UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna

The European One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS): WP8 Use Case focused on data editing

Steffen Moritz, Katja Bürk, Florian Dumpert
Federal Statistical Office of Germany
steffen.moritz@destatis.de

I. Introduction

1. The European One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS) is a collaborative project aimed at advancing the use Artificial Intelligence/Machine Learning (AI/ML) in official statistics (1). Work Package 8 titled **"Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on editing"** plays a crucial role in this initiative by focusing on how AI/ML can enhance data editing and imputation processes within statistical offices. This paper introduces WP8, highlighting its objectives and significance.

2. WP8 aims to explore the potential of AI/ML technologies in improving automation, efficiency, and the quality of data editing. By developing statistically valid and efficient editing methods, WP8 addresses the critical role that accurate data plays in official statistics, without compromising on quality standards.

3. Furthermore, WP8 will also foster collaboration among the project partners, enabling the sharing of knowledge and expertise in AI-driven data editing solutions. By learning from each other's experiences, statistical offices can adopt best practices and accelerate the development of effective tools for data editing and imputation.

4. If successful, this collaborative approach may lead to joint solutions that can be applied across the European Statistical System. The knowledge exchange between partners will ensure that WP8 not only improves individual processes but contributes to a broader, collective advancement in the use of AI/ML in official statistics.

II. One-stop-shop (1SS) for AI-ML for Official Statistics

A. Background and Project Goals

5. The European One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS) is an ESSnet collaborative project involving 16 European countries. The four-year project started in April 2024 and will last 4 years.

6. The main objective of AIML4OS is to explore the use of AI/ML for the production of official statistics and to implement innovative solutions for statistical products and processes.

7. The expected outcomes of the AIML4OS project focus on driving progress through practical solutions and collaboration, including:

(a) the delivery of a framework for developing AI/ML solutions to be used in the context of official and European statistics
(b) the provision of access for ESS staff and partners to established and proven AI/ML solutions/resources to be leveraged in the context of official statistics production
(c) the encouragement of engagement from ESS organisations with AI/ML for innovation purposes and to facilitate their understanding and realisation of the benefits of AI/ML
(d) the delivery of economies of scale and resources through cooperation inside and outside the ESS and the acceleration from ideas regarding AI/ML to actual production.

8. Similar to other ESSnet initiatives, the AIML4OS project promotes coordinated efforts to facilitate the exchange of experiences, identify best practices, and encourage the reuse of solutions across the European statistical system. The project presents an opportunity for significant improvements in the quality and efficiency of official statistics, delivering more accurate and timely data.

B. Involved Countries

9. On April 2nd and 3rd 2024, the kick-off meeting of the AIML4OS project was held at the Federal Statistical Office of Germany in Wiesbaden. During the meeting, the work package coordinators presented the planning activities and participants got to know about each other's plans.

10. These are the countries participating in the project:

COUNTRY	CODE	NSI
AUSTRIA	AT	Statistics Austria
CYPRUS	CY	Statistical Service of Cyprus
DENMARK	DK	Statistics Denmark
FRANCE	FR	Institut National De La Statistique Et Des Etudes Economiques
GERMANY	DE	Statistisches Bundesamt
IRELAND	IE	Central Statistics Office
ITALY	IT	National Institute Of Statistics
LUXEMBURG	LU	Statistiques Luxembourg
NETHERLANDS	NL	Statistics Netherlands
NORWAY	NO	Statistics Norway
POLAND	PL	Statistics Poland
PORTUGAL	PT	Instituto Nacional De Estatistica
SLOVENIA	SI	Statistical Office of the Republic of Slovenia
SPAIN	ES	Instituto Nacional de Estadistica
SWEDEN	SE	Statistics Sweden
SWITZERLAND	CH	Swiss Federal Statistical Office

Table 1: Participating countries AIML4OS project

C. Organization of work

11. The project is organized into 13 work packages (WPs), with 6 focused on supporting implementation and 7 dedicated to specific use cases. WP8 plays a key role in generating valuable knowledge and solutions by disseminating data editing use cases from participants. Furthermore, WP8 actively contributes to WP5, which is

centred on setting standards, and WP6, which focuses on creating a knowledge repository and developing training materials (2).

