ercim-news.ercim.eu Number 140 January 2025

ERCIM NEWS

Special theme:

Large-Scale Data Analytics

Editorial Information

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 2,000 printed copies and is also available online ,at https://ercim-news@ercim.eu.

ERCIM News is published by ERCIM EEIG
BP 93, F-06902 Sophia Antipolis Cedex, France
+33 4 9238 5010, contact@ercim.eu
Director: Dominique Hazaël-Massieux, ISSN 0926-4981

Contributions
Contributions should be submitted to the local editor of your country

Advertising
For current advertising rates and conditions, see https://ercim-news.ercim.eu/ or contact peter.kunz@ercim.eu

ERCIM News online edition: *https://ercim-news.ercim.eu/*

Next issue:
April 2025, Special theme: Cultural AI

Subscription
Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu

Cover photo: A picture of the night sky with a Pixel 4a in 'astrophotography' mode. See article by Olivier Parisot on page 29.

JOINT ERCIM ACTIONS

SPECIAL THEME

Introduction to the Special Theme

RESEARCH AND INNOVATION

ANNOUNCEMENTS / IN BRIEF

NEXT ISSUE

ERCIM News 141, April 2025
Special theme: Cultural AI

Call for contributions: https://ercim-news.ercim.eu/call

This issue was supported by the European project GLACIATION (see article on page 24).

Promoting Diversity and Inclusion in Computer Science Academia

Interview with María-del-Mar Gallardo, ITIS Software, University of Málaga

• Dear María, what is your role in your organization?

I am currently a full professor of Computer Science at the Institute for Software Engineering and Software Technology of the University of Malaga (ITIS) (Spain) (itis.uma.es). The institute comprises around 140 researchers and technicians, including permanent staff (approximately 60%) and trainees. Its activities are carried out in five research areas: Automatic Software Engineering, Data Science and Artificial Intelligence, Cybersecurity, Intelligent Networks and Services, and Applications. At the Institute, I co-lead the Morse group, which focuses on contributions in the areas of mobile communication networks and formal methods.

I have been teaching and researching different aspects of Software Engineering for more than 30 years. During these years, I have witnessed first-hand the predominance of men in the computer science profession. Specifically, women have been and continue to be a minority at universities and research centres. It is well known that the number of female professionals in computer science is alarmingly low, and although in recent years the presence of women seems to be growing, it is doing so very slowly.

I have carried out various management tasks at the University of Malaga. Specifically, for two years I chaired the commission for university accreditation in the area of computer science at the Spanish Agency for Quality Assessment and Accreditation (ANECA). In my opinion, while being aware of the limitations of the procedure that logically has to be gradually improved, the mere existence of a national agency that evaluates the competence of all Spanish researchers and teachers is an important step to guarantee access to university positions on equal terms for all applicants without deviations due to their gender.

• Why do you think it is important to promote inclusion and diversity in research institutes and universities?

From my point of view, policies should be formulated to promote inclusion and diversity in all areas of research. In the specific case of computer science, this promotion is essential because there are many economic interests related to computer technologies in which women are not participating as main actors in reasonable numbers. If work teams were more diverse, surely the areas of research and development would produce different contributions, making them in turn more attractive for all sectors of society.

What we are seeing now is only a slice of the pie, because the teams are mostly made up of male personnel. Work teams with more diversity would lead to more inclusive technologies, which benefit society as a whole. Consequently, yes, I think we have to continue working on promoting more diverse and inclusive work teams.

• Can you briefly explain some initiatives in which your organization has started to promote diversity and inclusion?
Any initiative that you are particularly proud of?

For about five years, I have been part of the coordination team of the "Como Tú" ("Like You") project at the University of Malaga (comotu.uma.es). About 150 researchers and professionals from the STEM fields participate in the project, which aims to make the role of women in science and technology careers visible to school-aged girls and boys (from 5 to 17 years old). For this purpose, the project participants go to interested schools to carry out practical workshops or give talks so that girls and boys have a direct contact with professional women in the different areas of science, technology, engineering and maths. The final objective of the project is to overcome gender stereotypes that, in many cases, discourage girls from following their vocation in science or technology.

The ITIS Institute is very involved in diversity actions and gender issues. In particular, the Institute contributes financially to the project, and most of the Institute's researchers participate in the activities described.

Depending on the age of the children to whom the activity is directed, participatory workshops are prepared in which students are challenged to use their analytical intelligence to programme more or less complex devices by applying algorithms they know or have developed themselves.

The experience with this type of initiative is very positive. The surveys carried out after the workshops show that, in general, students rate the activities very positively and are interested in learning more about technology.

• Have you faced any challenges in promoting inclusion and diversity?

No, in general, schools are aware of the problem and are increasingly including the participation of the "Como Tú" project in their annual activities. On the other hand, the experience of the researchers and professionals is so good that practically all of them repeat every year, preparing and carrying out the workshops in different primary and secondary schools in the province of Malaga.

The challenge is to obtain more funding to be able to reach more schools, including provinces bordering Malaga.

• Is there any "mistake" that is important to avoid?

I think that the main mistake when promoting diversity is creating the impression that it is about giving preferential treatment to women or other groups. Some men may think, and in fact do, that some women take advantage of equality policies.

Once you have reached university as a student or researcher, promotion should be the same for everyone. An effort should be made to establish mechanisms to avoid negative discrimination. This is obviously not at odds with the idea of giving visibility to women's work in professions that are mostly dominated by men. If this is not done, the stereotype that certain professions, such as IT, are for men will be perpetuated.

For this reason, the "Como tú" project attempts to address the problem from the time children are small, showing that neither science nor technology have a gender, that they are not "male activities".

However, when dealing with children, you have to be very careful not to segregate by gender to make sure they all feel equal. This way, all children participate in the same way in the activities and workshops carried out within the framework of the project, although the ultimate objective is to highlight that the instructors are female researchers.

• Are there lessons learned or best practices that you would like to share with other organizations that want to work on these themes?

I believe that any project to promote diversity must be implemented taking into account all groups involved. Male researchers should not feel attacked by policies; on the contrary, we should be able to convince everyone that it is good to have diversity in teams, and that diversity enriches us as researchers and as individuals.

A question that male researchers ask me from time to time is why science and technology need to be promoted among female students. They say that girls should be free to do what they want, they should not be "forced" to choose a technological career. I think that the issue cannot be addressed from that perspective. Of course girls must be free to choose the profession they want. The problem lies in the context in which that decision is made, and if it is linked to a stereotyped idea of what it means to be a researcher. Many years ago we would have been surprised to see a female police officer, a female judge, a female bus driver, or a female doctor, for instance. These were traditionally male professions. With a little effort and luck, in a few years' time it will be normal to see female computer scientists working in research and technological development.

The interview was conducted by Monica Divitini of NTNU, chair of the ERCIM Human Capital Task Group.

Please contact:
María-del-Mar Gallardo, ITIS Software, University of Málaga
mdgallardo@uma.es

Beyond Compliance 2024: Research Ethics in the Digital Age

by Anaëlle Martin (French National Advisory Ethics Council for Health and Life Sciences)

After Paris in 2022 [L1] followed by Porto in 2023 [L3], the third edition of the ERCIM Forum 'Beyond Compliance' was held in Budapest on 14-15 October 2024, at the HUN-REN Institute for Computer Science and Control [L3]. This year's event, which took place both in person and online, continued the discussion on the tough ethical issues faced by researchers in digital sciences. The scientific richness of these two days lay not only in the distinguished status of the speakers, but also in the wide range of cutting-edge topics covered. The diversity of contributions and the high calibre of Forum participants made it possible to explore digital issues from cultural, legal, (geo)political, historical, philosophical, and ethical perspectives.

The programme of the first day was marked by two particularly brilliant keynotes, masterfully delivered by Julian Nida-Rümelin ("Beyond Compliance: Digital Humanism") and Milad Doueihi ("Beyond Intelligence: Imaginative Computing"). While the first speaker focused on tracing the philosophical origins of Digital Humanism and describing its challenges through animism and mechanistic reductionism, the second one offered a historical and literary analysis of what we now refer to as thinking machines. These presentations revisited classic AI debates, drawing on the ideas of Turing, Gödel, Wittgenstein, and earlier thinkers such as Leibniz and Butler. Both speakers explored the intersection of humanity and digital technology, advocating for human-centered approach to AI. The German philosopher emphasized the centrality of human authorship, while the American historian discussed the transformative effects of digital memory on culture and knowledge. Ethically, both thinkers stressed the importance of responsibility in the use of technology, emphasizing that education should guide digital transformation. They both called for critical reflection to safeguard cultural values and advocated for the preservation of human relationships, while reflecting on how digital culture reshapes knowledge transmission.

The first session dedicated to the making of regulations featured three researchers. Firstly, Melodena Stephens discussed the complexities of AI regulation, emphasising the difficulty of implementing effective, intergenerational policies in a rapidly evolving technological landscape, and the need for a global, flexible, and ethically sound approach to address issues like human autonomy, security, and the future of jobs. Next, Anna Ujlaki critically reviews the political theory discourse on AI, focusing on its conceptual limitations, normative questions, and potential for addressing AI's integration into society, while highlighting the political risks and ethical dilemmas involved in AI regulation. Finally, Nikolaus Forgo discussed how, since the introduction of computers into public administration, lawmakers have re-

peatedly overestimated the short-term effects of new technologies while underestimating their long-term impacts, exemplified by the development of data protection laws and the recent AI Act.

The rest of the day featured two additional sessions dedicated to emerging topics and cultural influences.

Anatole Lécuyer opened the emerging topics session by discussing the paradoxical effects of virtual reality and metaverse technologies, highlighting their history and their growing impact on the population, particularly children and young adults, and the emerging ethical questions surrounding them. He explored psychological effects such as the sense of embodiment, agency, and the Proteus effect, which leads users to behave according to the stereotypes of their avatars, while also examining the potential harms and benefits of VR, from therapeutic uses to the risk of altering identity. This fascinating discussion was extended by the following speakers, who were present in person: Michele Barbier and Ferran Argelaguet. They presented a project exploring the ethical challenges of social interactions in the metaverse, focusing on issues such as harassment, privacy, and the legal status of avatars, with the goal of fostering empathy, improving safety tools, and addressing social and cultural concerns around digital identities and regulation. Finally, and in a slightly unconventional style, Jean-Bernard Stefani discussed the concept of "conviviality" from Illich to highlight the moral dilemmas in the digital world, including its ecological impact, surveillance capitalism, algorithmic discrimination, and digital divides, while arguing that these issues require a critical approach and a shift towards more human-centered and de-automated technologies.

Finally, the last two remote speakers addressed the issue of cultural influences. Rockwell Clancy discussed the relationship between cultural responsiveness, psychological realism, and global AI ethics, highlighting the importance of understanding both the normative and empirical components of AI ethics, the challenges posed by cross-cultural contexts, and the need for culturally informed policy frameworks in AI development. Marianna Capasso presented a project on algorithmic discrimination, approaching it from a cross-cultural perspective. She highlighted how algorithmic discrimination should be understood in a nuanced way, using examples such as Amazon's CV screening system, which discriminated against women due to biased historical training data. She examined various forms of algorithmic discrimination, including indirect and statistical discrimination, and explored how culturally specific norms influence discriminatory behaviours.

The second day began with a session on cooperative agents. Elias Fernández Domingos discussed the importance of studying delegation to AI, explaining its issues and presenting a behavioral experiment where AI delegation improved coordination in a collective risk scenario, emphasizing the need for well-designed systems that maintain human agency while delegating tasks. Rebecca Stower explored ethical and psychological implications of human-robot interactions, focusing on errors in robot behaviour, the impact on trust and risk-taking, and the challenges of balancing data privacy and user preferences in robot design. Finally, Michael Fisher discussed the importance of ensuring trustworthiness in autonomous systems, emphasizing the need for reliability, transparency, and ethical decision-making, while also addressing sustainability concerns related to both the environmental impact of AI and robotics, as well as the unnecessary deployment of technology.

At midday, the Forum participants had the opportunity to attend the Tutorial Training expertly delivered by Alexei Grinbaum. He emphasized the importance of operationalizing AI ethics and explained that ethics in AI should be viewed as a valuable framework rather than a constraint. The scientist addressed a range of ethical challenges, including security risks in robotics, and introduced tools to facilitate discussions between ethicists and engineers. He presented training courses featuring exercises on dilemmas and the evaluation of AI projects in sectors like healthcare. He also explored the issue of responsibility in personalised education, focusing on topics such as bias, fairness, and the role of teachers.

For the first time, the Forum left some space for an unconference session which allowed participants to discuss, in a more informal way, Open Science and Nobel Prize in Computer Science.

Finally, the Forum concluded with a session dedicated to democracy that gave the floor to four speakers. Natali Helberger argued that AI is a powerful political tool that can either strengthen or undermine democracy, highlighting concerns about misinformation and the influence of big tech, while also recognizing AI's potential to enhance communication. Siddharth Peter de Souza discussed the creation of data governance norms, emphasizing the role of civil society and advocating for a pluralistic approach to regulation that includes marginalized voices. Attila Gyulai explored the impact of AI on democracy, questioning the assumption that democracy is solely about autonomy, and suggesting that a more realistic understanding of democracy, which accounts for representation, manipulation, and the constructed nature of preferences, is necessary to address the challenges AI poses. Finally, Bjorn Kleizen examined the level of trust citizens have in AI systems used by governments, exploring how transparency and public perceptions influence trust, and emphasizing the need for long-term strategies to maintain trust in AI applications.

Links:

[L1] https://www.ercim.eu/beyond-compliance/beyond-compliance-2022

[L2] https://www.ercim.eu/beyond-compliance/beyond-compliance-2023

[L3] https://www.ercim.eu/beyond-compliance

Please contact:

Anaëlle Martin
Comité consultatif national d'éthique, France
anaelle.martin@ccne.fr

ERCIM "Alain Bensoussan" Fellowship Programme

The ERCIM Postdoctoral Fellowship Programme is one of ERCIM's principal activities. The programme is open to young researchers worldwide and focuses on a broad range of fields in computer science and mathematics.

> "The ERCIM fellowship has been a pivotal experience in my professional growth. It provided me the freedom to explore new ideas and pursue innovative projects, fostering an environment of creativity and learning. The connections I made with industry experts and fellow participants have been invaluable, opening doors to new opportunities and collaborations. I am grateful for the support and encouragement provided throughout the journey.
>
> **Vineeta JAIN**
> Former ERCIM Fellow

The fellowship helps promising scientists improve their knowledge of European research structures and networks, and gain insight into the working conditions of leading European research institutions. Fellowships last for 12 months (with a possible extension) and are spent at one of the ERCIM member institutes.

Where are fellows hosted?

Only ERCIM members may host fellows. When an ERCIM member is a consortium, the hosting institute may be any of its constituent organisations. When an ERCIM member is a funding body, the hosting institute may be any of its affiliates. Fellowships are offered according to the needs of member institutes and the available funding. Fellows are appointed either via a stipend (agreement for a research training programme) or a working contract. The contract type and monthly allowance or salary depend on the hosting institute.

ERCIM encourages researchers from academic institutions and those in industry to apply.

Why apply for an ERCIM Fellowship?

The Fellowship Programme enables outstanding young scientists from around the globe to tackle challenging problems at Europe's leading research centres. It also helps to widen personal ties and deepen mutual understanding among scientists. Through the programme, ERCIM fellows can:

- Work with internationally recognised experts,
- Improve their knowledge of European research structures and networks,
- Familiarise themselves with working conditions in leading European research centres,
- Foster cross-fertilisation and cooperation between research groups .

Equal Opportunities

ERCIM is committed to ensuring equal opportunities and promoting diversity. Applicants for a fellowship within the ERCIM consortium are not discriminated against on the basis of race, colour, religion, gender, national origin, age, marital status or disability.

Conditions

Candidates must:

- Have obtained a PhD in the past eight years (before the application deadline), or be in the final year of doctoral study with an outstanding academic record. Proof of the PhD qualification will be required before the fellowship begins;
- Be fluent in English.

Application deadlines

Deadlines for applications are 31 March and 30 September each year.

Since its inception in 1991, more than 800 fellows have participated in the programme. In 2024, eight scientists began an ERCIM PhD fellowship; over the course of that year, 33 fellows were hosted. The Fellowship Programme is named in honour of Alain Bensoussan, the former president of Inria, one of ERCIM's three founding institutes.

http://fellowship.ercim.eu

ERCIM Fellows Community Event 2024

ERCIM organized an online community event for its postdoctoral fellows and guests for the fourth time on 8 November 2024.

Due to significant demand and interest, ERCIM has decided to host its community event once again, marking the fourth edition of this initiative since its inception during the pandemic in 2021. The core aim remains consistent with the earlier events: fostering engagement and collaboration among ERCIM fellows. This latest edition broadened its scope to include not only fellows hosted between 2021 and 2024, their scientific coordinators, and representatives from ERCIM member organizations but also recommended external post-doctoral researchers. The virtual gathering brought together 40 participants, creating a dynamic platform for idea exchange and networking.

We designed the event to handle challenges, like having postdoctoral researchers from different countries who are at various stages in their fellowships. The aim was to keep the schedule short while still allowing 18 posters to be presented and ensuring the event was free for everyone.

The event commenced with a series of five engaging presentations in the keynote room. The main highlight of the event was the lively poster sessions spread across three virtual rooms on the "Gather Town" platform.

The event received overwhelmingly positive feedback, with participants valuing the networking opportunities and expressing enthusiasm for future events.

Introduction to the Special Theme

Large-Scale Data Analysis – Software Infrastructure and Application Domains

by Andras Benczur (HUN-REN SZTAKI) and Dominik Ślęzak (University of Warsaw)

Large-scale data analytics empowers organizations to harness the full potential of the vast amounts of data they generate and collect. By driving innovation, enhancing business operations, personalizing customer experiences, and improving risk management, insights derived from large-scale data analytics are critical for gaining a competitive advantage and making informed, data-driven decisions. With the exponential growth of data generated by businesses, consumers, and connected devices, it is essential to address key challenges in handling Big Data, processing real-time information, and enabling timely, actionable insights.

The ERCIM News Special Theme on Large-Scale Data Analytics focuses on two main areas. On one hand, it highlights cutting-edge techniques such as machine learning, predictive modeling, and advanced analytics methods. On the other hand, it explores applications across diverse sectors, industries, and societal challenges.

Articles on Big Data Infrastructure and Technologies delve into topics such as distributed data processing, the edge-cloud continuum, federated data analysis, and the integration of heterogeneous data sources. A special focus is given to data analytics for Open Science. From a technological perspective, machine vision and spatial data analysis emerge as key tools in several domains. Articles on Big Data applications span various verticals, including healthcare, energy, transportation, robotics, finance, agri-food, environment, sustainability, and science. In many cases, critical issues of data governance, privacy, and security are addressed. These include techniques for anonymization and de-identification of large datasets, as well as ensuring transparency, reproducibility, and explainability in large-scale data systems.

Data infrastructure and ecosystems

The first part of the articles focuses on data infrastructure and ecosystems. Barba-González et al. present a semantic-driven workflow automation system for large-scale analytics, with applications in areas such as machine learning and e-Science (page 10). The large-scale analytics platform developed by Assante et al. facilitates collaboration and advances reproducible research by enabling researchers to share, reuse, and build upon each other's work across diverse scientific disciplines (page 11). Tzagkarakis et al. discuss the decision intelligence platform of the TwinODIS Horizon-Widera project, which combines AI and operations research to address challenges in large-scale, uncertain systems (page 13). In the Italian National Center for Sustainable Mobility project, Millitarì et al. developed a data mining platform designed for predictive maintenance in the railway sector (page 14). Sartzetakis and Chamanara describe the results of the DataBri-X project, which introduced a trustworthy AI platform that aligns with European values and ethical standards, emphasizing the transformation of data-sharing ecosystems (page 16).

Scientific data

Articles focusing on scientific data are closely linked to initiatives such as the European Open Science Cloud (EOSC), the AI4EU on-demand platform and ecosystem, and the EGI Foundation's infrastructure services. This area partly overlaps with developments in data infrastructure, as illustrated in the work of Assante et al. (page 11). Scientific digital twins are explored by Manzi et al., with an emphasis on reusing modular components. (page 18). The article describes use cases including flood impact modeling and early warning systems, cyclone projections, the Virgo Gravitational Wave Interferometer Noise simulations, and high-energy physics particle detector simulations. Sipos and Schaap discuss an AI-driven platform developed in the iMagine project, designed to analyze vast amounts of image data for aquatic sciences (page 19). This platform connects with the EOSC and AI4EU initiatives and seeks to expand its scope through open calls for additional use cases.

Application verticals

The application verticals in this Special Issue cover a wide range of domains. Articles focusing on data infrastructure and open science address areas such as science (Barba-González et al., page 10; Sipos and Schaap, page 19; Manzi et al., page 18), transportation (Millitarì et al., page 14), and the environment (Tzagkarakis et al., page 14; Manzi et al., page 18), among others.

Medwenitsch et al. discuss how advanced data analysis can transform agriculture in Austria's climate-stricken Seewinkel region (page 21). In the energy sector, Klikovits and Fabianek propose solutions to overcome challenges such as security, privacy, and GDPR compliance, which often impede data sharing, analysis, and interpretation (page 23). Similarly, Rotskos et al. present the outcomes of the Glaciation project, which focuses on scalable anomaly detection within the edge-cloud continuum for grid management (page 24). In the realm of sustainability, Suta et al. analyze information on sustainability extracted from large-scale European digital financial reports (page 26).

Additional verticals are explored in articles addressing computer vision and spatial data applications. Finally, medical and health applications form a distinct and dedicated section of this Special Issue, emphasizing their unique challenges and contributions.

Computer vision or spatial data

Several articles highlight computer vision or spatial data as their technological focus. For example, Hubner and Nausner describe the Multimodal Fusion Architecture for Sensor Applications, a robust system for real-time sensor integration that enhances situational awareness and provides precise decision support for railway security (page 28). Parisot explores resource-aware detection of satellite streaks in deep sky image streams, utilizing lightweight machine vision on edge devices (page 29). Bouchal et al. present GLayer, a GPU-accelerated software platform designed for the fast aggregation, filtering, and visualization of large-scale spatial data, particularly traffic data (page 31). García-Nieto et al. discuss their big data workflow for processing and analyzing Earth observation remote sensing satellite data, showcasing its potential for large-scale geospatial analysis (page page 33).

Medical and health applications

Four articles focus on medical and health applications. Pejo et al. propose a privacy-friendly contribution evaluation technique designed to address the growing issue of selfish incentives in the self-evaluation of medical records (page 35). Al-Radhi and Németh explore methods to translate brain activity into clear and intelligible speech, aiming to restore communication abilities for individuals with severe speech disorders (page 36). Koltai et al. employ anomaly detection techniques to reduce false alarms in telemonitoring medical devices, achieving this without centralizing sensitive patient data (page 38). Finally, Zimeras integrates diverse data sources to create visualizations for digital epidemiology, enhancing the analysis and understanding of health trends (page 39).

The Special Theme demonstrates the wide verticals of large-scale data analytics addressed in Europe. Developments work in close collaboration with European data spaces and open science platforms, often addressing reproducibility. The presented results emphasize European values of openness, fairness, protection of personal values and privacy.

References:

- Sakr, Sherif, and Albert Y. Zomaya, eds. Encyclopedia of big data technologies. Springer International Publishing, 2019.
- Schintler, Laurie A., and Connie L. McNeely, eds. Encyclopedia of big data. Springer International Publishing, 2022.

