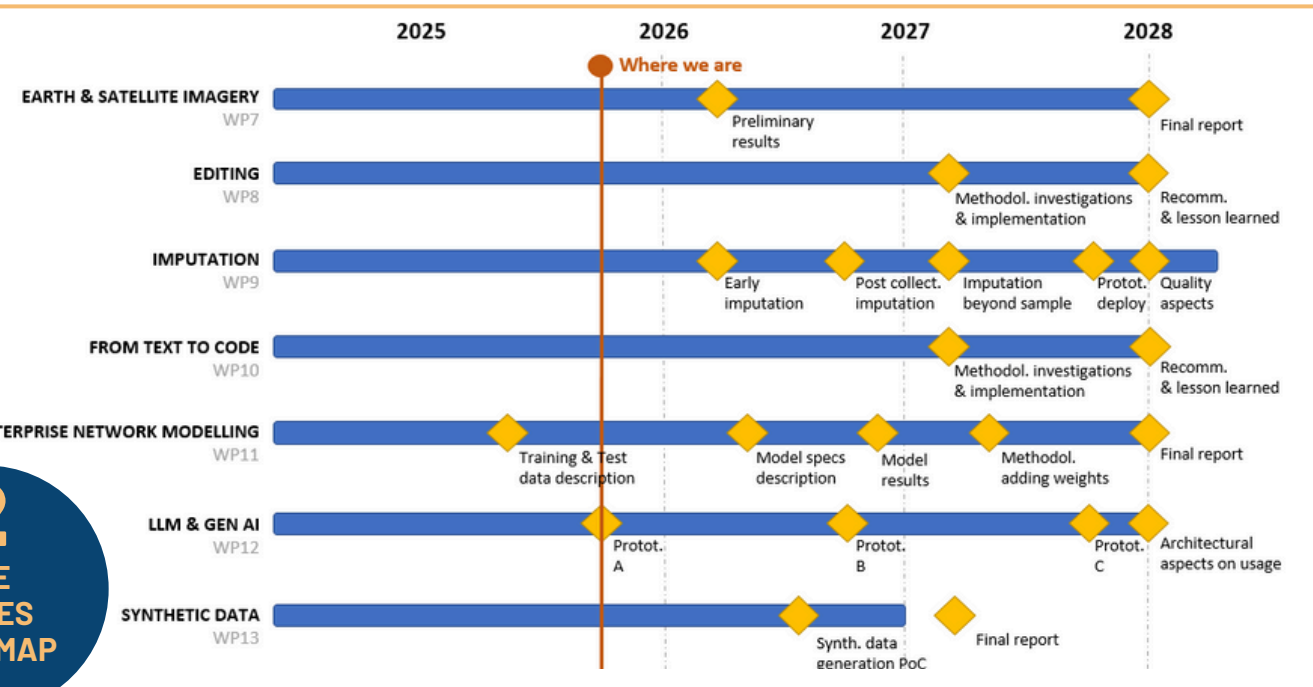
For each WP involved, the project is divided into several phases. Below are descriptions of what will be achieved and when.

PROJECT OVERVIEW

1 OVERARCHING WPS ROADMAP



2 USE CASES ROADMAP



Subscribe Newsletter

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

AIML4OS WEBSITE

FOLLOW US ON LINKEDIN