12. The interaction between the work packages is illustrated in the following figure:

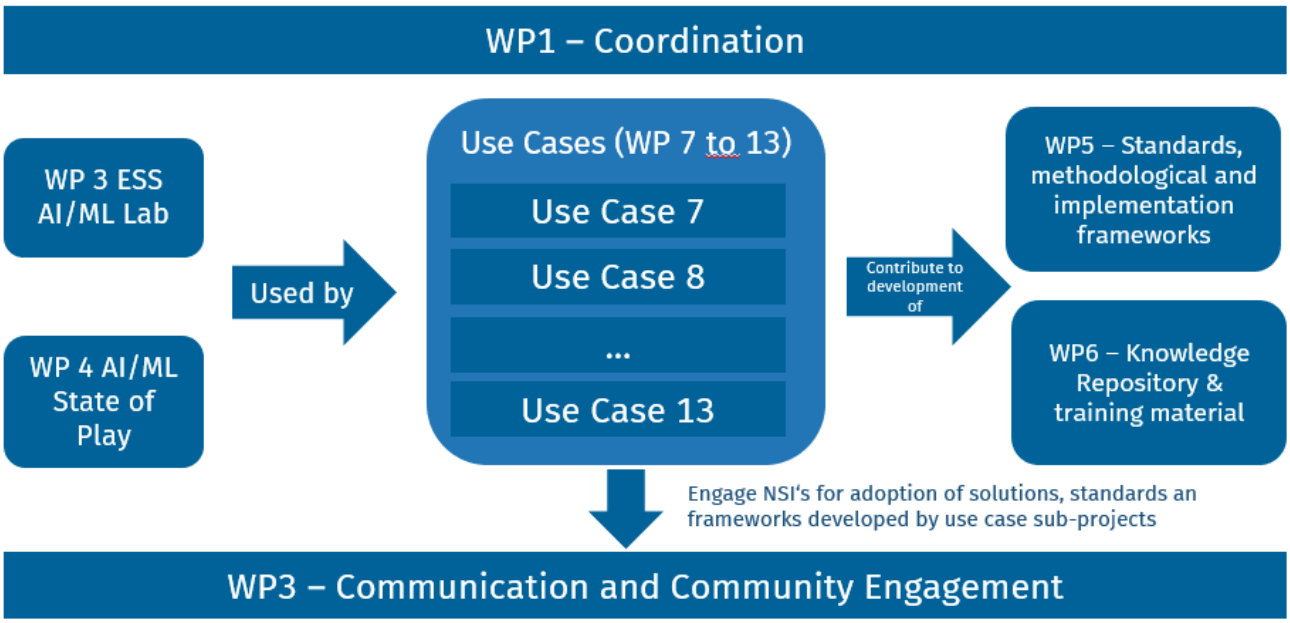

Figure 1: Interaction between work packages in the AIML4OS project

D. Work packages

13. The AIML4OS project consists of a comprehensive list of work packages, with interaction and collaboration between them being crucial for success.

14. Particularly important for WP8 are the close ties with WP9, which also focuses on data editing, but with a specialized emphasis on imputation. Use cases dedicated solely to imputation are typically handled by WP9, while WP8 mostly addresses use cases that combine both error detection and error correction. WP8 also has several use cases focused exclusively on error detection / error localization, such as anomaly detection.

15. This collaborative approach between ensures that the project maximizes synergies and benefits from shared expertise across work packages.

16. Below is a list with short descriptions of all work packages,

Workpackage	Title
WP1	Co-ordination - Organisational setup
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
WP7	AI/Ml on earth observation data, satellite imagery

WP8	Editing focus - 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 Potentials of the Use of AI/ML for Classifying and Coding
WP11	ML for enterprise network modelling Population study European Business Register
WP12	Text mining and NLP Open-source large language models (like ChatGPT)
WP13	Synthetic data SDC issues introduced by using AI/ML

Table 2: List and description of work packages in the AIML4OS project

III. WP 8: Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on editing

A. Goals and Background

17. The main mission of WP8 is to explore the potential of AI/ML to automate, enhance efficiency, and improve the quality of the data editing and imputation processes, with a special focus on data editing. By applying these technologies to practical examples, WP8 aims to strike a balance between innovation and maintaining the high-quality standards required in official statistics.

18. Data Editing deals with issues that are essential for the quality of official statistics (3) (related to WP5). This involves reliable information for traditional products as well as the production of good training data sets for machine learning. At the same time, machine learning can help to improve data editing (4). On the one hand, by enabling the use of methods that can recognise erroneous observations as such and, on the other hand, for the specific localisation of the error within an observation. The replacement of incorrect values also touches on the customisation of this work package and establishes the connection to WP9.

19. A standardised editing pipeline for all datasets and variable types would be of great benefit to official statistics, but seems hardly feasible. Different procedures and approaches must therefore be developed and analysed for different types of datasets, considering the fact that data is probably not shareable between different NSIs. Standard procedures can then be developed and implemented for specific constellations (reference to WP5). The interplay between development, testing and standardisation leads to an iterative process that will characterise the work of this work package.

20. The expected outputs of this work package will include methodological investigations and practical implementations. At best, this will lead to the development of a comprehensive methodological and implementation framework for data editing. This framework would be a key outcome of the work package, significantly improving the quality of official statistics.

B. Involved Countries

21. Not every country participating in AIML4OS is also involved in WP8. There are 11 countries contributing to WP8, each with varying levels of commitment.