Please contact:

Andras Benczur
HUN-REN SZTAKI, Hungary
benczur@sztaki.hu

Dominik Ślęzak
University of Warsaw, Poland
slezak@mimuw.edu.pl

Semantic-Driven Workflow Automation for Large-Scale Data Analytics

by Cristóbal Barba-González, José F. Aldana-Montes, and Ismael Navas-Delgado (ITIS, University of Málaga)

This article presents TITAN, a platform designed to enable the creation and execution of Big Data analytics workflows. Using semantic technologies, TITAN ensures the integration, validation, and reusability of data-driven components, empowering researchers and industries to handle large-scale data challenges more effectively. Through real-world case studies, we demonstrate its potential in transforming data processing workflows across various domains.

Processing, analyzing, and deriving valuable insights from large datasets is more critical than ever in today's data-driven society. Thus, large-scale data management, processing, and value extraction are significant challenges for organisations due to the exponential expansion of data generated by connected devices, businesses, and customers. However, conventional analytics techniques frequently find handling Big Data's volume, speed, and complexity to be challenging. Therefore, creating platforms that offer adaptable, scalable, and semantically rich Big Data analytics solutions is crucial to overcoming these obstacles.

TITAN [1] is a platform designed to support the creation, management, and deployment of data science workflows for Big Data analytics. What sets TITAN apart from other platforms is its innovative use of semantic technologies, which enables the definition and orchestration of complex workflows. The core of TITAN is the BIGOWL ontology [2], which provides formal definitions to annotate all workflow components semantically. Furthermore, these semantics enable consistent and interpretable definitions of each component. Hence, this semantic layer ensures that workflows are correctly defined and facilitates automatic validation of component compatibility, which is crucial in large-scale environments.

TITAN's design focuses on making Big Data workflows accessible, flexible, and reproducible. Its architecture allows for the modularisation of data processing tasks [L1], such as data loading, transformation, analysis, and storage. Each task within a workflow can be seen as a self-contained component that can be reused across different workflows. Based on Docker containers, this modular approach benefits researchers, data scientists, and organisations that must adapt their analytics pipelines to diverse use cases without rebuilding components from scratch (Figure 1).

A key feature of TITAN is its ability to integrate and process data from multiple sources, enabling end-to-end analytics workflows that span data collection, processing, and visualization. Additionally, the platform supports popular tools and frameworks for Big Data processing, such as Apache Spark, Kafka, and TensorFlow, ensuring seamless integration with existing infrastructures. This capability is valuable for dealing with large volumes of real-time or streaming data.

Another benefit is TITAN's ability to track data lineage through its semantic model. Using RDF (Resource Description Framework) triples, TITAN provides a transparent record of data's origins, transformations, and final outputs. This transparency is critical in ensuring data quality and reproducibility, especially in complex analytical scenarios involving multiple stakeholders or regulatory requirements. This feature makes it easy to publish the workflow results as FAIR data.

TITAN was initially applied to three case studies across various domains to demonstrate its versatility and power:

1. Iris Flower Classification (Machine Learning): The first case study uses a classic machine learning problem—classifying different species of the Iris flower based on attributes such as petal and sepal length. TITAN enables the workflow to comprise several modular tasks, such as dataset loading, data splitting, model training, and validation. Each task is semantically annotated, ensuring the components are compatible and data flows seamlessly. By leveraging TITAN, building and validating a machine learning model becomes more streamlined and efficient.
2. Human Activity Recognition (Deep Learning): In this case, TITAN handles large-scale data generated by wearable devices (30 TB of accelerometer data). The workflow includes training deep learning models (such as ConvNet and LSTM) for activity classification. TITAN integrates with Apache Spark to manage the massive volume of data, while its semantic layer ensures that the components used in the workflow are correctly configured and compatible. This case highlights TITAN's ability to support real-time data processing, scalability, and complex model training in Big Data environments.

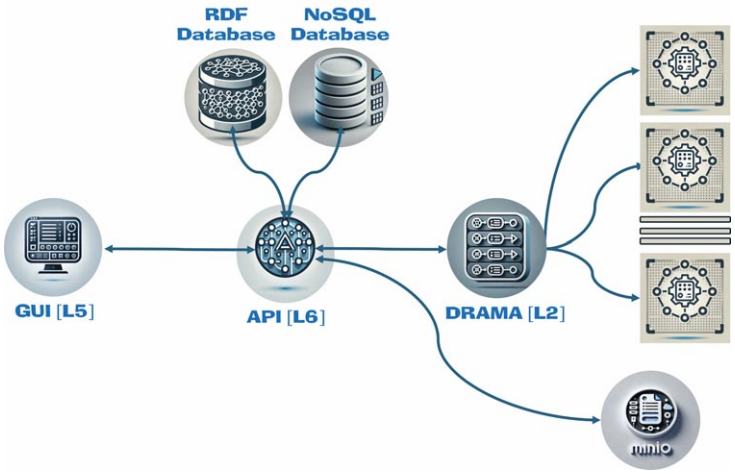


Figure 1: TITAN architecture. The users interact with a GUI that enables the creation of workflows using the available components in its RDF repository. Executing the workflows requires using the needed data and a set of workers to perform data analysis. This is orchestrated by DRAMA using a NoSQL database to track workflow execution and a MinIO distributed data storage to manage the input and output data of each component.

3. Automatic Monitoring of Earth Observation Satellite Images: The third case study involves classifying regions of interest in satellite images captured by the Sentinel-2 constellation. The workflow includes downloading satellite data, preprocessing images, training a support vector machine (SVM) model, and testing the model. Through TITAN, each step is defined as a task with precise inputs, outputs, and parameters, which are semantically validated. This case illustrates TITAN's potential for environmental monitoring, showcasing its flexibility in handling spatial data and complex analytical pipelines.

TITAN is an actively evolving tool successfully applied in various real-world scenarios, driving continuous improvement. Its core validation and evaluation framework, Drama [L2], has progressed into the more advanced DramaX [L3], enabling enhanced workflow support. This evolution has allowed TITAN to underpin the development of innovative infrastructures for scientific data analysis, including applications in environmental data processing [L4]. Notably, TITAN has facilitated the creation of impactful workflows for European environmentalists, such as a predictive model for pollen levels [3].

TITAN represents a significant step forward in Big Data analytics, offering a semantically enriched platform for building, deploying, and managing complex data workflows. TITAN makes creating scalable, reusable, and interoperable data pipelines easier through its modular architecture, semantic validation, and support for various data processing tools. Furthermore, TITAN ensures data quality and reproducibility by tracking data lineage through its semantic model. By supporting diverse case studies across different industries, TITAN demonstrates its versatility and potential to transform the way Big Data analytics are conducted, making it a valuable tool for researchers and industries.

Links:

[L1] https://github.com/KhaosResearch/TITAN-dockers
[L2] https://github.com/KhaosResearch/drama
[L3] https://github.com/KhaosResearch/dramaX
[L4] https://kwz.me/hFg
[L5] https://github.com/KhaosResearch/TITAN-GUI
[L6] https://github.com/KhaosResearch/TITAN-API

References:

[1] A. Benítez-Hidalgo, et al., "TITAN: A knowledge-based platform for Big Data workflow management," Knowledge-Based Systems, vol. 232, p. 107489, 2021, doi: 10.1016/j.knosys.2021.107489.
[2] C. Barba-González, et al., "BIGOWL: Knowledge centered Big Data analytics," Expert Systems with Applications, vol. 115, pp. 543-556, 2019, doi: 10.1016/j.eswa.2018.08.026.
[3] S. Hurtado, et al., "e-Science workflow: A semantic approach for airborne pollen prediction," Knowledge-Based Systems, vol. 284, p. 111230, 2024, doi: 10.1016/j.knosys.2023.111230.

Please contact:

Ismael Navas Delgado
ITIS Software, University of Málaga, Spain
ismael@uma.es

Empowering Collaborative and Reproducible Large-Scale Data Analytics with D4Science

by Massimiliano Assante (CNR-ISTI), Marco Lettere (Nubisware srl), Alfredo Oliviero (CNR-ISTI), and Pasquale Pagano (CNR-ISTI)

The D4Science platform is advancing reproducible research by providing scientists with robust, cloud-based tools for large-scale data analysis such as the Cloud Computing Platform (CCP). CCP enhances collaboration, allowing researchers to share, reuse, and build on each other's work across diverse scientific disciplines.

D4Science [1, 2] embraces the "as a Service" paradigm, offering Virtual Research Environments (VREs) [3] to streamline the research process, serving as a foundation for modern scientific collaboration, combining accessibility, innovation, and scalability within a single, cohesive framework. The VREs allow researchers to perform their data-driven research tasks without needing to manage the complexities of storage, computation, or deployment.

The Cloud Computing Platform (CCP) [L1], born from D4Science's more than 10 years of experience as an operational digital infrastructure, embodies the principles of FAIR (Findable, Accessible, Interoperable, and Reusable) data, advancing Open Science and reproducibility in research. At its core, CCP is designed to handle large-scale data analysis, promoting the widespread adoption of microservice-based architectures. This approach enhances the platform's flexibility and makes it highly interoperable and composable, enabling researchers to build and integrate their computational methods effortlessly. CCP incorporates several innovative features that make it particularly suited for data-intensive research. Its methods importer tool simplifies the integration of computational methods, allowing users to deploy custom algorithms and applications. The execution lifecycle tracker ensures that every step of a method's life, from creation to execution, is meticulously documented. Additionally, CCP includes a real-time execution monitor, which provides live feedback from the server logs during method execution, enabling users to track progress and identify issues promptly. Furthermore, CCP supports the archiving of executions, preserving the full configuration, parameters, and inputs of each execution. This capability ensures that methods can be repeated with fidelity in the future, enabling reproducibility and providing a robust framework for iterative research processes.

One of CCP's core strengths lies in its flexibility, enabling seamless integration across programming environments, languages, and execution infrastructures. This flexibility extends to the execution of methods by the adoption of standard REST/JSON APIs for interacting with CCP. Methods can be executed programmatically from web applications as shown in Figure 1, command-line interfaces, Jupyter Notebooks or Galaxy workflows. To further support users, automatic code

generators create stubs and templates for multiple languages and runtimes, including Python, Julia, Bash, Galaxy, and Notebooks.

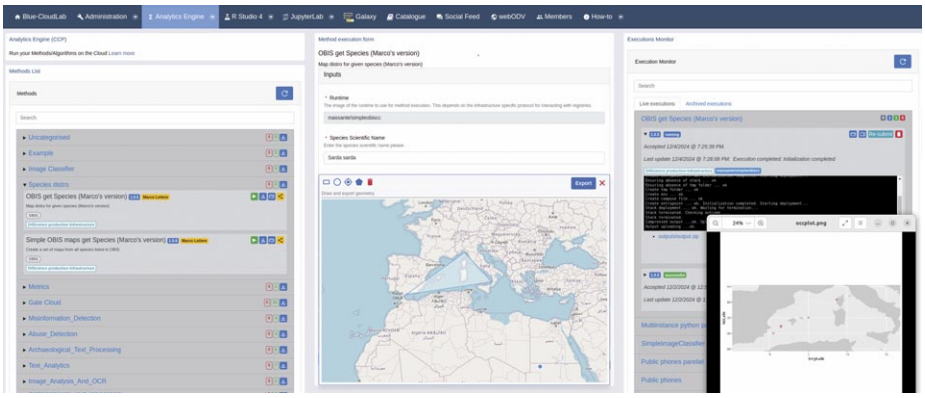


Figure 1 - The Cloud Computing Platform Web User Interface

CCP supports both containerised (e.g., Docker, Docker Swarm, LXD, Kubernetes, Singularity) and non-containerised infrastructures (e.g., Galaxy, Slurm). Execution infrastructures can be hosted on commercial platforms such as Google Cloud Platform, or non-commercial ones like D4Science's production environment. They also accommodate High-Performance Computing (HPC) clusters, as well as local environments such as personal laptops for experimental purposes.

The flexible and inherently distributed nature of CCP execution infrastructures allows users to design and run methods precisely aligning computational workloads with the most suitable resources thus fostering optimisation of execution environments based on data locality, computational demand and infrastructure capabilities. By accommodating a broad spectrum of execution contexts, from HPC systems to experimental environments, CCP facilitates distributed and scalable collaboration, enabling reproducibility and interoperability across diverse research domains.

D4Science offers a native execution infrastructure based on Docker swarm and enriched by a dedicated image registry based on Harbor. Especially when designing methods for such container-based infrastructures, scientists have virtually no limitations on what programming languages, environments, versions, or dependencies they are allowed to use. By encapsulating computational methods and their dependencies into isolated containers, CCP ensures reproducibility, portability, and scalability. This approach allows researchers to deploy methods without compatibility concerns, maintaining consistent execution environments regardless of underlying hardware or software variations. Additionally, the use of containers significantly reduces the overhead associated with traditional virtualisation technologies, thereby improving the efficiency of complex scientific workflows.

By integrating containerisation, flexible infrastructure management, and robust automation tools, CCP emerges as a critical enabler of large-scale distributed computing, driving innovation and reproducibility across research disciplines.

In addition to its technical features, CCP strongly aligns with the principles of Open Science, ensuring that all scientific outputs are transparent, repeatable, and reusable. Every execution within CCP is documented with comprehensive provenance tracking, which records the origins, transformations, and outcomes of data. This capability not only supports reproducibility but also provides a clear lineage for scientific discoveries, uplifting trust and attribution in research. This flexibility ensures that researchers across disciplines, regardless of their preferred tools, can adopt and benefit from the platform. Furthermore, the integration of dynamic resource allocation allows users to scale their computational resources based on demand, making CCP suitable for projects of varying sizes and complexities.

The platform's impact is already visible in large-scale scientific initiatives. For instance, the EOSC Blue-Cloud2026 project [L2] leverages CCP for ocean science research, utilising its robust infrastructure to perform collaborative data analytics on vast datasets. Similarly, the SoBigData Research Infrastructure [L3], focusing on social data mining and ethical Big Data analytics, integrates CCP to enable a multidisciplinary ecosystem for studying social phenomena. These examples highlight how CCP empowers researchers to address complex scientific questions by providing the tools necessary for effective collaboration, computation, and discovery.

By bridging the gap between technical complexity and scientific innovation, with its combination of containerisation, API-driven workflows, provenance management, and scalable infrastructure, CCP offers researchers a reliable and efficient environment for advancing their work. As scientific challenges grow in complexity, platforms like CCP will play an increasingly critical role in enabling reproducible and impactful research across disciplines.

Links:

[L1] https://ccp.cloud.d4science.org/docs/index.html
[L2] https://www.blue-cloud.org
[L3] http://www.sobigdata.eu

References:

[1] M. Assante, et al, "Enacting open science by D4Science," Future Gener. Comput. Syst., vol. 101, pp. 555–563, 2024, doi: 10.1016/j.future.2019.05.063.
[2] L. Candela, D. Castelli, and P. Pagano, "The D4Science Experience on Virtual Research Environment Development," Comput. Sci. Eng., vol. 25, no. 2, pp. 12–19, 2023, doi: 10.1109/MCSE.2023.3290433.
[3] M. Assante, et al., "Virtual research environments co-creation: The D4Science experience," Concurr. Comput. Pract. Exp., vol. 35, no. 18, pp. e6925:1–e6925:12, 2023, doi: 10.1002/cpe.6925.

Please contact:

Massimiliano Assante
CNR-ISTI, Italy
massimiliano.assante@cnr.it

Optimised Decision Intelligence in Data-intensive Environments

by George Tzagkarakis (FORTH-ICS), Rommert Dekker (EUR-DE), and Themis Palpanas (UPC- LIPADE)

Effective decision-making in large-scale, uncertain systems faces growing challenges in today's complex, data-rich environments. Traditional systems struggle to process vast datasets in real time while balancing conflicting objectives and ensuring fairness. The TwinODIS project introduces a transformative approach by combining Artificial Intelligence (AI) and Operations Research (OR) to create next-generation Decision Intelligence systems. This integration leverages advanced analytics, optimisation techniques, and AI-driven insights to transform large-scale decision-making, enabling sustainable development and economic growth through smarter, data-driven solutions.

Traditional decision support systems struggle to meet the demands of today's data-intensive and uncertain environments. While data provides valuable insights, the real challenge lies in transforming it into timely, impactful decisions. Existing systems often lack the agility and intelligence required to process vast, complex datasets and respond to dynamic conditions. As the need for autonomous, AI-driven decision-making grows, the European Union is investing heavily in AI and the data economy, with spending expected to surpass €70 billion by 2026. Regulatory frameworks like the AI Act [1] and Data Act [2] aim to establish Europe as a global leader in data-driven decision-making, fostering sustainable development and economic progress. The EU's "2030 Digital Compass" envisions resilient, agile societies built on intelligent systems that convert knowledge into action, driving innovation and workforce transformation.

The TwinODIS project [L1], funded by the EU under Horizon Europe HORIZON-WIDERA-2023-ACCESS-02-01 Twinning Programme, introduces advanced technologies to transform large-scale decision support systems in dynamic, data-rich environments. To address the challenge of managing complex big data, it will develop distributed storage systems and real-time predictive and prescriptive analytics to process vast volumes of data efficiently. For adaptive decision-making, TwinODIS combines AI techniques such as Reinforcement Learning with Operations Research methods like uncertainty-aware optimisation to ensure responsiveness in uncertain systems. It also tackles the complexity of conflicting objectives by using AI-driven multi-level optimization and metaheuristics, such as swarm intelligence, to solve high-dimensional problems while minimising computational demands. To promote fairness and transparency, the project incorporates bias-reducing data re-weighting methods and explainability tools, enabling human-understandable and unbiased decision-making. These innovations mark a significant advancement in scalable, intelligent, and equitable decision systems.

The transition to AI-driven decision-making in data-intensive and uncertain environments relies on a multidisciplinary approach that integrates Computer Science and Operations Research (see Figure 1). This is the focus of TwinODIS, which unites the Institute of Computer Science (ICS) at the Foundation for Research and Technology - Hellas (FORTH) with two renowned European institutions, namely, Université Paris Cité (UPC) - Laboratory of Informatics Paris Descartes (LIPADE) and Erasmus University Rotterdam (EUR) - Department of Econometrics (DE).

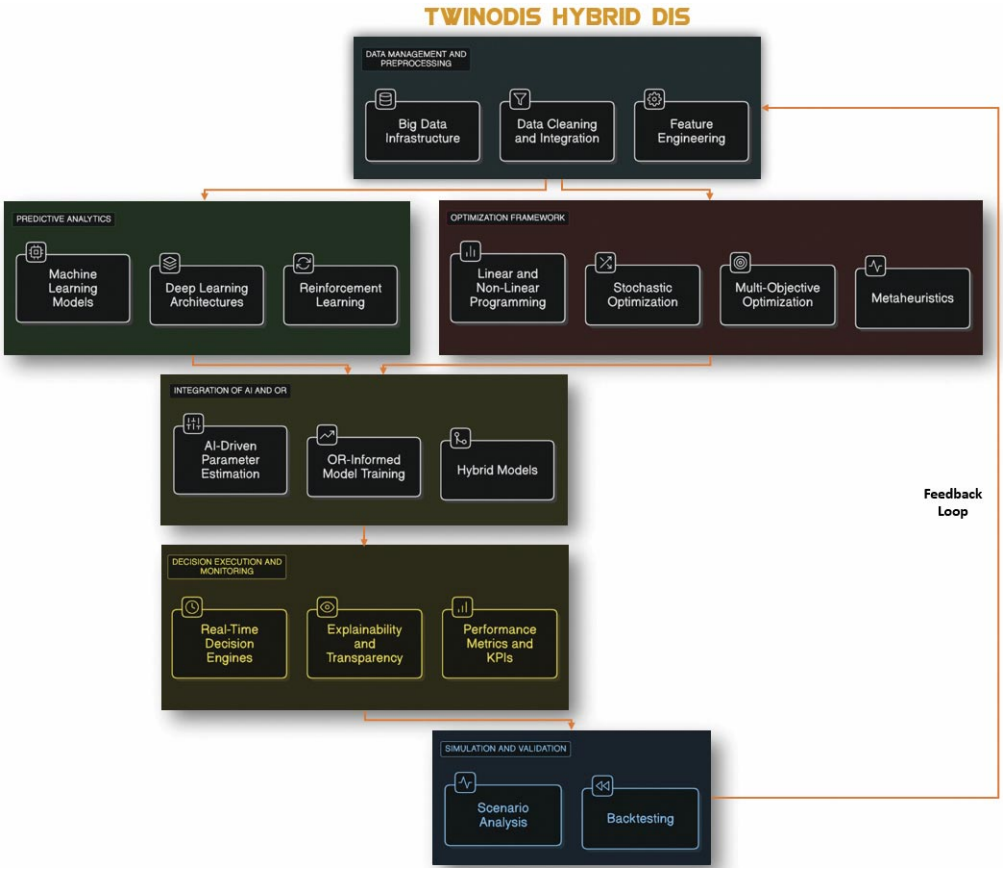


Figure 1: Architecture of TwinODIS decision intelligence system.

Real-time OR techniques are indispensable for decision-making in today's fast-paced and dynamic environments. Unlike traditional OR, which relies on historical data and fixed models, real-time OR utilises live data streams for swift re-

sponses to changing conditions, particularly crucial in sectors like energy, transportation, and emergency response. FORTH-ICS excels in computational methods for time series data in IoT and remote sensing applications. With EUR-DE's expertise in online optimisation, stochastic dynamic programming, risk and uncertainty management, and large-scale multi-objective problem-solving, this collaboration will yield innovative data-driven optimisation and adaptive decision-making, ensuring agility in uncertain data and model situations. Furthermore, the convergence of AI and OR creates powerful systems that enhance decision support capabilities in complex, data-rich environments. In TwinODIS, UPC-LIPADE's expertise in big data and distributed/explainable AI combines with EUR-DE's strengths in multi-objective optimisation and metaheuristic algorithms, to advance FORTH-ICS's machine learning (ML) and deep learning (DL) deployment capabilities. This collaboration integrates AI and OR by improving ML/DL training with advanced optimisation, using ML/DL models to inform OR methods like linear programming, and leveraging ML/DL for pattern recognition, reinforcement learning, and uncertainty handling in large-scale decision-making frameworks.

The TwinODIS project targets two critical applications: multi-energy systems management and smart ports planning, both of strategic importance to the EU. In multi-energy management, TwinODIS addresses the challenge of coordinating renewable and non-renewable energy sources, aiming to enhance energy autonomy and efficiency. The system uses real-time data and hybrid AI-OR models to optimise energy demand, supply forecasting, and grid stability, with key performance goals such as reducing energy costs and increasing renewable energy usage. In smart ports planning, TwinODIS will implement hybrid deep reinforcement learning models to optimise vessel scheduling, berth allocation, and resource distribution at ports. The system will leverage real-time data to improve port efficiency, reduce turnaround times, and decrease CO_2 emissions. Both applications will be evaluated through backtesting and simulations using real-world data from partners like the Dutch Energy Dashboard and the Port of Rotterdam Authority.

Link:

[L1] https://spl.ics.forth.gr/twinodis

References:

[1] Artificial Intelligence Act, European Parliament, Legislative Observatory (OEIL). [URL: https://kwz.me/hFY]

[2] Data Act, European Parliament, Legislative Observatory (OEIL). [URL: https://kwz.me/hFB]

Please contact:

George Tzagkarakis, FORTH-ICS, Greece
gtzag@ics.forth.gr

Rommert Dekker
Erasmus University Rotterdam, The Netherlands
rdekker@ese.eur.nl

Themis Palpanas, Université Paris Cité, France
themis@mi.parisdescartes.fr

Data Mining in Railway Diagnostic Data for Predictive Maintenance

by Giulia Millitarì (University of Pisa and CNR-ISTI), Alessio Ferrari (CNR-ISTI) and Giorgio O. Spagnolo (CNR-ISTI)

We describe the initial and crucial phase of an analysis for a project belonging to the Spoke 4 on "Railway Transportation" of the Italian National Center for Sustainable Mobility (MOST) [L1], which is part of the National Recovery and Resilience Plan (PNRR). The objective of the project is the implementation of a predictive maintenance strategy within the decision-making process of Trenord [L2], a railway company responsible for the operation of regional passenger trains mostly in Lombardy. Before conducting the analysis, it was essential to perform extensive data mining procedures to make the data from the remote diagnostic system truly usable for extracting meaningful insights and apply machine learning techniques effectively.