22. However, in the early stages, every country can provide valuable input by sharing their current situation, challenges, and successful experiences. In the later phases, not every country will contribute its own use case, as the focus will shift towards more specific implementations.

23. These are the WP 8 participants and their level of involvement:

GERMANY	Statistisches Bundesamt	Lead
AUSTRIA	Statistics Austria	Beneficiary
ITALY	National Institute Of Statistics	Beneficiary
NETHERLANDS	Statistics Netherlands	Beneficiary
NORWAY	Statistics Norway	Beneficiary
PORTUGAL	Instituto Nacional De Estatistica	Beneficiary
CYPRUS	Statistical Service of Cyprus	Associate
DENMARK	Statistics Denmark	Associate
SPAIN	Instituto Nacional de Estadistica	Associate
FRANCE	Institut National De La Statistique Et Des Etudes Economiques	Associate
IRELAND	Central Statistics Office	Associate
EUROPE	European Statistical Office	Observer

Table 3: Participating countries in WP8

IV. Planned Work and Current Status

A. Planned Tasks and Deliverables

24. The four-year project duration offers the opportunity to address complex challenges, but it also demands flexibility and openness to innovation. As additional experiences are made and technology advances, some parts of the project might need to adapting to shifts. This is why the tasks are not specified overly detailed at the outset, focusing instead on broader goals rather than a rigid to-do list, allowing space for adjustments as needed throughout the project.

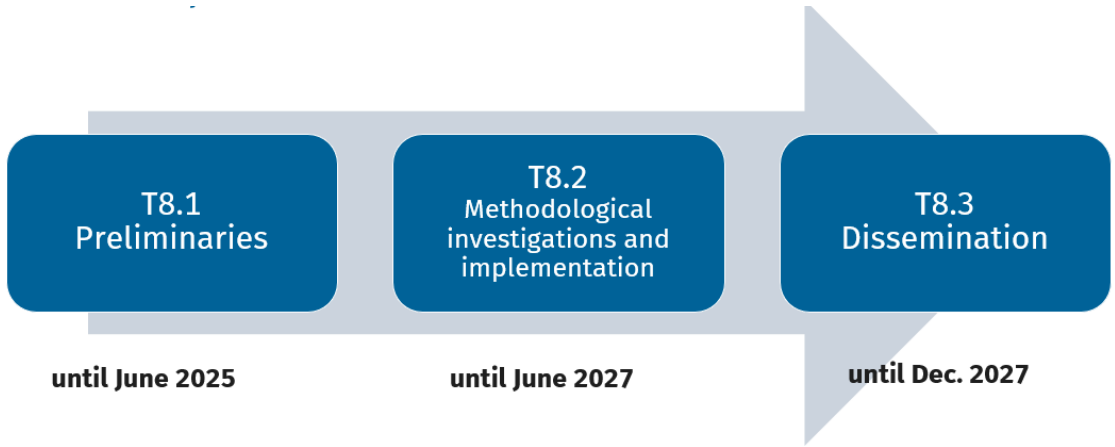

Figure 2: Planned tasks in work package 8

25. The first task T8.1 contains preliminary work, including a literature review and a systematic overview on the current experiences and challenges in the participating countries. Based on this internal and external view, the decision will be made which problem(s) to tackle from a methodological and an implementation point of view (technique, grade of automation, data, etc).

26. The second task T8.2 contains the methodological investigations and implementations of several approaches based on the decisions made at the end of T8.1, in which different solutions (usually in forms of prototypes) will be assessed.

27. The third task T 8.3 includes the final report and participation in dissemination events such as conferences and workshops, in particular the UNECE Data Editing Workshops/Expert Meetings 2024 and 2026, both virtual and physical with the aim of disseminating the outcomes and the lessons learned.

28. The project defines two key deliverables for WP 8 that represent the main objectives to be achieved, ensuring both practical outcomes and shared knowledge. These deliverables are designed to provide significant value to the National Statistical Institutes (NSIs) by fostering collaboration and improving data processes across the board.

(a) D8.1 - Methodological investigations and implementation
- Planned for April 2027
- Results will be public
- Delivered as prototypes / code: Share Python and/or R code for editing; Code and report on methods' assessment

(b) D8.2 - Recommendations and lessons learned
- Planned for
- Results will be public
- Delivered as document: Overall final report on the whole approach, the assessment of the results, the impact on processes, the quality and the costs – as well as the lessons learned from this WP

B. Short-term goals

29. In the short term, WP8 will focus on conducting a literature review to assess the state-of-the-art in data editing. This will highlight current methodologies and innovations, particularly regarding the integration of AI and machine learning in official statistics.

30. Simultaneously, WP8 will gather insights from National Statistical Institutes (NSIs) to understand their specific challenges and experiences. By creating a systematic overview, the project will identify common issues and best practices across the European statistical system.