Trains are a sustainable and efficient way to transport people and goods over long distances. The core pillars of rail transport are efficiency, safety, and reliability, all of which are guaranteed and managed through maintenance activities. Effective maintenance strategies help prevent equipment failures and unplanned downtimes, reducing costs and improving overall service performance.

In recent years, predictive maintenance [1] has emerged as an efficient and promising strategy, gaining prominence in the era of Industry 4.0, for which vast amounts of data are collected. This maintenance type, specifically based on data-driven approaches, exploits information such as sensor and maintenance data to anticipate failures and optimize maintenance schedules accordingly.

For the above mentioned project, Trenord provided data on the maintenance plan, as well as service and diagnostic events for trains in its fleet. Our analysis specifically focused on data related to the traction system of a TSR (Regional Service Train).

The data included two separate datasets on scheduled and corrective maintenance, specifying the type of maintenance, start and end dates, and the wagon with issues. Then, service data provided detail such as departure and arrival stations and kilometers traveled during each service. Additionally, diagnostic data related to the traction system (e.g., power supply loss, water temperatures exceeding a fixed threshold, absence of motor speed signal, and electrical current imbalances) were extracted from Trenord's remote diagnostics system. This data provides insights into the train's operational status, performance, and potential issues, while the system enables remote monitoring of the rolling stock by collecting diagnostic events detected by the train's control units. Figure 1 illustrates a simplified version of the system's structure. The diagnostic events dataset comprised 1,892,948 observations containing information such as the alert criticality, the affected wagon, timestamp, train speed, and latitude and longitude GPS coordinates of the train.

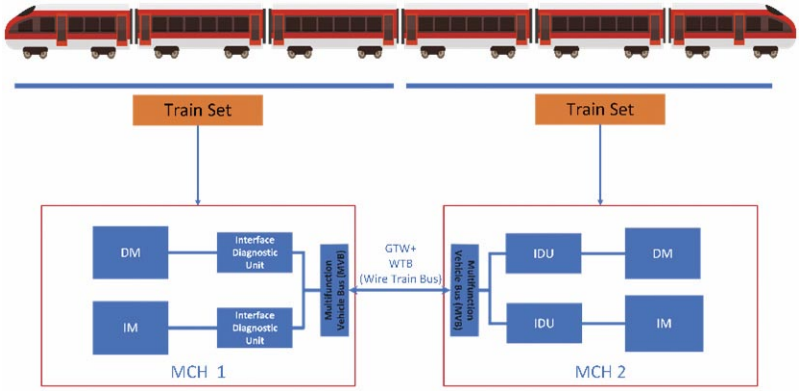


Figure 1: Diagram of on-board hardware structure.

Given the heterogeneity and suboptimal quality of the available datasets, we prioritized a comprehensive data analysis process focused on ensuring data quality. This involved necessary steps in data cleaning and pre-processing to achieve standards of accuracy, timeliness, and interpretability. Consequently, we were able to create a unified dataset containing all the information required to compute meaningful descriptive statistics and accomplish the research objectives. Figure 2 summarizes the entire data mining procedure performed specifically on the diagnostic events dataset.

Through both the documents provided by Trenord and a detailed analysis of the data, such as plotting variable distributions or performing cross-tabulation (contingency tables), we identified several issues related to data incompleteness and inconsistency, particularly in the diagnostic events dataset. The most significant data cleaning effort focused on addressing duplicate data and aberrant values, which had multiple underlying causes. Duplicates were not only identified through identical values across all variables, but also because of specific service-related situations and issues within the remote diagnostic system. Aberrant values, including false declarations, measurement errors, input mistakes, and inconsistencies, were detected by examining each variable's values and through consultation with Trenord staff, requiring a detailed and thorough analysis. Overall, we addressed these data quality issues by removing instances based on ad hoc rules and criteria, using algorithms that we developed for the specific problems identified.

Later, we carried out data pre-processing tasks, including data integration to integrate multiple datasets, data transformation to create new variables not included in the original dataset, such as seasonal effects and service-related factors, and data reduction to aggregate data using an a priori criterion to overcome the curse of dimensionality. Additionally, we applied an undersampling technique to convert the data from a timestamp-based frequency to a daily frequency, helping to regularize the time series and potentially transform the dataset into a cross-sectional one. As a result, all variables had to be reevaluated, as their meanings changed in the daily frequency format. In this format, each row represents a day's events, so, for example, the speed variable reflects the average speed of all events on that day.

In this way, all the different and heterogeneous datasets were combined into a single and smaller dataset, making it more suitable for computing summary statistics and performing statistical analyses, both cross-sectional and time series. This also facilitated achieving the final research objective of implementing a predictive maintenance strategy within Trenord's decision-making process. This initial phase of the project highlighted the importance of thoroughly examining the nature of the data, revealing weaknesses in the data collection system and emphasizing the challenges of enabling immediate predictive maintenance.

Spoke 4 on "Railway Transportation" of the National Centre for Sustainable Mobility (MOST) received funding from the European Union – NextGenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1033 17/06/2022, CN00000023). The project will run until February 2026 and is coordinated by Marco Bocciolone from the Polytechnic University of Milan. Further partners in T3.1 include the universities of Florence, Naples, Parma, and Roma "Sapienza", as well as the industrial partners Accenture, Hitachi, Lutech, and Trenord.

Links:

[L1] https://www.centronazionalemost.it/eg/

[L2] https://www.trenord.it/

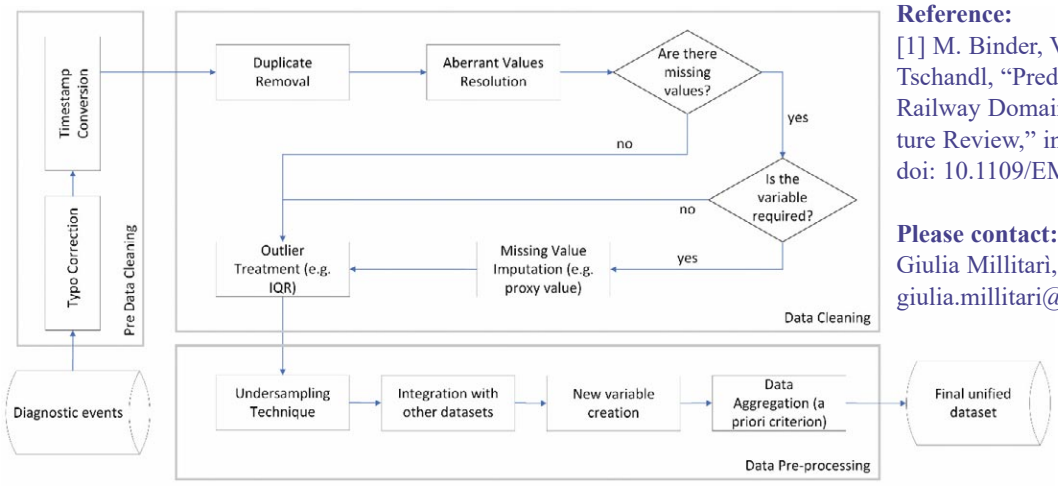


Figure 2: Diagram of data mining procedures on the diagnostic events dataset.

Reference:

[1] M. Binder, V. Mezhuyev and M. Tschandl, "Predictive Maintenance for Railway Domain: A Systematic Literature Review," in IEEE EMR, doi: 10.1109/EMR.2023.3262282

Please contact:

Giulia Millitarì, CNR-ISTI, Italy

giulia.millitari@isti.cnr.it

DataBri-X: Expanding Digital Value Creation in European Data Spaces

by Stelios Sartzetakis (ATHENA RC) and Chamanara, Javad (TIB)

In an era when data availability and AI technologies advance rapidly, the European data economy is poised for substantial growth, unlocking new opportunities and innovations. The DataBri-X project [L1] is motivated by the need to foster the development of trustworthy, "made in Europe" AI that embodies European values and ethical standards. DataBri-X focuses on transforming data-sharing ecosystems by advancing data lifecycle practices, tools, and governance frameworks.

Despite the potential of data, data sharing and interoperability are still in their nascent stages. To realise a truly cross-border and cross-sectoral data-sharing environment, DataBri-X employs comprehensive Data Spaces and data processing tools that allow for seamless processing of proprietary, personal, and open public data. Achieving this vision necessitates overcoming various technical, legal, and business challenges throughout the data lifecycle. DataBri-X not only focuses on conventional raw data and its transformations, but also encompasses metadata, models, and processing algorithms.

DataBri-X applies a fundamental rethinking of data lifecycle practices, focusing on the development and implementation of innovative strategies that prioritise transparency, efficiency, and collaboration in data sharing. By fostering an environment in which data sharing becomes standard practice, the aim is to build trust among stakeholders and establish a robust framework for sustainable data management. DataBri-X advances data tools and services by evaluating current offerings, identifying gaps, and addressing areas for improvement within data-sharing ecosystems. The goal is to create cutting-edge tools that ensure seamless interoperability, accessibility, and usability of data. These tools are designed to align with FAIR principles, reduce energy footprints, and adapt to diverse user needs while supporting innovative business models. FAIR not only in the context of data, but also for the governance process and execution. In DataBri-X, the processes applied on the data as well as the full execution workflow including software systems used, their versions, and the functions applied on data are preserved, can be shared and reused.

Focus is placed on addressing key challenges in data management. Clear frameworks are established to define data ownership rights and responsibilities within decentralised data-sharing landscapes. Mechanisms are implemented to ensure data provenance and verification, enhancing confidence in data quality. Decentralised technologies are explored to enable secure and efficient data sharing, promoting autonomy and privacy without reliance on centralised control. Additionally, robust strategies are employed to safeguard sensitive data through effective confidentiality measures and digital rights management, ensuring that digital rights are respected throughout the data lifecycle. Furthermore, energy-efficient practices are integrated into data processing and sharing to minimise environmental impact and promote sustainability.

Building on the results of relevant past and ongoing initiatives, DataBri-X aims to refine existing data management tools, sys-

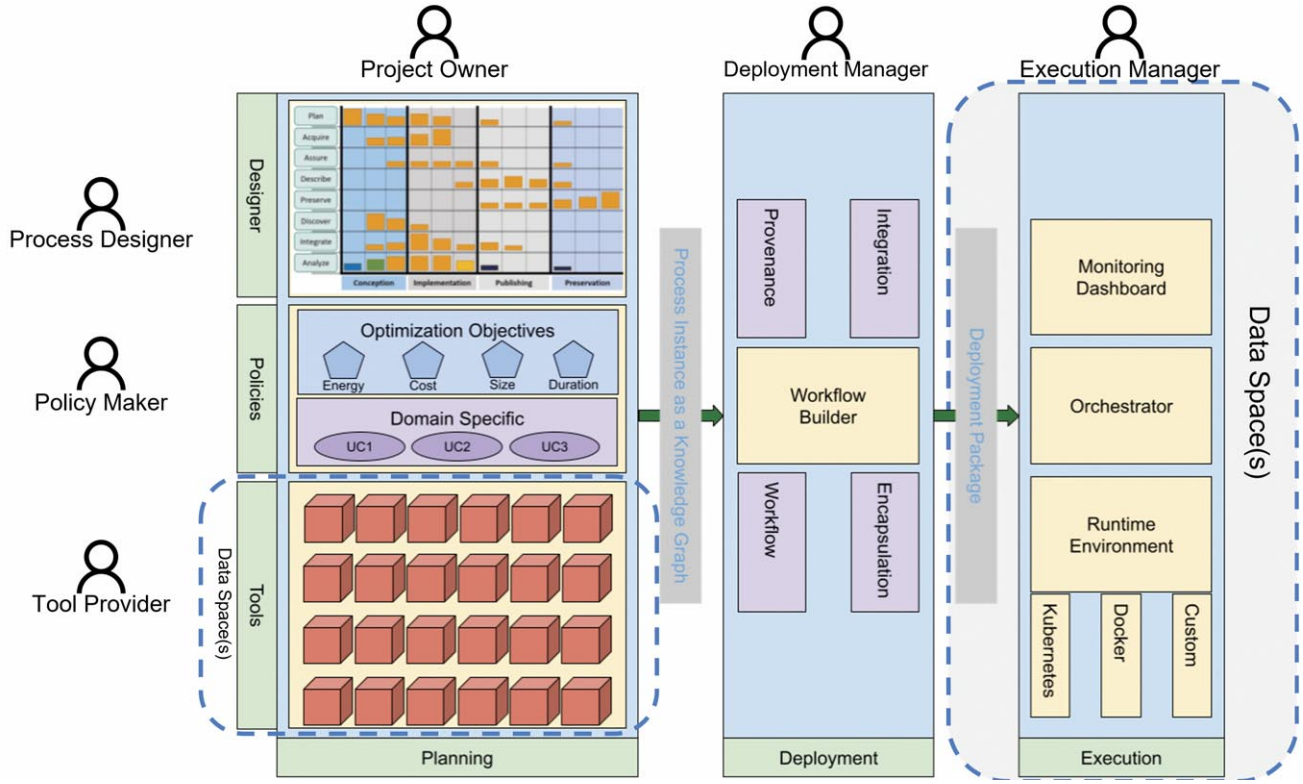


Figure 1: DataBri-X Architecture.

tems, and processes. This includes enabling and automating the creation and maintenance of common ontologies, vocabularies, and data models, as well as supporting automated authoring, co-creation, curation, annotation, and labelling of data. These efforts are geared toward enhancing the usability of data for artificial intelligence and other advanced applications, ensuring that data remains a valuable and adaptable resource for future innovations.

DataBri-X Outcomes

1. A comprehensive management model that redefines data life cycle practices, fostering a culture of responsible and efficient data sharing.
2. Enhanced maturity of data tools and services that are user-friendly, interoperable, and secure, empowering organisations to leverage data effectively.
3. A framework addressing critical challenges related to data ownership, provenance, confidentiality, and energy efficiency, paving the way for a sustainable and equitable data-sharing ecosystem.
4. Implement IDS-compliant Data Spaces.

DataBri-X Architecture

DataBri-X aims to revolutionise data lifecycle practices with a focus on data-sharing ecosystems. By proposing a comprehensive management model, JenPlane, the project enhances the maturity of data tools and services, ensuring they are equipped to meet the evolving demands of secure and efficient data sharing.

JenPlane is a governance model that advocates a process-based approach to data management. It reimagines traditional data lifecycle practices by introducing a dynamic, non-linear framework tailored to the complexities of modern data-sharing ecosystems. By conceptualising the data lifecycle as an interactive plane rather than a sequential cycle, JenPlane offers a flexible model that adapts to the diverse needs of users engaged in data management activities.

JenPlane consists of multiple elements, namely, the processes, the designer, the composer, and the builder. At its core, JenPlane empowers users to design their data lifecycle by customising one of the available process templates and then selecting the most energy-efficient and complementary tools tailored to their specific data management tasks with the help of an LLM-based recommendation engine. This process designer and composer address the challenges of tool selection, collaboration, and orchestration, enabling semi-automatic deployment, execution, and orchestration of data-driven projects. By creating structured working areas that encompass various disciplines such as planning, data collection, validation, semantic annotation, preservation, discovery, and integration, JenPlane facilitates a comprehensive approach to managing data.

While JenPlane serves as the governance backbone, it operates within a broader ecosystem of the DataBri-X project. This ecosystem comprises various software tools that provide essential services across different segments of data-intensive projects. These tools are designed to collaborate seamlessly, ensuring smooth data flows and workflows, ultimately enabling users to meet their project requirements effectively.

Figure 1 “DataBri-X Architecture” illustrates the overall JenPlane Data Governance. Key components include the JenPlane Process Designer and the JenPlane Composer. These components allow users to specify project requirements, select appropriate tools, and assemble an efficient workflow for data governance. JenPlane’s unique structure, which represents project phases on a two-dimensional axis, ensures flexibility and allows multiple activities to proceed in parallel, enhancing adaptability to different data-centric projects. The toolbox will also include a Policy Centre that stores customisable policies, ensuring compliance with security and privacy regulations, such as GDPR, while facilitating sustainability and energy-efficient data processing.

In total, 11 DataBri-X partners are bringing 25 data tools and services together for different areas of the data lifecycle. The tools are improved on the TRL level and integrated into the DataBri-X toolbox that can be configured along the project governance components for easy deployment in Data Spaces.

The effectiveness and practicality of the DataBri-X toolbox are validated through its implementation in three distinct pilot use cases spanning telecommunications, energy, and legal sectors, designed to showcase the toolset’s potential in driving innovation and improving decision-making across diverse domains.

DataBri-X project has received funding from the EU’s Horizon Europe research and innovation programme under grant agreement no. 101070069.

Link:
[L1] https://databri-x.eu

Please contact:
Stelios Sartzetakis, ATHENA Research Centre
stelios@athenarc.gr

Chamanara, Javad, TIB
Javad.Chamanara@tib.eu

interTwin: An Engine for Scientific Digital Twins

by Andrea Manzi, Raul Bardaji and Ivan Rodero (EGI.eu)

The paper reviews the need to create a 'Digital Twin Engine' to support developers in reusing modular components to speed up the building process of Scientific Digital Twins. It highlights how the interTwin project leads this effort by providing a robust framework that cuts development time, supports scalability, and promotes collaboration between domains. Furthermore, the article highlights some ongoing use cases based on this Digital Twin Engine, showing the use and potential impacts.

The increasing demand for accurate simulations and real-time decision-making in various fields has positioned Digital Twins (DTs) as a transformative technology. By creating virtual replicas of physical systems, Digital Twins enable simulation, monitoring, and prediction across domains such as environmental management, healthcare, and industrial operations. Despite their potential, the creation of Digital Twins often faces significant challenges. These include the complexity of integrating heterogeneous data sources, the high computational demands, and the lack of standardised frameworks that enable the reuse of components and collaboration across projects.

To address these obstacles, the European project interTwin introduces an innovative Digital Twin Engine (DTE). This modular framework is designed to simplify the development of DTs by enabling the reuse of existing components, ensuring interoperability, and providing a scalable infrastructure. The DTE Blueprint Architecture, rooted in open standards and co-designed with diverse scientific communities, supports the seamless integration of DTs into various applications.

This article explores the necessity of a DTE to accelerate the adoption of DTs, examines how the interTwin project has developed its DTE and presents case studies showcasing its use and impact.

The Digital Twin Engine from interTwin

The DTE developed by the interTwin project represents a significant step forward in addressing the challenges of creating and scaling DTs. By integrating modularity, interoperability, and federated infrastructures, the DTE provides a streamlined approach to overcome the complexity of data integration and the heterogeneity of computational resources, making the development of DTs more efficient and accessible.

Figure 1 provides an overview of the DTE, highlighting its main components made available to facilitate the development and operation of DTs.

At the core of the DTE, as depicted in the central section, are fundamental capabilities such as data fusion, big data analytics, and the integration of artificial intelligence and machine learning (AI/ML). These capabilities are organised within a workflow composition system. The DTE core modules include DT output quality verification, real-time data acquisition and preprocessing and well.

At the base of the DTE lie components dedicated to federated resource management. These include orchestration, federated computing, data management, connecting high-performance computing (HPC), high-throughput computing (HTC), and cloud storage systems. This modular design allows the DTE to adapt to the specific needs of each DT, ensuring scalability and operational efficiency.

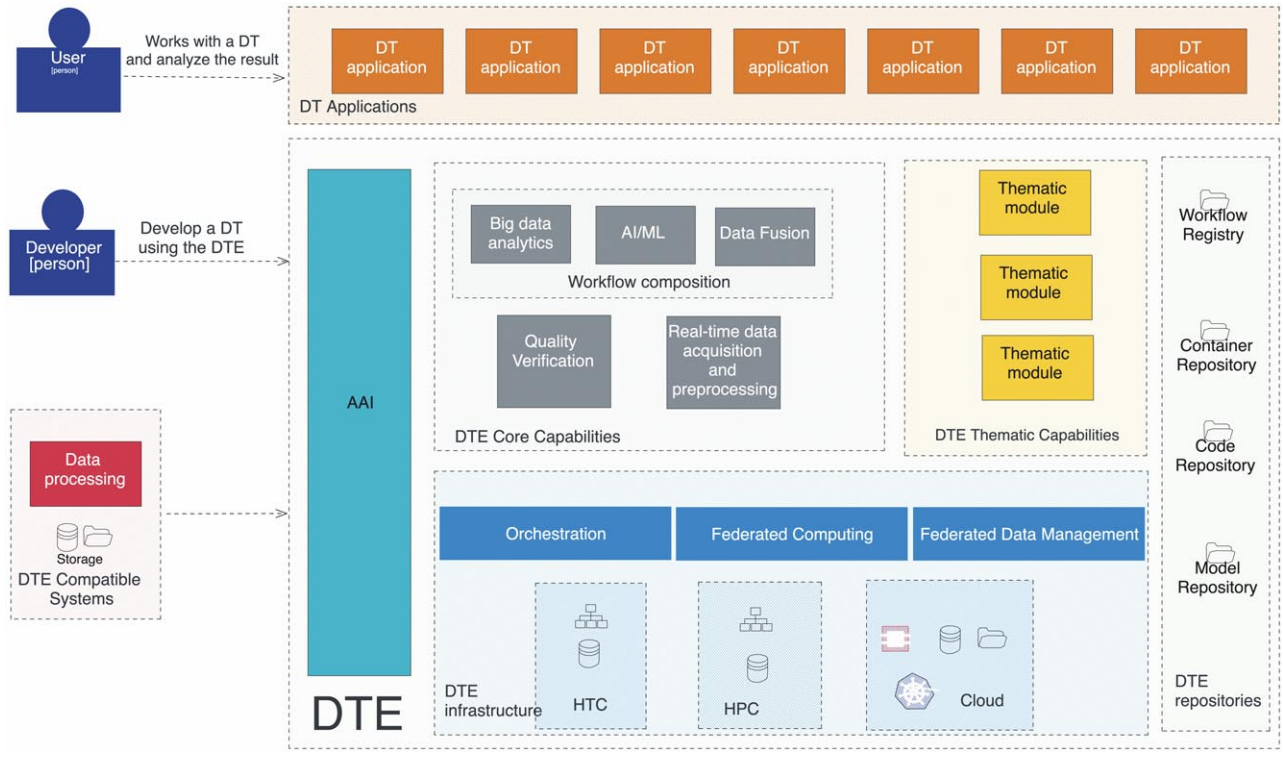


Figure 1: the interTwin Digital Twin Engine Blueprint architecture.

At the top, the DT applications, which serve as user-facing interfaces for analysing results and making decisions based on simulated data. These applications are supported by thematic capabilities, which include specialised modules tailored to specific domains, such as data integration, advanced simulations, and domain-specific tools.

The first version of the components has been released as Open Source and is available for installation and use via the project's website [L1]. A more detailed description of the DTE components can be found in [1].

Use cases with the Digital Twin Engine

The following use cases are developing DTs using the interTwin DTE and demonstrate the flexibility and adaptability of its modules to various needs, as well as their potential to drive innovation across various scientific disciplines:

1. FloodAdapt Deployment: This initiative focuses on creating a Digital Twin for flood impact modelling that can be implemented globally. By integrating rapid, physics-based compound flood modelling with detailed impact assessments, FloodAdapt enables non-expert users to evaluate various flood scenarios and adaptation strategies efficiently.
2. Virgo Gravitational Wave Interferometer Noise Simulation: In collaboration with Virgo, interTwin is developing a Digital Twin to simulate 'noise' within the interferometer. This simulation aids in understanding and mitigating noise sources, thereby enhancing the detection capabilities of gravitational waves.
3. Flood Early Warning Systems: interTwin is advancing components to establish Digital Twins for flood early warning in coastal and inland regions. These systems utilise complex models and satellite data processing pipelines to monitor and predict flood behaviours, providing stakeholders with timely and accurate information.
4. Tropical Cyclone Projection: By employing deep neural networks, interTwin is developing Digital Twins capable of detecting and projecting the occurrence of tropical cyclones under changing climate conditions. This approach enhances predictive accuracy and informs climate adaptation strategies.
5. High-Energy Physics Particle Detector Simulation: interTwin is setting up a Digital Twin for fast particle detector simulation, facilitating more efficient data analysis and experiment planning in high-energy physics research.

Conclusions

The DTE demonstrates how a modular, interoperable, and scalable framework can overcome important challenges in the development and deployment of DTs. By leveraging federated infrastructures, integrating advanced capabilities like AI/ML and data fusion, and adhering to open standards, the DTE simplifies the creation and operation of DTs across diverse domains. Use cases like flood impact modelling, gravitational wave interferometer noise detection, and high-energy physics detector simulations illustrate its flexibility and potential to drive innovation. Looking ahead, the DTE aims to serve as a foundation for collaborative and efficient DT development, enabling broader adoption and impact across scientific disciplines.

Acknowledgements

Although this paper accepts only three authors, the interTwin project represents the collective effort of all participating entities. This article summarises the work accomplished, and its content belongs to everyone involved in the project. We acknowledge the invaluable contributions of all team members and collaborators developing the interTwin DTE.

Link:

[L1] https://www.intertwin.eu/

Reference:

[1] https://www.intertwin.eu/intertwin-digital-twin-engine

Please contact:
Andrea Manzi
EGI Foundation, The Netherlands
andrea.manzi@egi.eu

iMagine: Revolutionising Aquatic Sciences with AI-Driven Image Analysis

by Gergely Sipos (EGI Foundation) and Dick Schaap (MARIS)

iMagine's AI-driven platform and modules empower researchers to analyse vast amounts of images, accelerating scientific discoveries from the micro to the macro level.

iMagine [1] is a European Union-funded project that deploys, operates, validates, and promotes an AI framework and thematic image analysis services dedicated to aquatic sciences and connected to the European Open Science Cloud (EOSC) and AI4EU initiatives. Its mission is to provide researchers in aquatic sciences with open access to a wide range of AI-based image analysis services and image repositories from various Research Infrastructures (RIs). The project focuses on the overarching theme of "Healthy Oceans, Seas, Coastal and Inland Waters."

iMagine enhances existing image analytical capabilities in the marine and freshwater sciences to improve research performance in the whole aquatic domain. Specifically, it (1) operates a generic computational platform, the 'iMagine AI Platform', that helps aquatic communities develop AI-based image analysis services, (2) develops and supports the development of AI-based image analysis services for 12+ specific scientific problems within aquatic research; (3) offers labelled images that can be used for the training and re-training of AI models and (4) captures and shares best practices related to imaging data and image analysis with artificial intelligence in aquatic sciences. Ultimately, iMagine delivers a portfolio of scientific capabilities that are targeted at researchers in marine and freshwater research.

The iMagine AI Platform [2], customised from the AI4OS framework [3] and supported by AI4EOSC [4], is a powerful

tool designed to revolutionise aquatic science research. The platform is hosted on four OpenStack clouds from the EGI e-Infrastructure Federation to deliver GPU and storage capacity at scale and according to the dynamically changing capacity requirements of the supported use cases. The platform offers a comprehensive range of tools and techniques, including image annotation, preprocessing, deep learning models catalogue, model training and performance evaluation, model inference for scientific end users. This platform supports the entire machine learning cycle, from model development to training, validation and delivery to end users, enabling easy sharing and collaboration among AI experts and domain researchers.

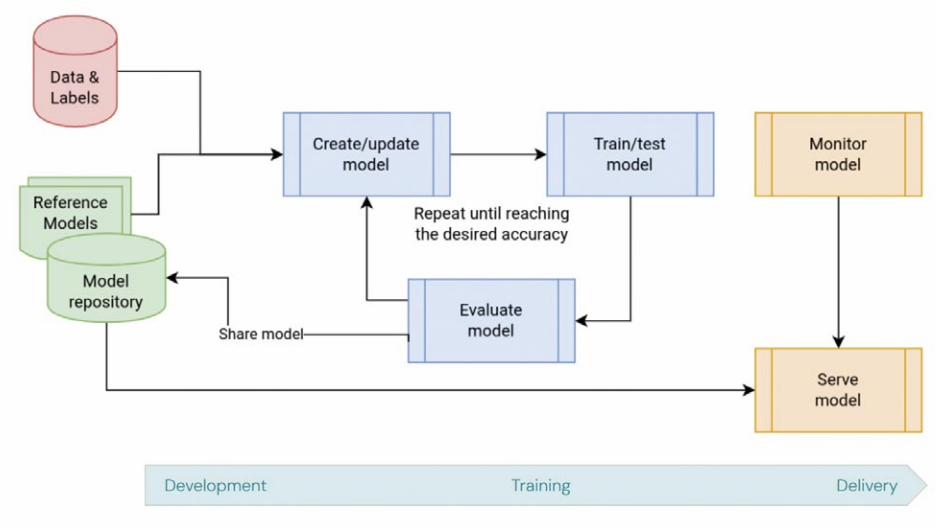


Figure 1: Overview of the steps supported by the iMagine platform, from development to training to delivery.

By utilising artificial intelligence and machine learning, researchers can analyse vast amounts of image data from various sources, including underwater platforms, webcams, microscopes, drones, and satellites.

Figure 1 provides an overview of the steps supported by the iMagine platform, from development to training to delivery.

iMagine supports a growing number of use cases with the AI platform, each focusing on a different problem within aquatic sciences, and demonstrating the power of AI for image analysis in the field. The use cases address issues such as

- classification and quantification of floating litter from drone images taken over open water surfaces,
- taxonomic identification of planktons from microscopic images,
- generating ecosystem statistics (e.g. fish occurrences) from images and video streams produced by underwater cameras,
- predicting the movement and spread of oil spills in the ocean from satellite images,
- estimating the location and movement of boats from images generated from underwater audio streams,
- monitoring the health of coral reefs from underwater images.

Figure 2 offers at a glance all available services in the platform. The services are also available in the service catalogue [5] on the iMagine website, where interested users can submit an access request before they are officially released.

The supported use cases collaborate with the platform providers using the so-called 'iMagine Competence Centre', a virtual support team that brings together generic AI experts,

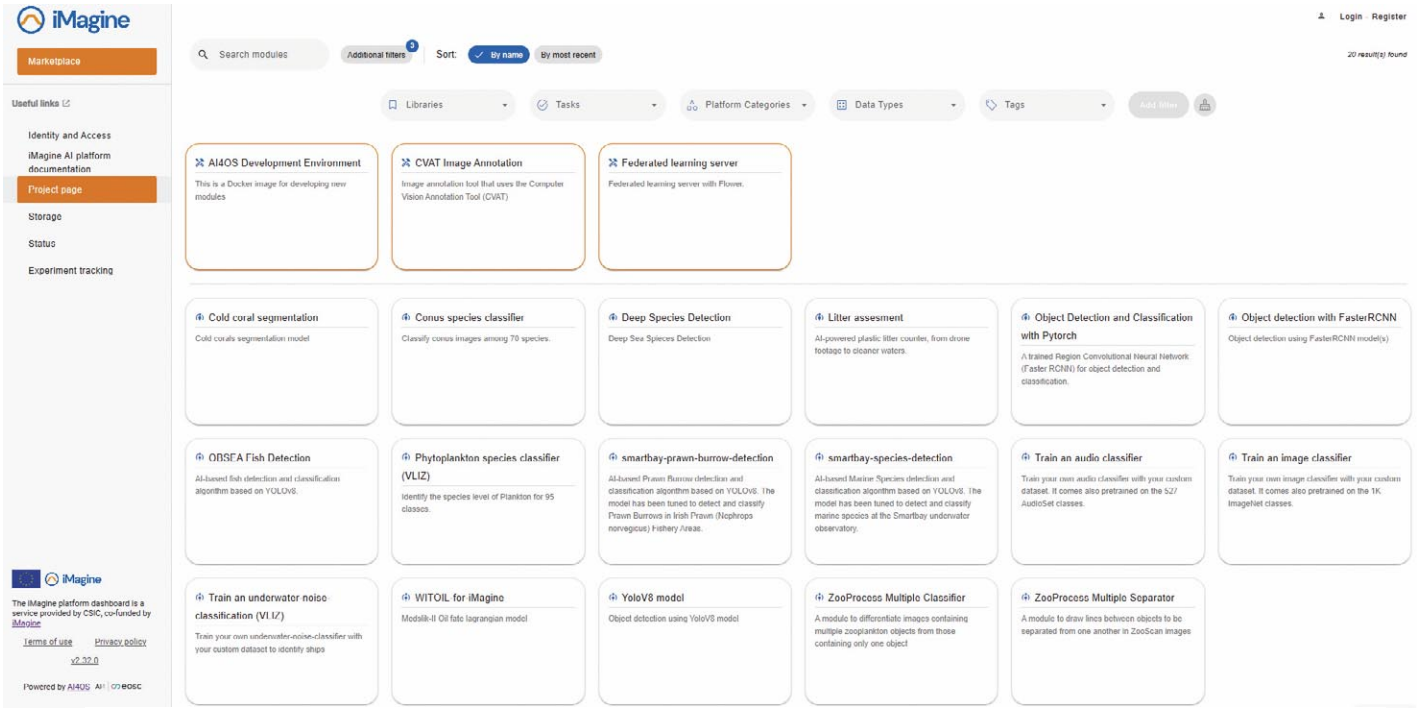


Figure 2: All available services in the platform at a glance.

domain scientists, and image data owners. The Competence Centre convenes the stakeholders through regular meetings, training sessions, and continuous feedback collection, all designed to facilitate experience sharing and progress towards reaching precise AI models and robust online services for end-user scientists.

Besides helping the use cases within the project, the Competence Centre also captures and publishes best practices, including a comprehensive overview of the AI tools and techniques used in various projects in a recent project deliverable titled "Best Practices for Producers and Providers of Image Sets and Image Analysis Applications in Aquatic Sciences" [6]. Despite this deliverable specifically focusing on aquatic sciences, it contains valuable insights that can be beneficial to researchers who face challenges in image analysis in other fields of science.

The Competence Center already supports 12 use cases; through the continuously open call on the website, we invite additional use cases to benefit from the iMagine support.

All use cases contribute to the image datasets available publicly on Zenodo [7] to validate the published AI models, retrain them, or train additional models from these data.

By adhering to best practices in data management, quality control, and model development, iMagine aims to enhance data quality, promote reproducibility, and facilitate scientific progress in aquatic research. The platform is dedicated to sharing its developments with other leading projects, such as EOSC, AI4EU and Blue-Cloud 2026, to maximise its impact and encourage the broader adoption of AI-powered solutions within the aquatic science community.

Links:
[1] https://www.imagine-ai.eu/
[2] https://dashboard.cloud.imagine-ai.eu/marketplace
[3] https://ai4os.eu/
[4] https://ai4eosc.eu/
[5] https://kwz.me/hFb
[6] https://doi.org/10.5281/zenodo.13864196
[7] https://kwz.me/hFt

Please contact:
Gergely Sipos
EGI Foundation, The Netherlands
gergely.sipos@egi.eu

Dick Schaap
Maris, The Netherlands
dick@maris.nl

Data-Enhanced Agriculture: Leveraging Analytics for Efficient Water Usage

by Karina Medwenitsch, Markus Schindler, and Christoph Klikovits (Forschung Burgenland GmbH)

How can advanced data analysis reshape agriculture in Austria's climate-stricken Seewinkel region? By combining IoT, AI, and real-time environmental analysis, researchers at Forschung Burgenland are pioneering innovative solutions to optimise water management and support the energy transition, ensuring resilience in the face of climate change.

The Green Sentry research project addresses the pressing challenges posed by climate change in the Seewinkel region of Austria, a vital agricultural area increasingly affected by extreme weather conditions. Starting in 2024, this initiative leverages advanced digital technologies and data analysis to develop sustainable solutions for water management and agricultural resilience. By combining innovative IoT sensors, real-time monitoring, and cutting-edge analytics, Green Sentry aims to enhance resource efficiency, reduce environmental impact, and strengthen the region's adaptability to climate stressors.

The Seewinkel region located in Eastern Austria is particularly vulnerable to climate extremes such as droughts and intense heat waves, which have drastically affected agricultural productivity. Farmers have reported crop yield reductions of up to 50% for staples like maize and soybeans. Groundwater depletion, high irrigation costs due to reliance on diesel-powered systems, and occasional irrigation bans exacerbate the challenges. These issues demand innovative and sustainable solutions to ensure agricultural viability and economic stability in the region.

Previous research has shown the significant potential of IoT technologies for improving water management, particularly in agriculture. In the Civis 4.0 Patria project [L1], IoT-based solutions were employed to monitor environmental conditions and manage water resources, providing a foundation for future efforts like Green Sentry. This project demonstrated the effectiveness of IoT technologies in real-time data collection, aiding in the efficient allocation and use of water resources in various sectors. Similarly, studies have highlighted the benefits of IoT-enabled modern technologies for irrigation management, emphasising their ability to provide real-time data, optimise irrigation schedules, and reduce water consumption by dynamically adjusting irrigation needs [1] [2]. These findings contribute to the ongoing development of IoT-based systems that can monitor and manage water resources efficiently, providing the basis for Green Sentry's approach to optimising water usage in Seewinkel.

The primary objective of Green Sentry is to address the aforementioned challenges by developing and implementing a scalable technological framework for sustainable water management and agricultural optimisation. The project aims to create precise systems for monitoring and regulating water use, par-

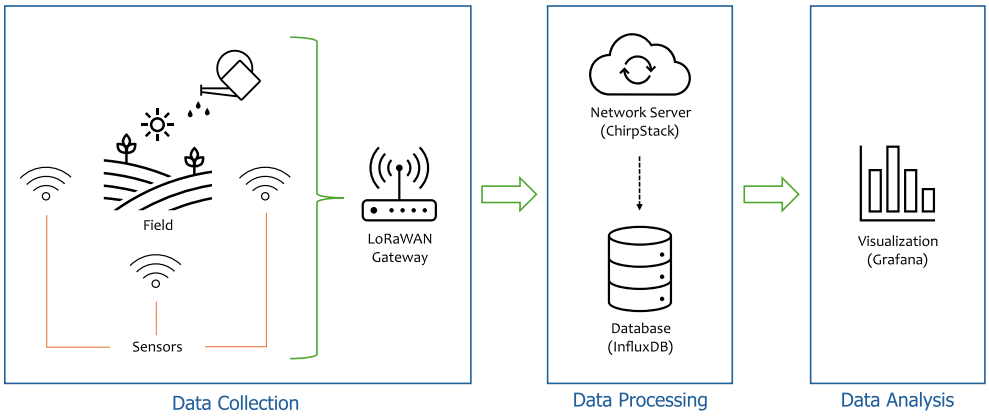


Figure 1: Collection, processing and analysis of the sensor data within Green Sentry.

ticularly in large-scale irrigation systems. This involves employing advanced technologies to collect and analyse environmental data, enabling data-driven decision-making that enhances irrigation efficiency, reduces water consumption, and preserves groundwater resources. Additionally, the project fosters collaboration among local stakeholders, ensuring the solutions are practical and tailored to the region's specific needs.

The methodology of Green Sentry integrates IoT technologies, cloud-based data processing, and advanced visualisation tools to facilitate comprehensive environmental monitoring and efficient resource management. IoT sensors deployed across wells and agricultural fields measure critical parameters, including water levels in wells, air humidity, air temperature, soil temperature, leaf temperature, light intensity, soil moisture, and leaf moisture. A Long-Range Wide Area Network (LoRaWAN) gateway installed in the Seewinkel region acts as the hub for transmitting sensor data. This data is collected in an open-source LoRaWAN Network Server (ChirpStack), where it is processed and stored in a time-series database (InfluxDB) for further analysis. To make the information accessible and actionable, the data is visualized via an open-source data visualisation and monitoring solution (Grafana), providing stakeholders with real-time insights into environmental conditions. The process of acquiring, processing and analysing the sensor data is illustrated in Figure 1. The technologies used in this project enable precise monitoring of water usage and environmental variables, allowing for the identification of inefficiencies and the optimisation of irrigation strategies. Additionally, the interoperability of these systems ensures they align with existing platforms, enhancing the utility and scalability of the solutions developed.

The data analysis techniques employed by Green Sentry play a central role in the project's success. By continuously collecting data from a variety of environmental sources, the project can generate a real-time, comprehensive understanding of water availability and environmental conditions. The data analysis helps identify patterns, predict water needs, and optimise irrigation schedules, all of which lead to reduced water consumption and more efficient use of resources. This data-driven approach can be particularly useful in areas like Seewinkel, where the risk of water scarcity is high, and resource management is critical.

While Green Sentry focuses on optimising water management in agriculture, it also opens avenues for further research. Future studies could expand real-time data analysis by incorporating environmental factors, such as the impact of weather patterns on soil and plant conditions, to refine predictive models for irrigation and automated water usage systems. Beyond agriculture, Green Sentry's approach could be applied to sectors like water treatment, urban planning, and climate resilience in vulnerable regions for large-scale data analytics. The integration of IoT technology, data analysis, and cloud solutions opens up new opportunities for efficiently managing resources and creating innovative approaches for smart cities. Data convergence techniques could also enhance disaster response systems, supporting real-time decision-making during extreme weather events.

Through its innovative approach, Green Sentry not only supports the immediate needs of the Seewinkel region but also provides a scalable model for tackling similar challenges in other climate-vulnerable areas.

Link:

[L1] https://forschung.hochschule-burgenland.at/projekte/projekt/civis-40-patria/

References:

[1] S. Ismaili, et al., "IoT-Based Irrigation System for Smart Agriculture," 2024 XXXIII International Scientific Conference Electronics (ET), pp. 1-6, 2024, doi: 10.1109/ET63133.2024.10721573.

[2] N. Nawandar and V. Satpute, "IoT-Based Low Cost and Intelligent Module for Smart Irrigation System," Computers and Electronics in Agriculture, vol. 162, pp. 979-990, 2019, doi: 10.1016/J.COMPAG.2019.05.027.

Please contact:

Karina Medwenitsch
Forschung Burgenland GmbH, Austria
karina.medwenitsch@hochschule-burgenland.at

Data Intermediaries Enabling Governance and Federated Analytics in Energy Communities

by Christoph Klikovits (Forschung Burgenland) and Christoph Fabianek (OwnYourData)

The energy sector generates a high volume of data, but data analysts face significant barriers due to issues like security, privacy, and GDPR compliance. These challenges often hinder data sharing, analysis, and interpretation, which are essential for unlocking the added value and insights that data can provide. How can accessible governance solutions help to overcome these obstacles in the energy domain?

As part of a research project, a model for a data service ecosystem is being developed to advance the energy transition. The overarching goal of the flagship project USEFLEDS is to create services that address key challenges in the energy sector. These include energy economics, governance, data security, privacy, data spaces, and data analytics, all crucial for fostering sustainable energy systems.

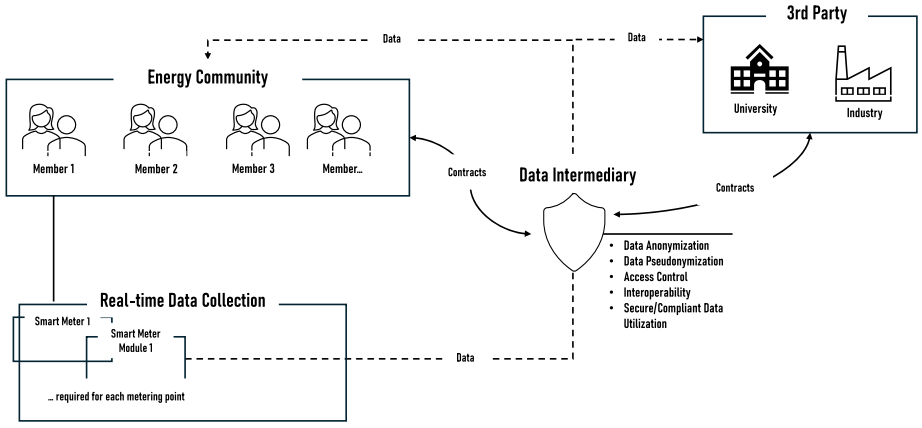


Figure 1: Data Intermediary architecture for energy communities.

The project involves various academic, industrial, and governmental institutions collaborating to ensure a holistic approach. These include energy providers, technology developers, legal experts, and research organizations specialising in data governance and security. The research is conducted in Austria, with pilot studies in specific energy communities that serve as testbeds for innovative data sharing and analytics models.

The energy sector faces significant barriers to data utilisation due to regulatory frameworks like the GDPR and the EU Data Governance Act. Challenges include managing sensitive data securely, mitigating liability risks, and addressing a lack of know-how and resources. These obstacles hinder the effective use of data for analysis and decision-making. The project seeks to demonstrate how data intermediaries can overcome these barriers, enabling secure data sharing and federated analytics.

The primary aim is to leverage data intermediaries to create transparent, secure, and privacy-compliant data ecosystems in the energy domain. Specifically, the project focuses on how intelligent, federated data processing can calculate energy surpluses in energy communities, enhancing efficiency and sustainability. [1]

The project employs advanced techniques to address the challenges of data management and governance in the energy sector. Real-time data is collected from smart meters, generating vast amounts of information that serve as the foundation for analysis. To facilitate secure and compliant data exchanges, data intermediaries play a central role, ensuring adherence to governance frameworks. Access control is managed through robust identification and authentication services, safeguarding data from unauthorised use. To protect sensitive information, anonymization and pseudonymization techniques are applied, maintaining privacy while enabling meaningful data utilisation (see Figure 1). Federated analytics further enhance this process by deriving valuable insights without the need to share raw data directly, preserving both security and confidentiality.

The project examines data intermediaries as enablers for accessible governance frameworks and federated large-scale data analytics in the energy sector. By addressing security, privacy, and interoperability challenges, data intermediaries empower stakeholders to utilise data effectively while ensuring compliance and trust [2].

Future initiatives could expand the role of data intermediaries to areas like peer-to-peer energy trading, intelligent flexibility management in power grids, and other innovations within energy communities. These advancements would further enhance the sector's ability to adapt to the increasing complexity of energy systems and accelerate the energy transition.

Through its comprehensive approach, this project highlights the transformative potential of data intermediaries in creating a sustainable and data-driven energy future.

Link:
[L1] https://usefleds.forschung-burgenland.at/

References:

[1] J. C. Schweihoff, I. Jussen, and F. Möller, "Trust me, I'm an intermediary! Exploring data intermediation services," 2023.
[2] A. Shaharudin, B. van Loenen, and M. Janssen, "Exploring the contributions of open data intermediaries for a sustainable open data ecosystem," Data & Policy, vol. 6, p. e56, 2024, DOI: 10.1017/dap.2024.63.

Please contact:

Christoph Klikovits, MSc.
Forschung Burgenland GmbH, Austria
christoph.klikovits@forschung-burgenland.at

Scalable Anomaly Detection in Renewable Energy Grids using the GLACIATION Platform

by Ioannis Rotskos (IPTO), Orestis Vantzos (IPTO) and Panagiotis Papadakos (ERCIM)

The transition towards sustainable energy systems is accompanied by huge data volumes generated by modern electricity grids. Within the GLACIATION EU project, Use Case 4 (UC4) demonstrates how scalable anomaly detection algorithms and the edge-cloud continuum enable efficient analytics in energy grid management. This use case highlights the deployment of advanced analytics frameworks for processing SCADA data, extracting actionable insights, and optimizing the energy utilization of renewable sources. UC4 employs the GLACIATION platform, which orchestrates distributed workloads across data centers, focusing on sustainability and cost efficiency by leveraging locally produced green energy.