31. These efforts will guide decisions on key methodological and implementation challenges. The project aims to present these findings in a paper, titled "Assessing Data Editing in European Official Statistics: Challenges, Future Directions, and the Role of Machine Learning," expected by mid-2025.

V. First insights into planned use cases

32. A list of use cases from participating countries has already been assembled. These countries aim to test innovative methods in data editing and imputation through these use cases. The cases will serve as practical applications for exploring advanced techniques and improving statistical processes.

33. Although this list is temporary, as the project is still in the exploratory phase, countries are actively communicating to determine what will work best and how everyone can benefit most. Some use cases may not be viable in the early stages and could be replaced as the project progresses. Therefore, the list can quickly become outdated as adjustments are made.

34. Below is the current list of use cases, including information showing the area or statistic where the method should be applied and the current status of each project. Keep in mind that the list may evolve as the project progresses and use cases are refined or replaced.

Country	Data to be edited	Project status
Ireland	Census data	In development
Italy	Administrative education data	In development
	Labour Force Survey	Idea stage
Norway	Business survey on R&D	In development
	Customs data	In development
Austria	Survey on travel behaviour	Beginning of project
Germany	Survey on structural business statistics	In development
Netherlands	Survey on Regional Employment Statistics	Beginning of project
Portugal	Employment Statistics	In development
France	Administrative data on wage and employment	Completed
	Business short-term statistics	Idea stage

Table 4: Planned use cases work package 8

35. As shown below, most use cases are currently focused on error detection and localization. This is largely due to WP9's emphasis on error correction, specifically through imputation. Additionally, many use cases address multiple phases of the data editing process, reflecting an integrated approach that encompasses several interrelated steps.

Country	Phase of the editing process: Error ...		
	Identification	Localisation	Correction
Ireland	✓		✓
Italy	✓	✓	
	✓		✓
Norway	✓	✓	
	✓		
Austria	✓	✓	✓
Germany		✓	✓
Netherlands	✓	✓	
Portugal	✓	✓	
France	✓		
	✓		

Table 5: Phase of the editing process of the planned use cases

VI. Conclusions

36. WP8 is a valuable part of the AIM4OS project because it fosters collaboration across participating countries and institutions. By recognizing that many of the challenges faced in data editing and imputation are shared, WP8 enables participants to pool their expertise and insights. This collaborative approach ensures that solutions are developed with broader applicability in mind, benefiting the entire European statistical system and driving progress in data processing practices.

37. Another key benefit of WP8 is its emphasis on code sharing. By encouraging participants to share the tools and methods they develop, the work package creates a more efficient environment where all countries can

learn from one another's efforts. This shared pool of knowledge and code minimizes duplication of work and enables faster adoption of best practices, ultimately leading to improved data quality across the board.

38. WP8 will try to consider the challenge of generalizability directly from the beginning. While it is already complex to create generalizable solutions within a single National Statistical Institute (NSI), developing frameworks that function across the European statistical system is extremely difficult. WP8's iterative approach, with its focus on testing, standardizing, and sharing results, tries to overcome these challenges.

39. To maximize its impact, WP8 is committed to producing outputs—such as reports and code—that are accessible and beneficial to others. By ensuring that the deliverables are clearly documented and aligned with the needs of different stakeholders, the work package facilitates the wider adoption of successful approaches. This focus on producing reusable, practical outputs strengthens the overall value of the project and ensures long-term benefits for participating countries.

40. In the most favourable outcome, once WP8 is fully developed, we can expect it to provide a robust framework for data editing that can be applied across various statistical systems. This would not only enhance collaboration and knowledge-sharing among participating countries but also streamline the use of AI/ML in official statistics. WP8's outputs—such as reusable code and standardized methods—can make it easier for countries to adopt efficient, high-quality data processes. In the future, this will help drive continuous improvement, ensuring more accurate, timely, and consistent statistics across Europe.

VI. References

(1) European Commission. (2024.). *AIML4OS: Artificial Intelligence and Machine Learning for Official Statistics*. CROS. https://cros.ec.europa.eu/dashboard/aiml4os

(2) Kay, F. (2024, April 3-5). *The AIML4OS project: A first overview* [Conference presentation]. Conference on Foundations and Advances of Machine Learning in Official Statistics, Destatis. https://www.destatis.de/EN/About-Us/Events/Machine-Learning/Slides/sp_kay.pdf

(3) UNECE. (2019). *Generic Statistical Data Editing Model, Version 2.0*. UNECE. https://unece.org/statistics/documents/2019/06/gsdem-v20

(4) ADSaMM Data Editing Task Team. (2024). *Organisational aspects of implementing ML-based data editing in statistical production*. UNECE. https://w3.unece.org/stories/2024/09/mlops_wp/