Large-scale data analytics is pivotal in addressing the challenges posed by the integration of renewable energy into electricity grids. Data from Supervisory Control and Data Acquisition (SCADA) systems, which monitor energy transmission infrastructure, require robust processing frameworks for anomaly detection and grid optimization. In the context of the EU-funded GLACIATION project [L1], the Use Case 4 (UC4) led by IPTO, explores the application of distributed algorithms and federated data processing on the GLACIATION platform to enhance grid resilience and operational efficiency, while at the same time taking advantage of available green energy in the grid.

The GLACIATION project is developing a platform that reduces energy consumption for data processing and analytics through AI-enforced minimal data movement operations. This platform enables organizations to deploy and manage analytics across the edge-core-cloud continuum in a secure, energy-efficient, and scalable manner. The GLACIATION platform has been built around the Kubernetes platform [L2] that automates the deployment, scaling and management of containerized applications. A cornerstone of the platform is the use of a Distributed Knowledge Graph (DKG) that spans across the edge-core-cloud continuum and takes advantage of the GLACIATION Metadata Reference Model (GLC-MRM) [L3]. GLC-MRM provides a conceptual model that allows making data ingestion and data processing interoperable for all the use cases of the GLACIATION project. The GLC-MRM model can be considered as a general conceptualization of a task scheduling problem that considers various measuring indicators over the deployed resources. Figure 1 provides an overview of the Metadata Service and its interaction with the core services of the GLACIATION platform.

In the context of the GLACIATION project, UC4 processes multivariate SCADA datasets encompassing current, voltage, and power measurements from three grid substations. Data are sampled at fixed intervals and are enriched with metadata, including sensor specifications, grid topology details and site RES (Renewable Energy Sources) capacity. Workloads run anomaly detection algorithms, generating annotated outputs for actionable insights. The focus is on detecting anomalies that signal faults or inefficiencies, such as a wind turbine disconnect or voltage irregularities. By aligning workload orchestration with green energy availability at each site, the system balances computational demand with renewable energy pro-

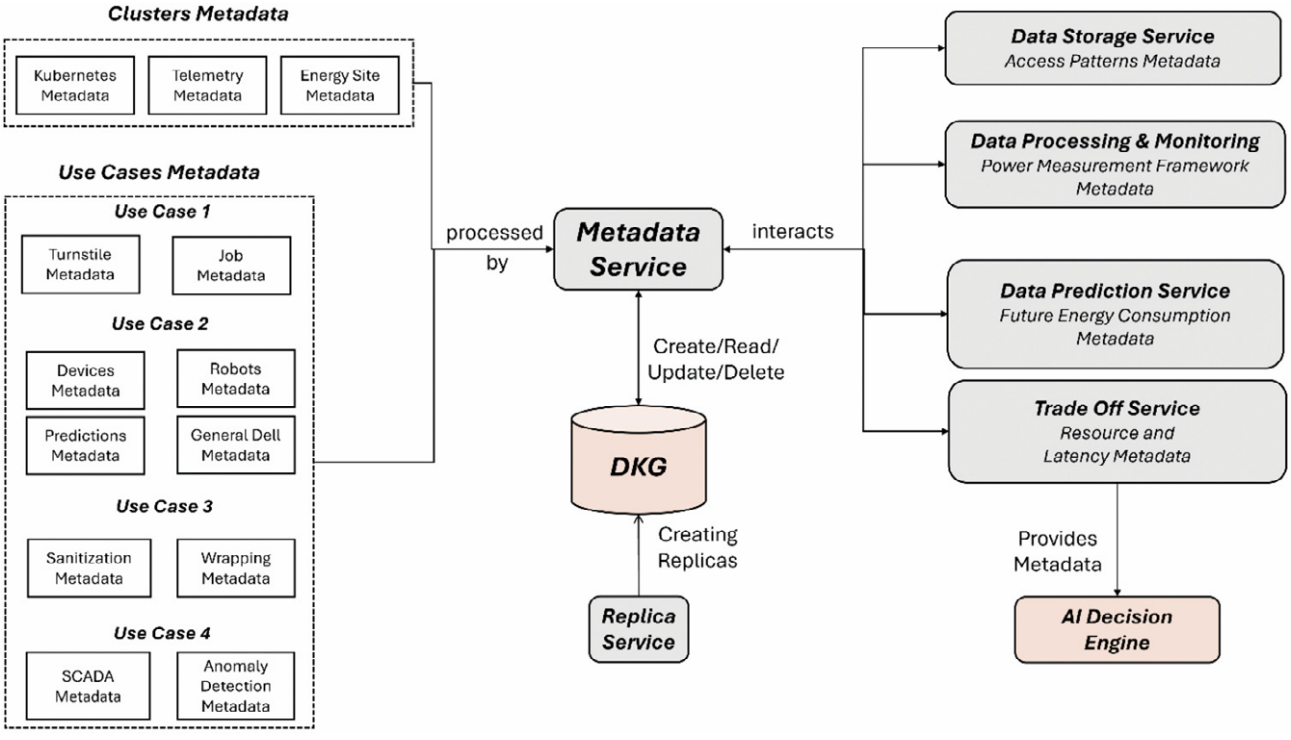


Figure 1: The Metadata Framework: An overview of the Metadata, the Metadata Service and its interaction with the DKG and the other services.

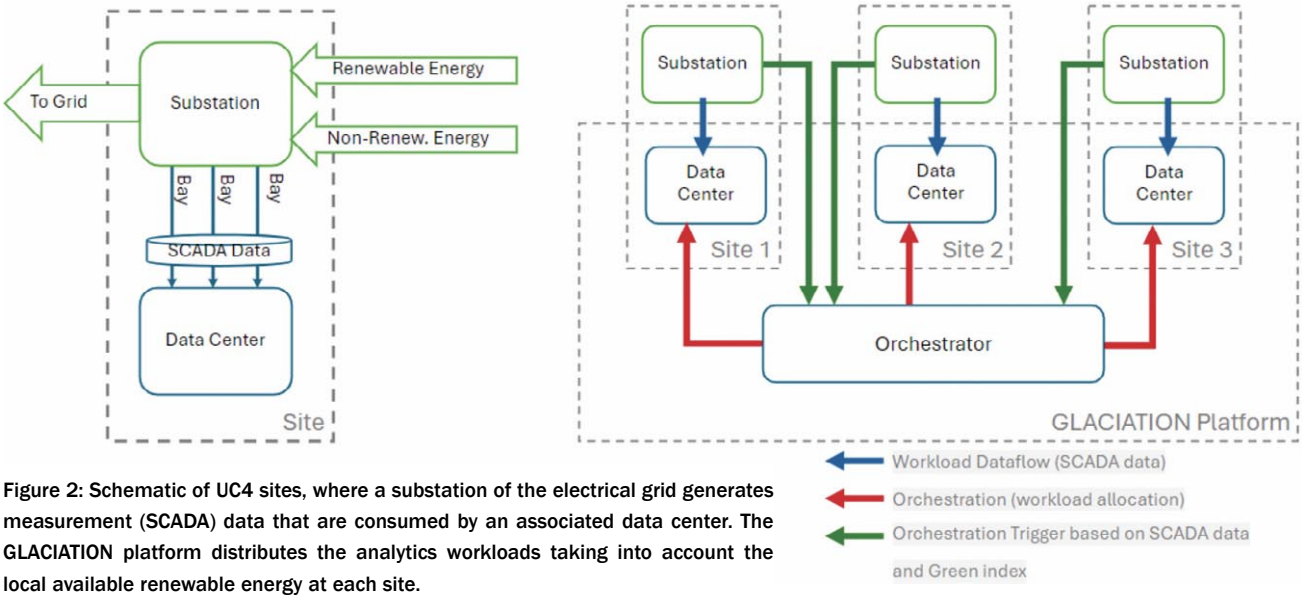


Figure 2: Schematic of UC4 sites, where a substation of the electrical grid generates measurement (SCADA) data that are consumed by an associated data center. The GLACIATION platform distributes the analytics workloads taking into account the local available renewable energy at each site.

duction, embodying a sustainable approach to data-intensive operations. These outputs are aggregated to inform grid operators and support predictive maintenance. In this way, the platform reduces reliance on non-renewable resources while alleviating grid congestion (see Figure 2).

The key contributions of UC4 are the following:

1. Advanced Analytics Techniques:
   The anomaly detection algorithms implemented within UC4 include Isolation Forest [R1], One-Class SVM [R2], and rule-based methods (see Figure 3 for an example). These algorithms are optimized for distributed execution, handling the inherent variability and scale of grid telemetry data.
2. Edge-Cloud Continuum:
   Data centers near grid substations operate as part of a Kubernetes-based cluster. This infrastructure supports dynamic workload assignment based on real-time energy availability, showcasing the potential of federated data analytics in renewable energy applications.
3. Competency Queries for Operational Insights:
   The system addresses critical queries, such as computing average renewable energy usage, evaluating workloads based on the Green Index, and identifying grid segments that meet specific renewable production quotas. These insights support operational decisions and enhance grid management. GLC-MRM was extended to capture the requirements of UC4, such as energy resources and grids, physical locations, the anomaly detection workloads and the SCADA data (see Figure 3).

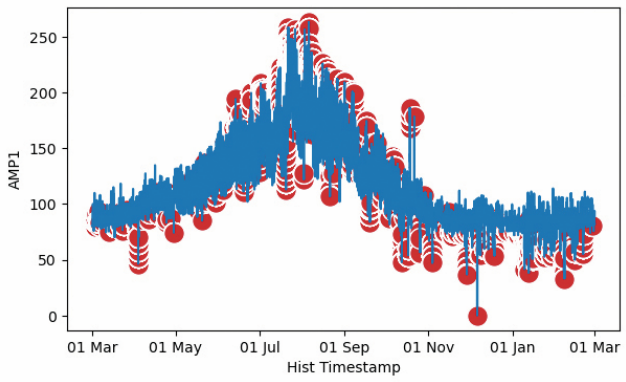


Figure 3: The graph illustrates the results of an anomaly detection algorithm applied to the current measured by a Remote Terminal Unit (RTU) over a specified time-period. Red dots highlight detected anomalies in the current production.

UC4 demonstrates the transformative potential of large-scale data analytics in the renewable energy sector. The integration of scalable anomaly detection algorithms, edge-cloud architectures, and green energy optimization showcases a novel approach to sustainable grid management. These insights contribute to the broader discourse on Big Data ethics, energy analytics, and the social implications of mass data processing.

The following colleagues also contributed to this article: Dimitrios Brodimas (IPTO), Theodoros Grigorakakis (IPTO), Dimitrios Skipis (IPTO), Michalis Mountantonakis (ERCIM), and Rigo Wenning (ERCIM).

Links:

[L1] https://glaciation-project.eu/
[L2] https://kubernetes.io/
[L3] https://glaciation-project.eu/MetadataReferenceModel/1.1.0/

References:

[1] W. S. Al Farizi, I. Hidayah, and M. N. Rizal, "Isolation forest based anomaly detection: A systematic literature review," in 2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), Sep. 2021, pp. 118–122.
[2] N. Seliya, A. Abdollah Zadeh, and T. M. Khoshgoftaar, "A literature review on one-class classification and its potential applications in big data," J. Big Data, vol. 8, pp. 1–31, 2021, doi: 10.1186/s40537-021-00446-9.

Please contact:

Orestis Vantzos, IPTO, Greece
o.vantzos@admie.gr

Panagiotis Papadakos, ERCIM, France
panagiotis.papadakos@ercim.eu

Extraction of Structured Information from Large-scale European Digital Financial Reports

by Alex Suta (Széchenyi István University), Loránd Kedves (Széchenyi István University), Árpád Tóth (Széchenyi István University)

The adoption of eXtensible Business Reporting Language (XBRL) for annual corporate disclosures is reshaping data accessibility and analytical methodologies. This technical study explores how European companies use XBRL to enhance data standardisation, which facilitates large-scale financial and sustainability analyses for practitioners and researchers.

The role of financial reporting in decision-making

Financial reporting serves as a cornerstone for corporate transparency and accountability, providing essential information to investors, regulators, and other external stakeholders. According to the International Accounting Standards Board, its primary objectives are decision-usefulness—enabling informed investment and resource allocation—and stewardship, ensuring accountability for resources entrusted to companies [1].

General-purpose financial reports include structured financial data (e.g., balance sheets and income statements) and qualitative disclosures, such as risk management and sustainability initiatives [1]. Annual reports support critical activities like trend analysis, benchmarking, and regulatory compliance. However, the increasing complexity of global business environments and regulatory demands has necessitated a shift toward digitisation, paving the way for the global adoption of innovative reporting frameworks such as XBRL [2][3].

A framework for digital transformation

XBRL was developed to transform the accessibility and comparability of corporate reporting. Its taxonomy-based framework provides a standardised approach to structuring and tagging financial and non-financial data. The International Financial Reporting Standard's (IFRS) taxonomy used in European digital reports includes more than 7,600 tags (concepts) that cover numeric, textual, and structural elements. These tags align with key report sections such as the statements of financial position or cash flow, and notes to the financial statements [2]. By encoding reports in XBRL, companies ensure that their disclosures are machine-readable, enabling automated processing and large-scale analysis.

Despite its potential, the adoption of digital reporting has faced significant challenges. Many companies rely on custom tags alongside the standardised IFRS taxonomy, which leads to inconsistencies and complicates comparisons. Additionally, a lack of robust IT infrastructure for processing XBRL reports limits its full utility. Errors in tagging and data duplication are common, requiring extensive cleaning and validation. The complexity of the taxonomy itself poses challenges for report preparers, increasing the likelihood of misinterpretation and non-compliance [4].

To address these challenges, our research developed a comprehensive framework for processing XBRL reports. This included retrieving 7,842 annual reports from the XBRL filings portal [L1] as of December 31, 2023, and employing a Java-based processing pipeline integrated with Python 3 tools. The extracted dataset was standardised to Euros using exchange rates from the European Central Bank, ensuring comparability across companies and periods. Stock-type data, such as balance sheets, utilised final-day exchange rates, while flow-type data, such as income statements, used period averages. The methodology also resolved inconsistencies in tagging by mapping custom tags to the closest IFRS equivalents and ensuring high data alignment.

Highlights from the total population analysis of digital reports

The analysed dataset comprised 3,469 companies with an average of 1.7 reports per company, and more than 3.2 million data points. Numeric disclosures accounted for 86.6% of the dataset, forming the backbone of financial analyses. Textual information, constituting 10% of the dataset, was primarily found in the notes sections of reports, which offer qualitative

Tag name	# Tags	Unique Reports	% Reports	Avg. Occ. / Report	Mean (EUR M)	St. dev. (EUR M)	Min. (EUR M)	Max. (EUR M)
ProfitLoss	44,741	7,639	97.47%	5.9	395	2,084	***-18,225***	***40,714***
CashFlowsFromUsedInFinancingActivities	15,031	7,614	97.15%	2.0	-203	1,887	-39,841	33,943
Equity	36,808	7,612	97.13%	4.8	3,024	11,502	-6,783	200,893
Assets	14,777	7,605	97.04%	1.9	22,543	132,768	***0***	***2,781,295***
CashAndCashEquivalents	36,062	7,563	96.50%	4.8	1,711	14,454	-1,243	489,097
ComprehensiveIncome	29,377	7,544	96.26%	3.9	366	2,176	-18,625	42,689
CashFlowsFromUsedInInvestingActivities	14,894	7,531	96.10%	2.0	-327	1,857	-41,460	22,349
CashFlowsFromUsedInOperatingActivities	14,837	7,460	95.19%	2.0	574	4,765	-79,875	102,276
NameOfReportingEntityOrOtherMeansOfIdentification	7,430	7,373	94.08%	1.0	-	-	-	-
IncomeTaxExpenseContinuingOperations	14,613	7,370	94.04%	2.0	134	1,133	-3,926	47,349

Table 1: Descriptive statistics on the most frequently used XBRL tags in 2022 reports.

	Numeric (Monetary)	Text	Other
General information about financial statements	0	7,486	339
Notes	1,473,424	388,061	25,144
Statement of cash flows	343,552	0	0
Statement of changes in equity	419,107	21	0
Statement of comprehensive income	426,213	235	16,230
Statement of financial position	323,961	0	0

Table 2: Estimated data point distribution by annual report sections.

insights into corporate activities, such as accounting policies and even sustainability-related activities.

Key financial metrics, including Profit Loss and Equity, demonstrated high adherence to IFRS taxonomy standards, with 87% of tags aligning with the taxonomy. The analysis identified significant variability among companies, as presented by the descriptive analysis in Table 1. The transparency of data is achieved by the traceability of any data point to actual reports in their traditional human-readable format. According to the highlighted extreme values, Shell reported a maximum profit of € 40.7 billion, while Electricité de France (EDF) posted a record loss of € 18.2 billion from the population. Similarly, HSBC Holdings reported assets of € 2.78 trillion, contrasting sharply with the nil values reported by smaller entities like CoreCapital I K/S.

The identification of sustainability-related disclosures was another critical focus. Tags such as "DescriptionOf AccountingPolicyForEmissionRights" provided insights into companies' environmental practices. Provisions related to environmental remediation appeared in over 11,000 instances, while contingent liabilities tied to sustainability risks were tagged more than 11,200 times. Such disclosures indicate that XBRL enables the integration of financial and sustainability reporting, which support a comprehensive approach to corporate transparency and responsibility [5]. Moreover, the distribution of data points across report sections reveals that the notes section contains the majority of numerical and textual disclosures, highlighting its importance in detailed financial and sustainability analyses, as presented in Table 2.

Although prior research has established the utility of the XBRL corporate reports as a primary source of data, their application to the entire European population has not been examined. As a contribution, the present work investigated the current state of digital business information disclosed in the recently accessible European electronic reports. The created knowledge base proves beneficial in extracting structured financial as well as business data presented in the notes to financial statements. As a result, the research questions set a good foundation for exploring potential uses of the data, with three potential application areas identified:

1. Business information users and academic researchers can both benefit from the generated data tables and central knowledge base. The dataset is particularly useful for producing financial ratios and data visualisations for specific companies or industries and can facilitate more complex machine learning analyses, text mining, and Natural Language Processing (NLP) techniques directed at extracted qualitative data.
2. In line with the European Union's goal of expanding the implementation of XBRL by more companies, data-driven or free-text searches of existing reports could support report preparation capabilities and benchmarking activities.
3. Large-scale report browsing for taxonomy preparers, including automatic detection of best practices, with additional support for standard creation consultation.

Links:

[L1] https://filings.xbrl.org
[L2] https://kwz.me/hFd

References:

[1] R. Ball, "What is the actual economic role of financial reporting?," Accounting Horizons, vol. 22, pp. 427–432, 2008.
[2] IFRS Foundation, "Climate-related Disclosures—IFRS S2," 2022. [Online]. Available: https://www.ifrs.org/projects/work-plan/climate-related-disclosures
[3] ESMA, "European Single Electronic Format," 2022. [Online]. Available: https://www.esma.europa.eu/issuer-disclosure/electronic-reporting
[4] A. Suta, et al., "Overview of XBRL Taxonomy Usage for Structured Sustainability Reporting in European Filings", Chemical Engineering Transactions, vol. 107, pp. 577-582, 2024.
[5] P. Molnár, et al., "Linking sustainability reporting and energy use through global reporting initiative standards and sustainable development goals", Clean Technologies and Environmental Policy, pp. 1-9, 2024.

Please contact:

Alex Suta, Vehicle Industry Research Center, Széchenyi István University, Hungary
suta.alex@ga.sze.hu

A Multimodal Fusion Architecture for Sensor Applications

by Michael Hubner and Jan Nausner (AIT Austrian Institute of Technology)

In this article, we introduce our Multimodal Fusion Architecture for Sensor Applications - MuFASA, which our research group has developed at the Austrian Institute of Technology. It offers a robust fusion architecture for real-time sensor applications, providing situational awareness and precise decision support.

The Austrian Institute of Technology (AIT) has developed an innovative multimodal fusion architecture for sensor applications, known as MuFASA. This technology is designed for real-time sensor applications, aiming to enhance the safety and security of end-users such as firefighters, paramedics, and other practitioners operating in challenging and dangerous conditions.

Technological Overview

MuFASA introduces a robust fusion architecture that integrates heterogeneous data from various sensors and external sources to establish situational awareness in real-time. This system ensures precise decision support by offering modular fusion modules that exploit sensor data on different levels. The levels are categorised as Signal Level Fusion, Feature Level Fusion and Decision Level Fusion. At each level, specific challenges are addressed to ensure the full potential of the available sensor information is exploited.

To illustrate, at the Signal level, raw sensor data is employed to enhance or even create new features based on homogeneous and heterogeneous sensor modalities. Signal level fusion occurs in real-time fusion scenarios or may be an additional step in the pre-processing of signals to yield features of interest. A representative example would be image fusion of RGB images and infrared images for the purpose of detecting persons under difficult lighting conditions.

At the feature level, the objective is to fuse the sensor features with the intention of establishing spatio-temporal coincidence across all sensor modalities. This is a crucial step in providing situational awareness, as demonstrated by the generation of heat maps. The fusion modules dedicated to this level are also designed with the aim of improving detection accuracy and reducing false alarms compared to a system that does not utilise any fusion modules. At the decision level, fusion models operate with the refined features of the previous level. Thus far, only sensor data has been used in the fusion process. A primary objective of the decision level is to incorporate external information, such as end-user experience and contextual or tactical information, into the fusion process. This output represents the highest level of fusion and is essential for precise decision support.

MuFASA selects its methodologies with a focus on real-time capability and ease of interpretation by those responsible for operational management. This provides immediate response capabilities to the acting personnel. The system is based on well-known fusion architectures described by JDL and Lou and Kay [1]. The software modules employ state-of-the-art statistical methodologies in the field of soft computing, such as Probabilistic Occupancy Mapping, Bayesian Inference, and first-order logic. Figure 1 illustrates these methodologies. These techniques enable the system to process and interpret vast amounts of sensor data efficiently.

MuFASA is being developed by the Austrian Institute of Technology (AIT), which is drawing on insights from previous and ongoing research projects. AIT's dedication to developing technology for public safety and security has been a key factor in the success of this project. The primary objective of the research group at AIT is to enhance the safety and security of end-users, including firefighters, paramedics, and other practitioners. In search and rescue operations, unmanned systems are increasingly used to assess situations and locate victims under complex and dangerous conditions. These tasks require numerous sensors, which generate a vast amount of data that can be difficult to interpret. MuFASA addresses this challenge by providing a robust fusion architecture that integrates data from different sensors and external sources, ensuring real-time situational awareness and precise decision support. The work is the culmination of years of research, which commenced in 2016. The project's objective is to provide a comprehensive approach to multimodal sensor data fusion, ensuring the utilisation of technologies is not limited to specific use cases. This flexibility and generalised approach enables MuFASA to be employed in a range of real-time sensor applications, enhancing its versatility and effectiveness. The technology is currently being evaluated in use cases across various sectors, in-

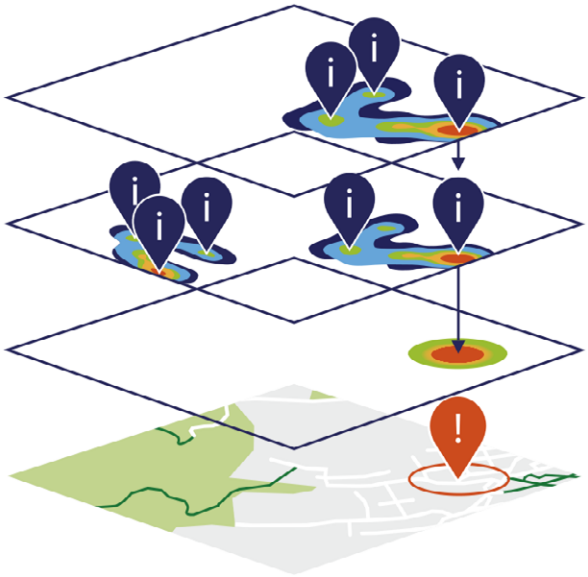


Figure 1: Illustration of MuFASA´s employed feature level methodologies utilizing Bayesian Inference and heat maps. The result of the statistical fusion is represented as an alert (red marker).

cluding border security, surveillance of critical infrastructure, search, and rescue operations, and CBRN reconnaissance.

Future activities

MuFASA represents a significant advancement in sensor data fusion technology, offering real-time situational awareness and precise decision support. Developed by the Austrian Institute of Technology, this innovative system is poised to enhance the safety and security of end-users across various sectors, ensuring that practitioners can operate more effectively in challenging and dangerous conditions.

Its capabilities will continue to evolve with new use cases in these sectors. The main goal is to ensure precise communication of results and minimise false alarms. To achieve this, ongoing research focuses on optimization procedures to find optimal fusion configurations necessary for designing effective fusion systems. This will result in high-quality, interpretable data that can be easily understood by practitioners, ultimately increasing the safety and security of personnel in dangerous missions.

MuFASA's capabilities have been evaluated in the context of railway safety in a national research project (completed in 2023) called MOBILIZE [L1]. This project evaluated the performance of MuFASA in three typical critical railway security events.

Link:

[L1] https://projekte.ffg.at/projekt/4105746

References:

[1] Meng et al., "A Survey on Machine Learning for Data Fusion". Elsevier, Volume 57, p115-129, https://doi.org/10.1016/j.inffus.2019.12.001

[2] M. Hubner et al., "A Bayesian Approach - Data fusion for robust detection of vandalism and trespassing related events in the context of railway security," 27th Int. Conf. on Information Fusion (FUSION), Venice, Italy, 2024, pp. 1-7, DOI: 10.23919/FUSION59988.2024.10706430.

Please contact:

Michael Hubner
AIT Austrian Institute of Technology, Austria
michael.hubner@ait.ac.at

Resource-aware Detection of Satellites Streaks in Deep Sky Images Streams

by Olivier Parisot (Luxembourg Institute of Science and Technology)

Capturing deep sky video streams has become accessible and inexpensive thanks to recent hardware and software advances, but the growing number of satellites in Low Earth Orbit (LEO) generates undesired light pollution. Thus, we are currently developing a resource-aware AI system for automatically detecting specific targets like satellite streaks in video streams produced with affordable observation stations.

Amateur and professional astronomers can witness that the number of satellites in Low Earth Orbit (LEO) is constantly increasing, as it has a major impact on both visual observation and telescope imaging [L1]. The public is not necessarily aware of this, due to a lack of interest on the subject, but also because of the supposed absence of technical means to verify it.

However, recent advances in astronomical equipment allow to obtain high-resolution astronomical image streams with little effort. For wide-field observation, there's now no need to have access to a professional observatory: recent smart telescopes are equipped with sensitive CMOS sensors, and can therefore obtain decent images with exposure times of a few seconds, making it possible to produce continuous streams of raw images - which are either combined later to obtain final quality images (astrophotography), or used as they are for live sky monitoring purposes (transient events like meteors, supernovas). These automated instruments, or all-sky devices (e.g., cameras equipped with fisheye lenses), are relatively easy to use and affordable (prices now start at just a few hundred euros). Even recent smartphones are equipped with sensors sensitive enough to capture images of the night sky, the Milky Way... and satellite streaks (Figure 1). This is accessible to schools, science outreach associations, and even individuals.

These installations can generate large data streams, with a considerable number of high-resolution images. Processing these streams requires specialized software (such as ASTRiDE [2]) to detect and filter satellite streaks: this is often a resource-intensive and offline process, needing significant data storage. In a forthcoming article, we reported that 0.16% of the astronomical images we captured between 2022 and 2023 from Luxembourg Greater Region were affected by satellite streaks. Processing these images required heavy computations on several dozen gigabytes of data.

Nowadays, algorithms and hardware improvements mean that complex calculations can be performed efficiently on data streams for a whole range of tasks: it is now possible to run accurate AI models to process images and videos with low resource usage and latency.

As part of research projects at Luxembourg Institute of Science and Technology (LIST) about AI-powered technolo-

gies for the Space domain [L2], we are working on an online detection system to analyse continuous streams of sky images, and we are particularly interested in satellite streaks: this problem is obviously being addressed for large observatories [L1], but we are more specifically targeting methods that can be implemented on systems with limited capacity. In other words, we are targeting stream analysis that does not require significant storage requirements, using a lightweight computing device (such as a Raspberry Pi or even a smartphone).

To this end, we are currently working on this workflow:

- The first step consists of designing and training a supervised deep learning model for detection, for instance by using YOLO (You Only Look Once) [1]: model training and evaluation is based on images that we have collected ourselves over the years using various observation instruments, and then annotated. The tiniest model architectures (with low parameter count, ~6M or less) are preferred: they offer fast inference, lower computational and energy requirements, making it ideal for real-time applications on resource-constrained devices, though with slightly reduced accuracy (it's a trade-off to find). To go further, an incremental technique like DeepSORT can help avoid recalculating annotations for each image by considering the result of the previous image in the stream [2].
- Then, the trained deep learning model(s) must be adapted to the environments/devices on which it will be executed. This step is done via the compression and the quantification of the trained YOLO models, in order to obtain different lightweight versions of the trained models (with weights encoded in 8-bit integers or 16-bit floats, instead of 32-bit floats).
- Finally, the final application must be based on an intelligent strategy to execute the models obtained according to the resources of the targeted computing device. To this end, we apply 'Resource-aware control', focusing on efficient algorithms to process continuous streams of images under constrained computational, memory, and energy consumption [3]. When processing video streams, it is therefore a question of using models according to the actual occupation of the device's memory (bearing in mind that the largest models are potentially slower and more demanding) and adapting the size of the inputs if necessary for each inference (reducing the size of the inputs also reduces CPU/GPU and memory consumption). All this must be done without compromising the accuracy of detection (by minimising false negatives/positives).

Next step: finalizing and embedding the whole process (in a Raspberry Pi or a smartphone), and then interfacing with a setup capable of generating streams of astronomical images (smart telescope or other), showing in near-real time if satellite streak appears. When running, it will allow to highlight the importance of the number of satellites in the sky (for example, through local impact studies, such as for light pollution linked to urban lighting). This is unlikely to influence the actors the satellite industry, but it could enable educators and enthusiasts to raise awareness among the general public, and why not among politicians.

Figure 1: A picture of the night sky with a Pixel 4a in 'astrophotography' mode (August 22, 2022) – at the right of the image, a satellite track was captured.

Links:

[L1] https://kwz.me/hFy
[L2] https://kwz.me/hFc
[L3] https://github.com/dwkim78/ASTRiDE

References:

[1] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2023.
[2] R. Gandhi et al., "Multiple Object Detection and Tracking Using DeepSORT," in Communications in Computer and Information Science, Springer Nature Switzerland, Nov. 2024, pp. 438–448.
[3] A. M. Shiddiqi, E. D. Yogatama, and D. A. Navastara, "Resource-aware video streaming (RAViS) framework for object detection system using deep learning algorithm," MethodsX, vol. 11, p. 102285, 2023.

Please contact:

Olivier Parisot, Luxembourg Institute of Science and Technology, Luxembourg
olivier.parisot@list.lu

Visualising Big Traffic Data in GLayer to Guide Policy Development

by Jiri Bouchal (Digital Resilience Institute), Hugo Matousek (InnoConnect), Jan Ježek (University of West Bohemia)

GLayer is GPU-accelerated backend software designed for fast aggregation, filtering and visualisation of spatial data. Modelled on the OpenGL technology, GLayer is capable of performing analytical queries on large-scale datasets with millions of data points in a matter of milliseconds. At present, the tool is being tested in Aarhus as part of the BIPED project to support the city's transition to net-zero emissions.

GLayer exploits thousands of lightweight processing cores in a GPU unit to achieve its performance. Upon entering GLayer, a dataset is split into smaller segments. Because each is handled simultaneously by one of many GPU cores, aggregation-based analytics can be performed exceptionally fast. Even with standard GPU hardware, such as that used in laptops, it takes less than 100 milliseconds to process tens of millions of records [1].

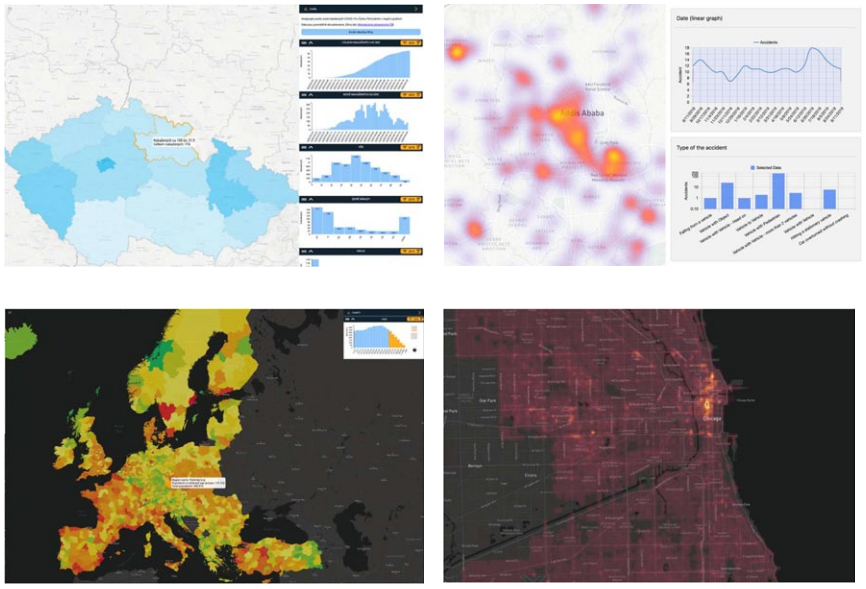


Figure 1: GLayer-based maps of COVID-19 infections in Czechia (top left), traffic incidents in Ethiopia (top right), Europe demographic data (bottom left), and crimes in Chicago (bottom right).

In addition to a fast response time, GLayer provides graphically rich output to aid visual analysis of processed data, in the form of dashboards and heatmaps. These front-end tools can aid decision-making in a variety of sectors, from transport and the environment to public health and safety (see Figure1).

Architecture

The GLayer Server software architecture is designed as a distributed, modular system that integrates diverse data sources to create and manage a high-performance, GPU-accelerated index (see Figure 2). This index is optimised for scalability and operates either as a single node or in a multiple node configuration, leveraging parallel processing capabilities to handle large-scale datasets efficiently.

At its core, the GLayer system features a robust data-ingestion layer that normalises and preprocesses incoming data from various sources, such as traditional relational databases via JDBC, NoSQL databases, with information then fed into the GPU-based indexing engine.

The project management user interface (UI) [L1] enables users to manage data sources, configure indexing parameters, and coordinate collaborative workflows seamlessly. Specifically, it serves three main purposes:

- Datastore configuration: data connectors can access data from permanent storages such as SQL databases (local or remote) or csv files
- Project configuration: a project can define how the dataset is going to be visualised and filtered. This includes data type definitions, aggregation strategies, filtering capabilities and cartographical outputs such as choropleth maps, heat maps, histograms etc.
- Project visualisation: The system includes a web-based visualisation client, providing users with an intuitive interface for exploring indexed data through dynamic, interactive visualisations.

The overall system is built with a microservices approach to ensure modularity, fault tolerance, and scalability, while the frontend leverages modern web technologies for a responsive and user-friendly experience.

Besides the GUI frontend, GLayer supports integration with other systems via an Open Rest API. The API offers extended capabilities for accessing third-party data without having to alter the data source or reconfigure the underlying project. The case study below illustrates how this was done in practice using TomTom data to support the development of a traffic model for Aarhus.

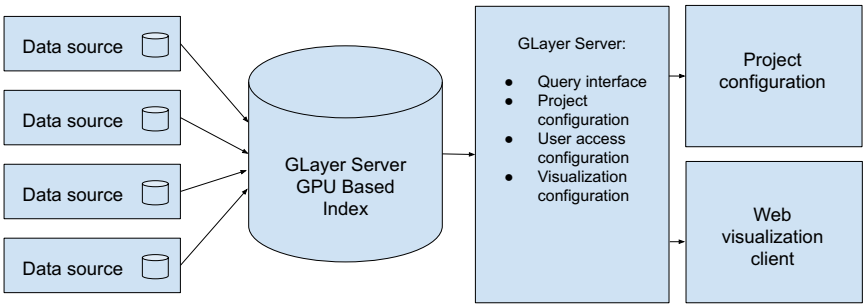


Figure 2. GLayer Server architecture.

Figure 3. A traffic-speed dashboard for Aarhus with segment visualisation and charts based on GLayer.

Aarhus use case

Aarhus wants to become climate-neutral by 2030. To achieve this goal, the city needs to offset its current footprint of about 1.3 million tons of CO2e. A macroscopic traffic model is currently being created as part of the BIPED project [L2] to provide scenario modelling for decarbonisation-oriented planning measures.

The model relies on several inputs. One is the manually enhanced traffic network from OpenStreetMap. Another is the origin-destination (OD) matrix, which was developed using big traffic data from TomTom, such as data from SIM cards and transportation records [L3].

Information on average speeds and travel times was extracted from TomTom reports for selected road segments and time periods (15-21 January 2023, 15-21 April 2023, 15-21 July 2023, 30 October 2023 - 6 November 2023). The aim was to provide a representative picture of the traffic situation in the city during a full week (i.e. without public holidays) in each season (winter, spring, summer, autumn).

For each day of the week, a full traffic report covering all the main road segments in Aarhus was generated, with data points aggregated on an hourly basis. Spatio-temporal data was then parsed via a custom configurable script to enable its mapping onto a corresponding database entry.

Policy makers can access this data via GLayer's interface (see Figure 3) without writing any code. They can filter data according to a desired period or day of the week to get aggregated insights for several or all road segments or view the traffic situation at a particular point in time in a single segment.

With the current data from TomTom, the traffic data from which the OD matrix was created covers the city's traffic only partially (15-25%). Additional data acquisition is planned to further improve the model and provide accurate insights into city-wide traffic patterns that can be used as a basis for cross-sectoral 'what-if' simulations [2].

The GLayer use case in Aarhus is developed as part of the BIPED project, which has received funding from the EU Horizon Europe Research and Innovation programme under Grant ID 101139060.

Links:

[L1] https://innoconnect.net/docs/GLayer/project-configuration/
[L2] https://www.bi-ped.eu/
[L3] https://www.tomtom.com/products/traffic-stats/

References:

[1] J. Ježek, et al., "Design and Evaluation of WebGL-Based Heat Map Visualization for Big Point Data," in The Rise of Big Spatial Data, I. Ivan, et al., Eds., Lecture Notes in Geoinformation and Cartography, Springer, 2017, pp. 118–122, https://doi.org/10.1007/978-3-319-45123-7_2.
[2] J. Ježek, K. Jedlička, and J. Martolos, "Visual analytics of traffic-related Open Data and VGI," in Proc. ICIST 2015 Conference, 2015.

Please contact:

Jiri Bouchal
Digital Resilience Institute, Czech Republic
jiri@digitalresilienceinstitute.org

Knowledge-driven Strategy for Scalable Land-cover Mapping Using Earth Observation Data

by José García-Nieto (ITIS, University of Málaga), Virginia García Millán (ITIS, University of Málaga), and José F. Aldana-Montes (ITIS, University of Málaga)

Researchers from ITIS Software work on projects for the generation of big data workflows for processing and analysis of earth observation, remote sensing satellite data. These handle massive amounts of data to obtain value-added applications in agroforestry, the environment, smart cities, and for society in general. As a use case, this paper provides an example of the use of Sentinel-2 satellite data for the generation of a land-cover map over a large area, the Mediterranean basin, using machine-learning algorithms and big data analysis.

Remote sensing-based earth observation (EO) is becoming increasingly important as it provides a robust technological framework for developing innovative applications in diverse areas, including climate change, precision agriculture, smart urban planning, and land cover evolution. Within this context, projects such as GreenSenti [L1] and EnBiC2-Lab2 (LifeWatch ERIC) [L2] are being developed by experts from ITIS Software (University of Málaga) to provide EO-based research tools to support investigations into the functions and services of agroforestry and ecosystems, aiding society in addressing critical challenges.

One of the key tasks, where EO plays a major role in the development of these projects, is Land Cover mapping (LC), which provides a meaningful way to describe the Earth's surface. Spatially detailed land-cover data are essential for local, national, and international decisions regarding natural resource management. It influences the functional relationship between topography, climate, and soil while offering biophysical insights into the environment and the factors driving change. In an increasingly digital world, LC mapping has evolved into a big data challenge, with the sheer volume of data requiring management becoming a demanding task. Remote sensing generates extensive datasets characterised by unique properties, such as being multi-source, multi-scale, high-dimensional, dynamic, and nonlinear. Moreover, satellite remote sensing has long been considered the most effective method and data source for large-scale land cover classifications.

Mapping land cover on a large scale presents significant challenges due to spectral heterogeneity and the complexity of ter-

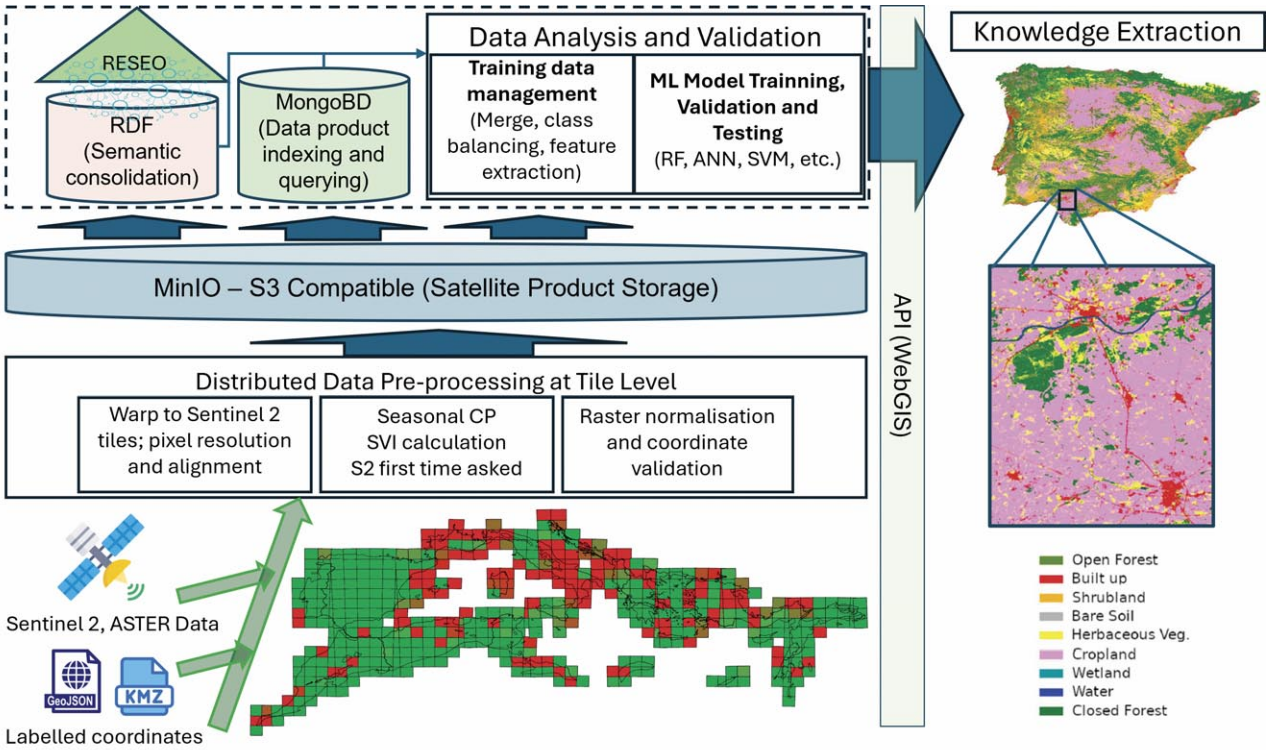


Figure 1: Global overview of the strategy for large-scale Earth Observation data integration and analysis in the context of land cover in the Mediterranean basin scenario. An approximate number of 450 Sentinel-2 and around 1200 ASTER tile products were processed (> 4TBs). Data are collected, pre-processed and stored in a high-scale MinIO repository. MongoBD is used for data indexing and flexible querying. Additionally, RDF mapping is also performed according to RESEO ontology, so reasoning tasks can be defined for supporting knowledge extraction. Final phases consist in training data management and machine learning model generation and testing. As a result, land cover maps are obtained with identified labels: Closed and open forests (green and olive), shrubland (orange), herbaceous vegetation (yellow), herbaceous wetland (cyan), bare vegetation (grey), cropland (pink), built-up (red) and permanent water bodies (blue). This graphical scheme is a composition of figures found in [1] and [2].

rain. The European Copernicus programme, supported by the Sentinel satellite missions, provides an effective solution for mapping vegetation on global, regional, and local scales, with periodic and repeated observations. Since 2013, Sentinel-2 has continuously collected optical imagery, delivering high-resolution multispectral images (10–60 m) every 2 to 5 days, enabling comprehensive global monitoring. Other international initiatives, such as the Advanced Spaceborne Thermal Emission and Reflection Radiometer satellite (ASTER) deliver multispectral imagery with spatial resolutions ranging from 15 to 90 meters. These satellite's dual cameras enable the generation of a stereoscopic digital surface model of the earth with a nominal resolution of 25 meters per pixel.

These satellite platforms are widely used today for large-scale land cover mapping, incorporating multiple images and other data to support monitoring, mapping, and modelling activities. To do so, big data workflows are orchestrated and deployed in high-performance processing environments, which carry out tasks of massive data collection and preprocessing, classification strategies, stratification, integration of auxiliary data, optimization of processes and accuracy assessment. Classifying land cover over vast regions (involving several countries) requires methods that are both reliable and reproducible, presenting unique challenges that go beyond those of traditional large-area image classifications. The immense volume of unprocessed remote sensing data introduces the "four Vs" of big data: volume, variety, velocity, and veracity, which are recognized as its core challenges.

In order to identify effective techniques for mapping land cover patterns, the remote sensing community has explored various approaches, where ML supervised methods for classification are commonly employed. Among others, Random Forest (RF) has been shown to perform efficiently in land cover mapping, thanks to its ability to manage high-dimensional data and multicollinearity, while being both fast and resistant to overfitting.

In the context of the previously presented initiatives, a methodology to streamline Big Data workflows for land cover classification over extensive areas has been implemented [1]. As a use case for validation, the complete Mediterranean Basin is mapped, which is covered by more than 450 Sentinel-2 tiles and around 1200 ASTER tiles. Three seasonal datasets from Sentinel-2, along with several derived products from both Sentinel-2 and ASTER are used, which involve more than 4TBs of information thqt considerably increase the computational demands for processing. To tackle this issue, several specific parts of the workflow have been parallelized using Dask, a tool designed to optimise algorithm execution.

The overall approach to big data storage and management is illustrated in Figure 1, along with metadata protocols. The first phase comprises data collection and distributed processing at tile level. Resulting data are then stored in a MinIO S3-compatible object storage system, while metadata are stored in a NoSQL MongoDB database for efficient indexing and querying. After this, data analysis and validation are performed, which implies high-quality training set generation, ML model generation and test. This last phase is refined until reaching an accuracy higher than 90%, so land-cover maps are obtained with low error rates and identified labels: Closed and open forests (green and olive), shrubland (orange), herbaceous vegetation (yellow), herbaceous wetland (cyan), bare vegetation (grey), cropland (pink), built-up (red) and permanent water bodies (blue) (Figure 1, right). Final land-cover maps are then available for visualisation in WebGis visualisation services.

Another key component in this knowledge-driven approach involves providing human experts with domain knowledge representation, supported by data standardisation and semantic integration of sources. To this end, the use of ontologies and semantic web technologies have shown high success in knowledge representation in many fields, in which the earth observation is not an exception. This is tackled in this project by the RESEO [2] ontology, which considers the special nature and structure of different satellite and airborne data products. It is implemented in standard OWL 2, according to which, an RDF repository has been generated to allow advanced SPARQL querying. This component allows the integration, reasoning and linking of heterogeneous data, such as meteorology or historical crop records and land use. This will enhance the construction of advanced on-top applications for future experts.

Links:

[L1] https://khaos.uma.es/green-senti
[L2] https://enbic2lab.uma.es/
[L3] https://itis.uma.es/

References:

[1] A. M. Burgueño, J. F. Aldana-Martín, M. Vázquez-Pendón, et al., "Scalable approach for high-resolution land cover: a case study in the Mediterranean Basin," Journal of Big Data, vol. 10, p. 91, 2023. https://doi.org/10.1186/s40537-023-00770-z

[2] J. F. Aldana-Martín, J. García-Nieto, M. M. Roldán García, and J. Aldana Montes, "Semantic modelling of Earth observation remote sensing," Expert Systems with Applications, vol. 187, p. 115838, 2021. https://doi.org/10.1016/j.eswa.2021.115838

Please contact:

Virginia García Millán (ITIS, University of Málaga, Spain)
virginia.garcia@uma.es

José García-Nieto (ITIS, University of Málaga, Spain)
jnieto@uma.es

Inferring Contributions in Privacy-Preserving Federated Learning

by Balázs Pejó (Budapest University of Technology and Economics) and Delio Jaramillo Velez (Chalmers University of Technology)

To what extent do individual contributions enhance the overall outcome of collaborative work? This question naturally arises across scientific fields and is particularly challenging in Federated Learning. It remains largely unexplored in privacy-preserving settings where individual actions are concealed with techniques like Secure Aggregation.

Federated Learning (FL) [1] enables multiple parties to develop a machine learning model collaboratively without sharing their confidential training data. For instance, millions of mobile devices can collectively train a predictive text model without exposing their personal texts, or multiple hospitals can jointly train a model to predict various health-related risk scores without revealing specific patients' sensitive medical records. Unlike centralized learning, where the training data from all participants are collected by a trusted entity, FL only involves the exchange of model parameter updates.

As such, FL comes with some built-in privacy protection. Yet, information could still leak through the model updates, which require additional techniques to reduce the privacy risk. Secure Aggregation (SA) [2] is a frequently used privacy-preserving mechanism that hides individual model updates via a lightweight cryptographic protocol. On the other hand, advanced privacy protection, such as SA, also helps malicious parties to remain undetected within the federation. For instance, adversarial (Byzantine) participants can degrade model performance stealthily through (data of model) data poisoning attacks due to the utilized privacy-enhancing technology (PET). Moreover, unintentional manipulations can even occur, such as when participants have biased or noisy training data they are unaware of.

To tackle these issues, the usefulness of individuals should be determined. Contribution Evaluation (CE) schemes allow participants to assess each other's value, thus, helping to identify potential harmful actors that could degrade the model's performance. The Shapley value (SV) [3] is a prominent candidate for CE, as it considers the marginal contributions of the participant to all possible coalitions. However, the SV is computationally demanding, thus, it is not feasible for large models or big datasets. Consequently, numerous approximation methods exist to relax its computational demands. On the other hand, they also rely on marginal contributions, which is incompatible with PETs such as SA. Indeed, there is a fundamental tension between them: privacy, in general, aims to conceal individual-specific information, while CE, conversely, seeks to obtain the individual's usefulness (e.g., measured in data quality).

A possible solution to this problem was envisioned in [4], where the marginal contributions based on the coalitions with the two extreme cardinalities are considered. These are the grand coalition (when everybody trains, i.e., the output of SA) and the empty coalition (when no one trains, i.e., the untrained model). These can be combined with individual coalitions, which are available to the participants but not to others due to SA.

Following the notation in Table 1, these scores are denoted as *Include-One-In (I1I)* and *Leave-One-Out (L1O)* and computed as follows for the *i*th participant:

$$I1I_i = \text{eval}(M_0 + U_i) - \text{eval}(M_0)$$

$$L1O_i = \text{eval}(M) - \text{eval}(M - U_i)$$

A shortcoming of this approach is the centralized nature of the evaluation: the marginal contributions for I1I and L1O are based on individual local updates, but they are measured via a global, pre-agreed test set. In reality, the participants' data might come from different distributions, hence, inconsistencies might arise, like disagreements on contribution scores: a model assessed as 'good' for someone might be 'bad' from another's point of view. Therefore, one's contribution should not be determined by themselves, as that would lead to misaligned incentives and introduce bias into the assessment. Instead, the participants should be evaluated by everybody else.

Symbol	Description
M_0	Untrained model
U_i	Model update of participant *i*
M	Trained aggregated model (i.e., $M = M_0 + \text{Sum}_i[U_i]$)
eval()	Evaluation function (based on a global shared test set)
$\text{eval}_i()$	Evaluation on *i*'s test data

Table 1: Notations.

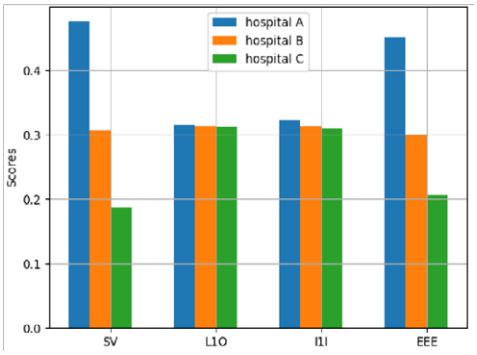


Figure 1: CE scores for three FL clients. The model is logistic regression, the dataset is Breast Cancer Wisconsin with added Gaussian noises: σ is 0.1, 0.3, and 0.5 for the clients, respectively.

Our proposed scoring algorithm is called Evaluate-Everyone-Else (EEE), as all participants evaluate each other except themselves. More precisely, they evaluate the others together as one instead of individually, which is prevented by SA. Formally, EEE for participant i is the sum of all the marginal contributions of everybody else from all participants' points of view, as shown below.

$$EEE_j = Sumi_{\neq j} [\ evali\ (\ M\ ) - eval_i\ (\ M_0 + U_i\ )\ ]$$

For example, if there are three hospitals (named *A, B,* and $_C$) performing FL together with SA, the *EEE* score of *A* can be evaluated by *B* and *C* as follows:

- $EEE_A = [\ eval_B\ (\ M\ ) - eval_B\ (\ M_0 + U_B\ )\ ] + [\ eval_C\ (\ M\ ) - eval_C\ (\ M_0 + U_C\ )\ ]$
- $EEE_A = [eval_B\ (M0 + UA + UB + UC) - eval_B\ (M_0 + U_B)] + [eval_C\ (M_0 + U_A + U_B + U_C) - eval_C\ (M_0 + U_{C)}]$
- EEE_A=[Marginal cont. of {A,C} according to B]+[Marginal cont. of {A,B} according to C]

Thus, the *EEE* score for *A* is the sum of the local views of *B* and *C* about everybody else. Although this score encompasses others' contributions explicitly, they are counted fewer times as the participant score is assigned to: EEE_A is influenced by *{A,C}* and *{A,B}*, so the impact of *B* and *C* are also included, their influence is less than that of the target *A*. To visualize the exact values for the envisioned scenario, we split a medical dataset into three (assigned to *A, B,* and *C*) and artificially injected different amounts of noise into them. In Figure 1, we show the baseline *SV*, the privacy-preserving self-evaluation methods *111* and *L1O*, and the proposed (both privacy- and incentive-aware) *EEE*. It is visible that our approach approximates the ground truth *SV* much better.

Our proposed technique is the first privacy-friendly contribution evaluation technique that mitigates the rising selfish incentives concerning self-evaluation. Furthermore, the evaluation is completely distributed as it does not require any common representative dataset of the participants for a coherent evaluation.

References:

[1] P. Kairouz, et al., "Advances and open problems in federated learning". Foundations and trends in machine learning, 2021.
[2] M. Mansouri, et al., "Sok: SA based on cryptographic schemes for FL". Proceedings on Privacy Enhancing Technologies, 2023.
[3] A. E. Roth, "The Shapley Value: Essays in honor of Lloyd S. Shapley". Cambridge University Press, 1988.
[4] B. Pejó, et al., "Measuring contributions in privacy-preserving federated learning". ERCIM News 126, 2021

Please contact:

Balázs Pejó;
CrySyS Lab, HIT, VIK, BME, Hungary
pejo@crysys.hu

Breaking the Silence: Brain-to-Speech Innovations

by Mohammed Salah Al-Radhi and Géza Németh
(Budapest University of Technology and Economics, TMIT-VIK, Budapest, Hungary)

How can brain activity be turned into clear, intelligible speech? An ambitious research initiative in Hungary is addressing this question by developing cutting-edge methods to decode neural signals into speech, aiming to restore communication for individuals with severe speech disorders.

Speech disorders caused by neurological conditions can shatter lives, cutting individuals off from their loved ones and the world. A pioneering Hungarian project (EKÖP-24-4-II-BME-197), funded by NKFI [L1] and ENFIELD [L2], is tackling this crisis by developing revolutionary brain-to-speech technologies, aiming to restore voices to those who have lost them.

The project centres on decoding speech envelopes—patterns of neural activity that convey essential information about speech articulation. These envelopes represent the rhythm and amplitude of speech, capturing the dynamic features necessary for intelligible and natural communication. By using state-of-the-art signal processing and advanced machine learning techniques, the research team is unlocking the potential of brain-computer interfaces for real-time communication restoration. The goal is ambitious but deeply impactful: to enable individuals with severe speech disorders to express themselves again.

Our team has introduced novel prosody-aware feature engineering methods and a transformer-based speech synthesis model to enhance Brain-to-Speech reconstruction. The developed pipeline begins with preprocessing raw neural data (e.g., EEG) using wavelet denoising and time-frequency analysis. This ensures the preservation of critical neural features such as intonation, pitch, and rhythm—elements essential for natural speech synthesis [1]. Unlike traditional pipelines, our approach prioritizes prosodic features, improving the emotional and expressive quality of reconstructed speech. At the core of the system lies our transformer encoder architecture, specifically designed to decode speech envelopes from neural signals. This model integrates attention mechanisms [2], enabling it to dynamically focus on the most relevant neural patterns. By incorporating prosodic features into the decoding process, the model achieves superior intelligibility and expressiveness compared to baseline methods such as bidirectional RNNs and Griffin-Lim vocoders [3]. Once the speech envelope is decoded, the final step involves synthesizing synthesising the predicted speech signals into intelligible and natural-sounding audio. This is achieved through cutting-edge neural vocoders such as AutoVocoder and BigVGAN [L3]. These vocoders are engineered to convert the decoded speech envelopes into high-fidelity waveforms with precise control over pitch, tone, rhythm, and even emotional expressiveness.

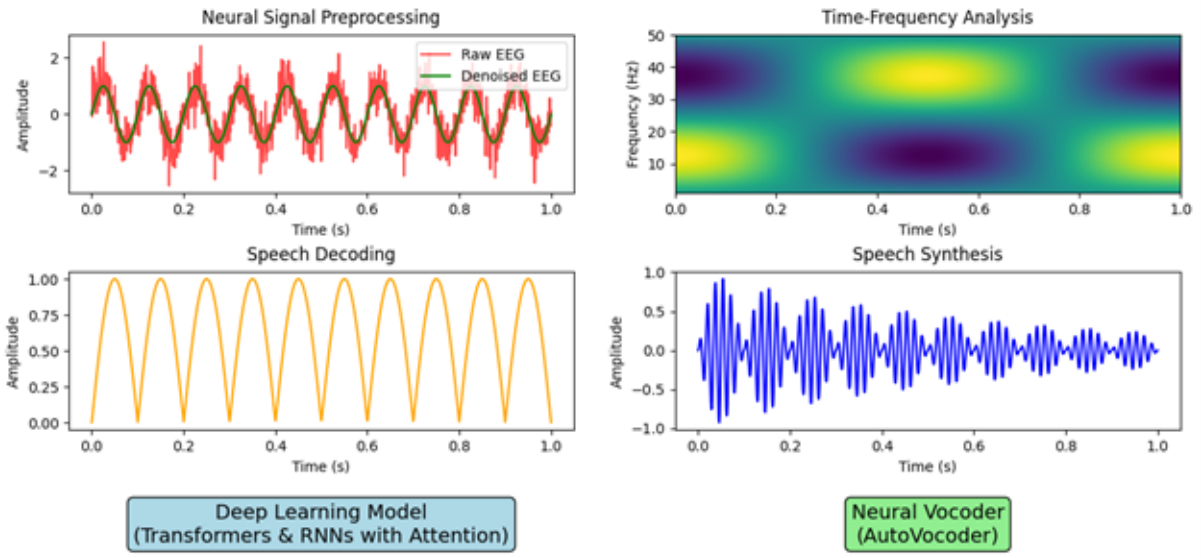


Figure 1: The neural decoding and speech synthesis pipeline. The figure illustrates the process of transforming raw neural signals (EEG) into intelligible speech. Starting with raw and denoised EEG signals, the system performs time-frequency analysis to extract key neural features. These features are then decoded using advanced deep learning models (Transformers and RNNs with attention mechanisms), which predict speech envelopes. Finally, the predicted speech envelopes are synthesized into audible speech using neural vocoders, ensuring both accuracy and expressivity.

Figure 1 illustrates the system's journey from raw neural signals to intelligible speech. It highlights the sequential processes of signal preprocessing, time-frequency analysis, neural decoding, and speech synthesis. The visualization underscores the complex pipeline that transforms noisy EEG recordings into clear, meaningful speech. By leveraging these advanced synthesis tools, the system not only achieves technical accuracy but also ensures that the reconstructed speech conveys the nuances necessary for effective and engaging communication. This technology offers a lifeline for individuals affected by ALS (amyotrophic lateral sclerosis), strokes, traumatic brain injuries, or other conditions that impair speech. It restores their ability to communicate, bringing autonomy and connection back into their lives. For individuals unable to speak due to neurological disorders, even a few words can rebuild bridges to loved ones, healthcare providers, and society.

From a scientific perspective, the project breaks new ground in neuroprosthetics and brain-computer interfaces (BCIs), providing a foundation for future assistive technologies. The methodologies developed here could extend to broader applications, such as neural control of robotics and virtual assistants. However, significant challenges remain on the path to bringing this technology into everyday use. One key difficulty is making the system work in real-time. While the current models are highly accurate, they require substantial computing power, which makes instant processing difficult. The team is refining these models to run faster and more efficiently without losing accuracy, a crucial step for practical applications. Another challenge lies in tailoring the system to each user. Everyone's brain is unique, with differences in structure, neural activity, and how speech disorders affect them. To address this, the researchers are developing ways for the system to "learn" and adapt to each individual's brain patterns. This personalized approach will ensure that the technology works well for people of all backgrounds and needs.

The vision for the future is a brain-to-speech system that is as easy to use as current assistive devices, enabling people to speak their thoughts naturally and effortlessly. By combining innovations in neuroscience, artificial intelligence, and signal processing, this project is breaking down the barriers of silence, offering a voice to those who need it most.

Links:

[L1] https://www.bme.hu/EKOP
[L2] https://www.enfield-project.eu/
[L3] https://github.com/NVIDIA/BigVGAN

References:

[1] Verwoert, M., Ottenhoff, M.C., Goulis, S. et al., "Dataset of Speech Production in intracranial Electroencephalography," Nature Scientific Data, 9, 434,2022. https://doi.org/10.1038/s41597-022-01542-9
[2] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes", Neural Comput & Applic 34, 13371–13385,2022. https://doi.org/10.1007/s00521-022-07366-3
[3] M.S. Al-Radhi, G. Németh, "Brain-to-Speech: Prosody Feature Engineering and Transformer-Based Reconstruction", Book Chapter: Artificial Intelligence, Data and Robotics: Foundations, Transformations, and Future Directions, under preparation, 2025.

Please contact:

Mohammed Salah Al-Radhi,
Budapest University of Technology and Economics,
Budapest, Hungary
mohammed.alradhi@vik.bme.hu

Anomaly Detection in Telemonitoring Using Sensor Correlation

by Beatrix Koltai, Gergely Ács, and András Gazdag
(Budapest University of Technology and Economics)

In telemonitoring, false alarms from medical devices can overwhelm doctors and desensitize care teams to critical issues. How could we reliably detect subtle yet critical changes in a patient's health status without false alarms or missed anomalies? By correlating data across multiple sensors, our solution improves detection accuracy and results in fewer false positives. Utilizing federated learning, our model is collaboratively trained across multiple hospitals, each with potentially limited data. This improves the model's performance without centralising sensitive patient data.

In telemonitoring, remote medical devices gather sensitive clinical data to detect abnormal changes, potentially life-threatening indicators of a sudden shift in a patient's health status, known as anomalies. These data could be used to train machine learning models to solve this anomaly detection task. We applied a similar approach within the SECURED project. [L1] While traditional solutions trigger alarms based on changes in individual sensor readings, our approach leverages correlations between multiple sensor data streams.

This approach results in fewer false positives and false negatives, because it detects patterns that might be missed when relying on a single data stream. By correlating these signals across sensors, the system can also identify the warning signs sooner.

Anomaly detection exploiting multi-sensor data correlation

Our approach, which applies a similar concept presented in our previous paper [3], consists of two interconnected models:

1 Forecasting Model: A time series forecasting model that predicts the subsequent sensor measurement based on prior readings of all the correlating sensors.
2. Detection Model: Anomalies are flagged by the detection model when the difference between the predicted and actual values of any sensor exceeds a predefined detection threshold.

In other words, an anomaly is detected if any sensor's measurement is "unpredictable" based on prior measurements, using a model trained on data predominantly from healthy patients (see Figure 1).

The forecasting model is a Temporal Convolutional Network (TCN) [1] combined with a linear output layer that simultaneously predicts the next measurement of all sensors from their previous measurements while exploiting correlations across sensors. Convolutional networks, which form the basis of a TCN, are highly effective for structured data such as images and time series that exhibit local spatial correlation.

Two-dimensional convolutional filters can capture such local dependencies across different sensor signals, extracting meaningful features for predicting future measurements. By stacking multiple convolutional layers, these networks can progressively model broader, global dependencies in the data capturing complex patterns over larger contexts. TCNs employ dilated convolutions to model both short-range long-range dependencies in a sequence while maintaining efficiency and parallelism in training [2].

Dataset preparation

After cleaning and anonymizing an initial data set from medical devices taken in a real environment, the dataset consisted of five time series of eleven patients, each from a different device. Such a time series corresponds to one sensor's readings.

Measurements were normalized and resampled to one-second intervals by keeping the very first measurement of every second, ensuring uniformity across data streams.

The optimal time range for each patient, where data coverage across sensors is highest, is identified and retained. Training samples are then generated using a three-minute sliding window, with the final value in each window as the prediction tar-

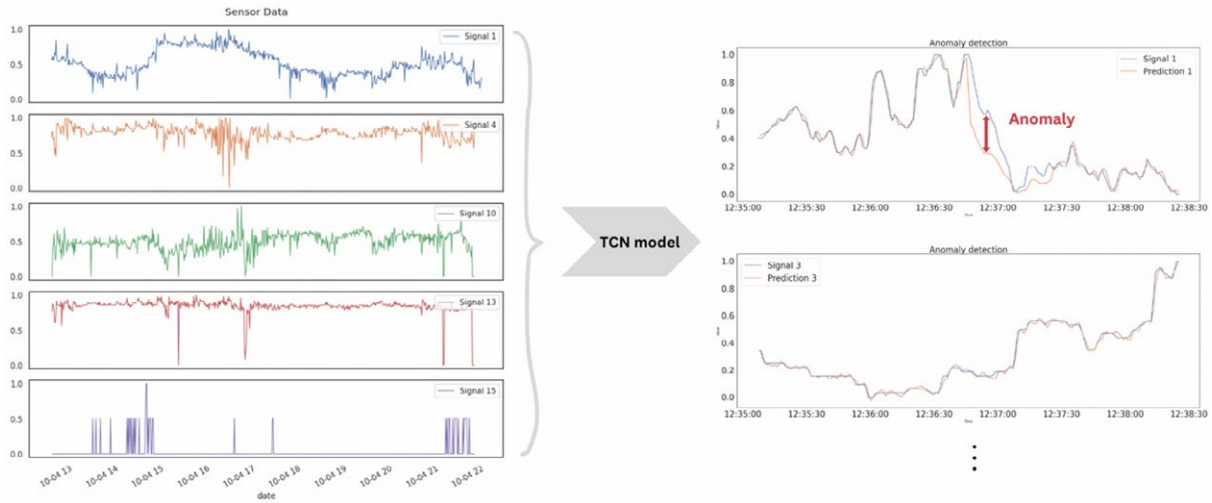


Figure 1: Illustration of the combined prediction of sensor data and anomaly detection when the measurement of one sensor is significantly different from its prediction.

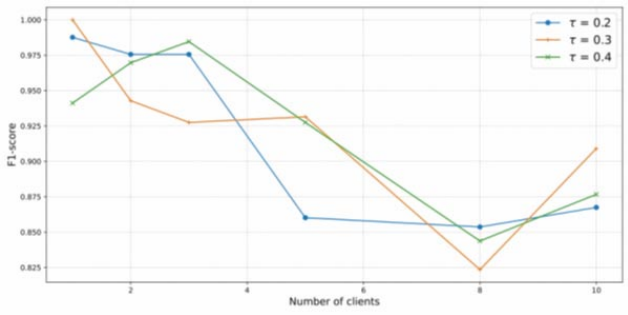


Figure 2 - Illustration of the F1-score results of the detection for different numbers of clients, with different ground-truth thresholds.

get. Windows lacking sensor data are discarded, while gaps in partially missing data are filled using linear interpolation.

Federated learning

In practice, hospitals often lack sufficient data to train accurate predictive models and are reluctant to share it due to GDPR concerns. As a result, centralized training of an anomaly detection model using shared data is not feasible.

Federated Learning (FL) offers a solution by training models collaboratively without directly sharing training data. Since hospitals can access only the shared model and not each other's data, the risk of unintended information leakage is reduced, while achieving better accuracy than training models on local data alone.

To showcase the viability of federated learning for anomaly detection in telemonitoring, the forecasting model is trained by federated learning.

It is simulated by partitioning the preprocessed dataset randomly among clients in an independent and identically distributed (IID) setting. Each client participates in every federated round, performing one local training epoch on the shared forecasting model. For demonstration, the server calibrates the threshold for the detection model using held-out data in a supervised manner. A measurement value is labeled as anomalous if the difference from its preceding value exceeds a predefined ground-truth value, tau (τ). For each sensor, a distinct detection threshold is calibrated to maximize the F1-score on the labelled held-out data.

After calibration, the detection model, along with the collaboratively trained forecasting model, is tested on a separate dataset that was not used for either calibration or training.

Results and future work

We compare FL with centralized training, when one client has all the data. Figure 2 shows that the F1-score for anomaly detection approaches 1 when the number of clients is fewer than three and remains above 0.8 even with 10 clients, where centralized training corresponds to a single-client scenario. This indicates that Federated Learning (FL) incurs only a 10% degradation in detection accuracy. These findings demonstrate that FL is a promising approach for anomaly detection in telemonitoring. However, further evaluation with larger datasets and non-IID settings is needed to confirm its robustness and generalisability.

Link:

[L1] https://secured-project.eu

References:

[1] Y. Liu et al., "Time series prediction based on temporal convolutional network," in Proc. IEEE/ACIS 18th International Conference on Information Systems (ICIS), 2019, doi: 10.1109/ICIS46139.2019.8940265.

[2] Y. He and J. Zhao, "Temporal convolutional networks for anomaly detection in time series," J. Phys.: Conf. Ser., vol. 1213, p. 042050, 2019.

[3] B. Koltai, A. Gazdag and G. Ács, "Supporting CAN Bus Anomaly Detection with Correlation Data", in Proc. of the 10th Int. Conf. on Information Systems Security and Privacy ICISSP; 2024, doi: 10.5220/0012360400003648

Please contact:

Beatrix Koltai
CrySyS Lab, HIT, VIK, BME, Hungary
bkoltai (at) crysys.hu

Data Visualisation for Big Data: Digital Epidemiology

by Stelios Zimeras (University of the Aegean)

At the University of the Aegean, we are developing advanced visualisation techniques and innovative algorithms to enhance digital epidemiology, enabling more effective disease monitoring and response through the integration of diverse big data sources.

The management of big data is an important but also demanding process where important decisions must be made. An essential role in the data processing process is both the collection and the organization of data in such a way that they efficiently bring about the best results.

Characteristics of big data are their diversity, speed of production and renewal. Based on the above, we are led to the development of analysis and visualisation techniques in order to analyse both the relationships and the structures set included in each dataset. Due to the multitude of features that govern big data, it is difficult to visualise and work with all of them. This has created the need to develop algorithms that help process and visualise multidimensional data.

New algorithmic techniques have enabled the creation of new tools for data processing, such as artificial intelligence, machine learning, and natural language processing. All of this has contributed to the emergence of a new branch of epidemiology called digital epidemiology [1]. Digital epidemiology embraces the goals of clinical epidemiology but takes a different approach to their implementation. Instead of relying solely on data from the health sector, it makes use of these alternative data sources. These new data sources, such as social media, are also called "big data" sources and are characterised by very large volumes of data that are complex in structure and highly heterogeneous [2]. Thus, one can search for information that

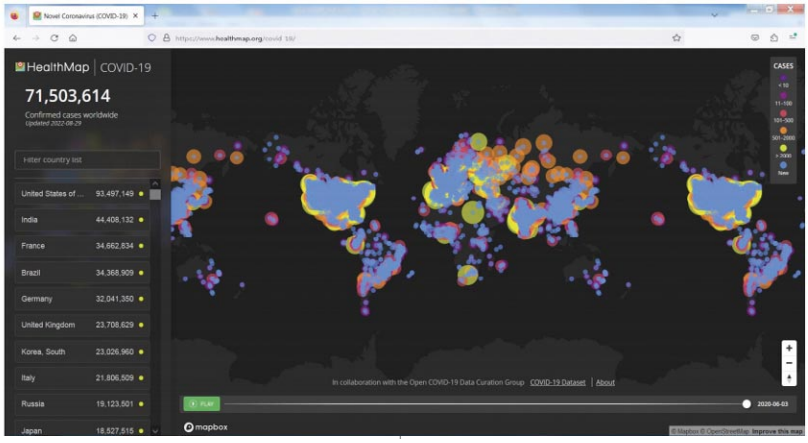


Figure 1: HealthMap's interface with disease outbreaks around the world.

may be related to a disease. Therefore, the electronic search data can provide information about the spread of a disease in a place when the electronic searches by the residents are largely related to the health, diseases and medical personnel of the area. This development was accompanied by several implementations in this field. Some of them are [3]:

HealthMap [L1] is a non-profit implementation that aims to collect and visualise important information related to public health risks and, basically, information about impending or evolving epidemics, with the aim of informing the human population. The HealthMap interface consists of an interactive world map, which provides information on any public health risks in each region, but also in the world in total. The system also detects the location of each user and, initially, provides information for the area near the user, thus achieving the correct information of the user on urgent issues that directly concern them (see Figure 1).

BioCaster [L2] is a subscription-based, non-governmental, free-to-use, research-based data stream monitoring program from various sources with the aim of safeguarding public health. The aim is to identify new outbreaks of diseases and monitor their spread. To display confirmed events, BioCaster uses a Google Maps-based interface in which it displays confirmed events based on their

Figure 2: BioCaster Interface.

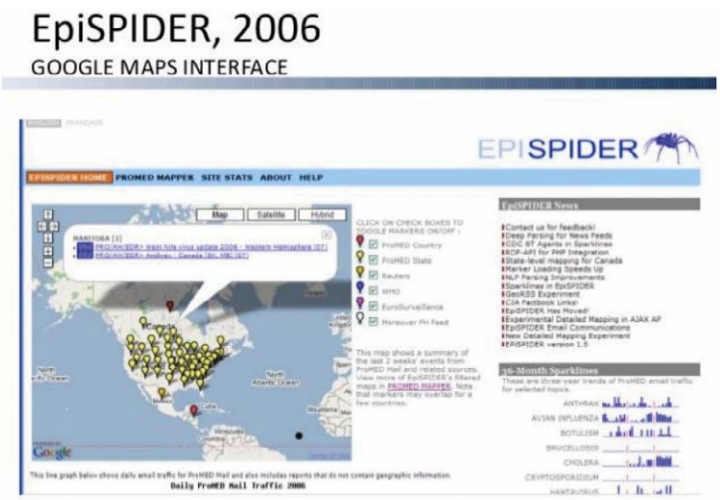


Figure 3: EpiSPIDER Interface.

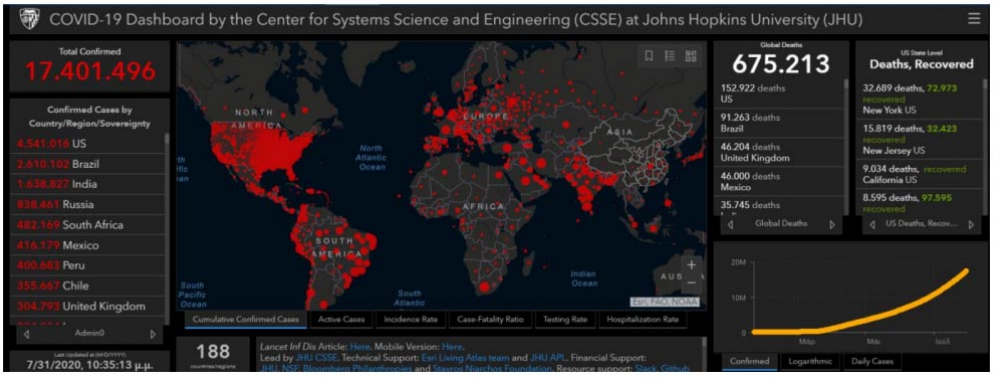


Figure 4: COVID-19 Dashboard Johns Hopkins University Interface.

geographical location, while it is also possible to update its subscribers via e-mail messages. (see Figure 2).

EpiSPIDER - Semantic Processing and Integration of Distributed Electronic Resources for Epidemics and Disasters [L3] is an automated program for collecting information on human and animal diseases, while also collecting information on natural disasters. Also, as complementary data, demographic information and information related to public health are collected. The system then processes, analyzes and automatically classifies the data into various groups. This process is enhanced by the use of a natural language processing system (see Figure 3).

The JHU COVID-19 Dashboard [L4] is a platform for recording and visualizing confirmed information about the spread of the coronavirus in various regions of the world. The application is part of the Coronavirus Resource Center, Johns Hopkins University, which collects information about the virus and its spread for various purposes. The COVID-19 Dashboard features an interactive global map that displays information on the total number of cases and deaths due to the coronavirus in various regions of the world. The map is also accompanied by graphs that show the growth of cases of the virus, and by figures that relate to cases and deaths worldwide and for specific countries (see Figure 4).

Links:
[L1] https://healthmap.org/en/
[L2] https://earlywarning.wordpress.com/2009/02/14/detecting-rumors-with-web-based-text-mining-system/
[L3] https://www.slideshare.net/hermantolentino/2009-epispider-cdc-gis-day
[L4] https://coronavirus.jhu.edu/map.html

References:
[1] G. Lippi, C. Mattiuzzi, and G. Cervellin, "Is Digital Epidemiology the Future of Clinical Epidemiology?", J. Epidemiol. Glob. Health, vol. 9, no. 2, p. 146, 2019.
[2] A. H. Ali and M. Z. Abdullah, "A survey on vertical and horizontal scaling platforms for big data analytics," Int. J. Integr. Eng., vol. 11, no. 6, pp. 138–150, 2019.
[3] A. Gribas, "Big Data Processing in Digital Epidemiology," M.Sc. Thesis, EMP, 2020.

Please contact:
Stelios Zimeras
University of the Aegean, Greece
zimste@aegean.gr

Call for Participation

SAFECOMP 2025 and DECSoS Workshop

Stockholm 9-12 September 2025

The international SAFECOMP Conference takes part this year in Stockholm, Sweden, at KTH, from 9-12 September 2025, the first day being reserved for several parallel workshops.

SAFECOMP has contributed to the progress of the state-of-the-art in dependable application of computers in safety-related and safety-critical systems. SAFECOMP is an annual event covering the state-of-the-art, experience and new trends in the areas of safety, security and reliability of critical computer applications. SAFECOMP offers a platform for knowledge and technology transfer between academia, industry, research institutions and standardization bodies. It provides ample opportunity to exchange insights and experience on emerging methods, approaches, and practical solutions. It is a single-track conference allowing easy networking. Proceedings are published in the Springer LNCS series.

Important Dates (Main Conference):
- Workshop proposal submission: 7 February 2025
- Abstract submission: 7 February 2025
- Full paper submission: 14 February 2025 (an extension is expected)
- Notification of acceptance: 12 April 2025
- Camera-ready submission: 1 June 2025

DECSoS Workshop

Of particular interest for the ERCIM community will be the co-located 20th DECSoS Workshop, co-organized since many years by the ERCIM DES (Dependable Embedded SW-intensive Systems) Working Group (Erwin Schoitsch, Amund Skavhaug). The workshop proposals are underlying an evaluation process, afterwards separate "Call for Papers" will be announced in March 2025 for each workshop. All papers for the workshops will be peer reviewed by at least three independent reviewers and published by Springer, LNCS series, as SAFECOMP Workshop Proceedings, separate from the Proceedings of the main conference.

More information:
https://safecomp2025.se/

IDIMT 2025 - 33rd Interdisciplinary Information Management Talks

Hradec Kralove, Czech Republic
3-5 September 2025

"ICT in Business: AI Everywhere? Glory and Disgrace of AI"

The 2025 edition of the IDIMT conference- an interdisciplinary forum for the exchange of concepts and visions in the area of software intensive systems - is organized by the University of Economics and Business in Prague and JKU in Linz, Austria. Papers are peer reviewed and indexed by Scopus and Web of Science.

The conference is organized in ten sessions with several sessions of particular interest to the ERCIM community, such as "AI and Autonomous Systems". The session is co-organized by Erwin Schoitsch and Abdelkader Shaaban, AIT, Austria.

More information: https://idimt.org/

A Cost-Benefit Analysis of Additive Urban Manufacturing

by Igor Ivkić (University of Applied Sciences Burgenland, AT | Lancaster University, UK), Burkhard List (b&mi GmbH & Co KG)

Traditional manufacturing means that a product is mass-produced in distant countries and then shipped long distances to customers, leaving a very large carbon footprint. In this article, we present an approach to disrupt these traditional value chains and replace them with urban manufacturing (or local production) using three-dimensional (3D) printing technology. This approach allows products to be printed locally in an environmentally friendly way, rather than being manufactured far away and flown in. At the heart of this idea is a cloud-based Manufacturing as a Service (MaaS) platform [1] that manages the entire process from online purchase to 3D printing, promoting sustainability and strengthening local economies.

In many areas, the industrial manufacture of products is characterised by resource-intensive production methods, long logistics chains, and ever-increasing over-production beyond actual demand [2]. Typically, products are conceived, designed, and developed in first-world countries, only to be mass-produced in large quantities in low-wage countries. This type of industrial production is neither environmentally friendly nor sustainable, nor does it help to strengthen local economies. Studies have shown that 27% of the environmental impact is due to the transport of these goods, and that these goods spend 90% of their production time in storage or in transit (unproductive). Furthermore, the Covid-19 pandemic has shown how dependent our society is on "offshore production" [L1] from developping countries, especially when the associated supply chain [3] is not available or functioning as usual. Another effect since the pandemic has been the conscious preference of customers for fair, regional, and sustainable products [L2].

To meet the new trend of local and sustainable purchasing in manufacturing, without resorting to "offshore production" and long supply chains, a new approach is needed [L3]. From the customer's perspective, products should be produced locally and "on-demand". Due to the technological advances in additive manufacturing, a wide range of products can now be produced both locally and cost-effectively using 3D printing technologies. At the heart of the MaaS approach is a Cloud Crafting Platform (CCP) that allows 3D printer operators to integrate their 3D printers online and offer 3D printing as a service. With the help of this urban CCP platform that enables MaaS, any 3D printer operator can become a local "on-demand" producer, ststrengthening the local economy and giving customers the opportunity to buy products made in their nearby urban area. Another goal of the MaaS platform is to break down traditional supply chains and manufacture products where they are purchased. This eliminates long transport routes and positively impacts the environment.

The CCP follows a serverless architecture approach based on the Function as a Service (FaaS) cloud service model, allowing both web shops and 3D printer operators to be integrated. Figure 1 shows how the CCP connects the web shop (Point of Sale) and the 3D printer operator (Point of Manufacturing). The production process is only initiated when a customer purchases a product from a web shop and selects the option to have the product manufactured (or 3D printed) locally. The CCP thus connects the web shop (Point of Sale) with the 3D printer operators (Point of Manufacturing), enabling on-demand production as opposed to traditional mass production of products. The CCP is designed to be scalable, allowing multiple web shops and 3D printer operators to be integrated.

However, the platform is not limited to 3D printer integration alone; it identifies additive manufacturing using 3D printing as the first technological opportunity for on-demand production. Subsequently, additional technologies such as Computerized Numerical Control (CNC), laser cutting, plotter cutting, robotics, and augmented reality could be integrated to expand the range of what and how products could be produced.

The CCP idea has also been published in a first position paper [1], in which we describe both the MaaS approach and its technical implementation via the cloud. Furthermore, we proposed a cost-benefit analysis including metrics [1]. The focus of the position paper was to describe the architectural building blocks of the proposed CCP, including a description of a lab environment, where the MaaS approach is divided into three zones with different responsibilities. Zone 1 contains the "Point of Purchase," where a product is purchased by a customer. This could be a web shop connected to a CCP in Zone 2. This zone operates in the cloud and acts as a link between the customer who purchases a product (Zone 1) and the 3D printer operator who manufactures the product using 3D printers (Zone 3). The CCP in Zone 2 provides the necessary interfaces to connect both web shops and 3D printer operators, thus enabling production based on a MaaS approach. Figure 2 shows the lab environment described:

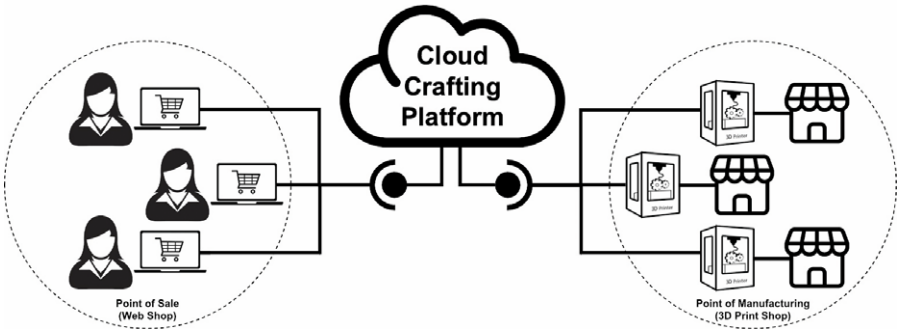


Figure 1: A purchased product triggers the on-demand manufacturing process by sending the order to a local 3D printer operator via the CCP (adapted from [1]).

In addition to the laboratory setup, the following metrics were used in the cost-benefit analysis, where a specific product (a ring) was produced simultaneously using three

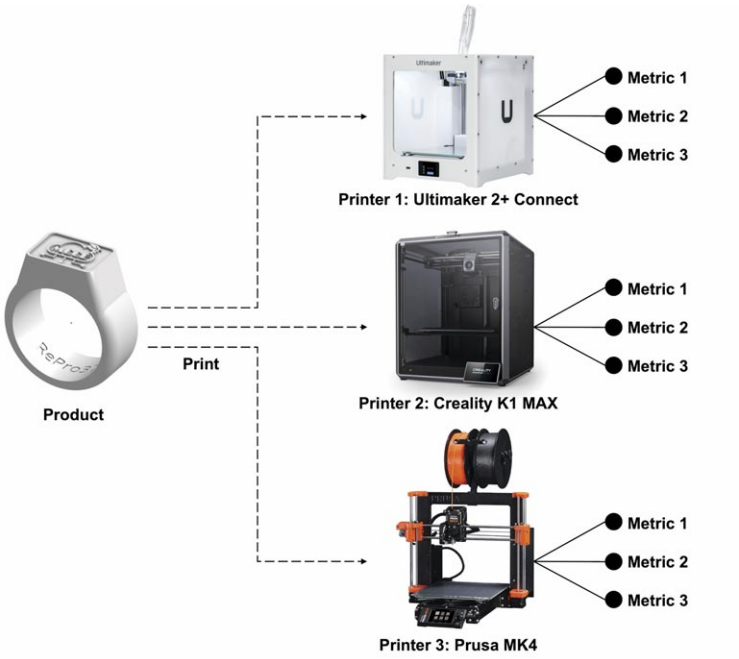


Figure 2: Lab Environment for the Cost-Benefit Analysis of the MaaS Approach using the CCP (adapted from [1]).

different 3D printers, and the production costs were calculated based on a 1-year operational simulation in a Multi-Level Constribution Margin Accountig (CMA) scheme:

- Metric 1: Print time per 3D print job
- Metric 2: Material usage per 3D print job
- Metric 3: Power consumption/energy cost per 3D printer
- Metric 4: Human operation time per 3D print job (post-processing time).

These metrics allow for a comparison to be made between the three 3D printers to determine which require more or less time, material, and energy (and ultimately money) to produce the same product. The lab experiment showed that the selected product could be 3D printed at a cost of €3.96 (Ultimaker 2+ Connect), €2.41 (Prusa MK4) and €1.18 (Creality MK1 MAX) in a 1-year simulation.

These results and projections include direct labour, electricity, material costs and CCP-commission (costs) at CM I level, rent and depreciation at CM II level, local IT-infrastructure at CM III level and general overheads and profit towards an extrapolated operating profit.

In summary, by providing services for the integration of web shops (point of sale) with 3D print shops (point of manufacturing), the CCP is at the heart of future urban manufacturing. This enables both individuals and local small and medium-sized enterprises (SMEs) to become urban on-demand manufacturers of products purchased online. Similar to the Uber ecosystem, where anyone with a driver's licence and a car can become an Uber driver, the CCP helps to connect the web shops with the local 3D production sites. In this new CCP ecosystem, the customer has the power not only to choose a product, but also to trigger on-demand production after the purchase (as opposed to mass production in traditional manufacturing), thus strengthening the local economy.

Total	Revenues
-	Variable Costs
=	Contribution Margin I (CM I)
-	Product Fixed Costs (Costs directly attributed to a product)
=	Contribution Margin II (CM II)
-	Division Fixed Costs (Costs attributable to a specific business division)
=	Contribution Margin III (CM III)
-	Company Fixed Costs (Costs that are not attributed to a specific division but apply to the entire company)
=	Operating Profit

Table 1: Multi-Level Contribution Margin Accounting Scheme used to calculate Production Costs based on a 1-Year Simulation.

Links:

[L1] https://kwz.me/hFK
[L2] https://kwz.me/hFS
[L3] https://youtu.be/elPRPR_oLmQ?si=W7Js0picVlWpfrcz

References:

[1] I. Ivkić, et al., "Towards a Cost-Benefit Analysis of Additive Manufacturing as a Service", in Proc. of the 14th Int. Conf. on Cloud Computing and Services Science - CLOSER; ISBN 978-989-758-701-6; ISSN 2184-5042, SciTePress, pages 338-345. DOI: 10.5220/0012733500003711

[2] E. Westkämper, C. Löffler, "Strategien der Produktion: Technologien, Konzepte und Wege in die Praxis", Springer Vieweg, 2016.

[3] S. Chopra, P. Meindl, "Supply chain management", Strategy, planning & operation (pp. 265-275), Gabler, 2007.

Please contact:

Igor Ivkić
Univerisity of Applied. Sciences Burgenland, AT and Lancaster University, UK
igor.ivkic@fh-burgenland.at | i.ivkic@lancaster.ac.uk

Burkhard List
b&mi GmbH & Co KG
b@bandmi.at

HIGHER: European Heterogeneous Cloud/Edge Infrastructures for Next Generation Hybrid Services

by Manolis Marazakis and Stelios Louloudakis (ICS-FORTH)

The HIGHER project provides cloud and edge infrastructures using European data centre-ready processor technologies and system design conforming to Open Compute Project (OCP) standards. The innovative framework aims to accelerate the energy transition and drive sustainable technological growth.

The European Cloud and Edge computing markets are projected to experience substantial growth, with an estimated increase from $56.85 billion to $470.13 billion between 2022 and 2032, for the cloud market and from $15.54 billion in 2023 to $147.38 billion by 2032 for the edge computing market, reflecting a compound annual growth rate (CAGR) of more than 25% . These markets will play a pivotal role in European economic growth and social development. Notable technologies driving this growth include multi-core CPUs and accelerators, the vast majority of which are US-made, which are expected to achieve a Compound Annual Growth Rate (CAGR) of 49.47% from 2022 to 2032. By 2025, cloud and edge infrastructures are anticipated to cater to over 50 billion users. Unfortunately, the European Cloud Provider Share has significantly reduced in the last years and one of the main reasons behind it is the lack of European Cloud equipment and software. It is thus evident that Europe requires an advanced, energy-efficient, competitive European Cloud and Edge Computing Architecture driven by EU technology and infrastructures.

HIGHER [L1] aims to prototype and demonstrate the first all-European next-generation data center-ready processor and management modules and integrate them into cloud and edge infrastructures using European technologies and Open Compute Project (OCP) standards (See also the European Processor Initiative [L2]. This effort aims to create novel server platforms capable of efficiently deploying cloud and edge applications and services. Specifically, HIGHER will adhere to the OCP Data Center Stack (DC-Stack) standards, providing data center-ready integrated systems for edge, private cloud, and large data centers. Following the OCP Data Center – Modular Hardware System (DC-MHS), which offers specifications for modules compatible across servers, chassis and vendors, HIGHER will offer the following hardware modules:

- Processor Modules - Two OCP-compliant processor modules, which will be based on the OCP Host Processor Module (HPM) or the OCP Universal Baseboard (UBB) standard, one hosting the RHEA2 EPI chip and the other hosting the EPAC2.0 EPI chip and the pin- compatible EUPilot chip [L3].
- Management Module - A Data Center-ready Secure Control Module (DC-SCM), hosting a RISC-V processor inside an FPGA for server management, security, and control features.

These hardware modules will be integrated into the HIGHER server chassis, utilising commercial off-the-shelf OCP-compliant components for power distribution, network connectivity, and storage (see Figure 1). The specific mechanical aspects of the server will be developed within the project. The resulting platform will provide a modular hyperconverged infrastructure, offering distributed resources interconnected directly in hardware for efficient utilisation and management. The modular infrastructure will be easily configurable, and by integrating different HIGHER processor modules, we will be able to provide cloud systems incorporating multiple OCP sleds with high-end processors and accelerators, or smaller edge systems with weaker processors, based on requirements and the position of the server in the cloud-edge continuum. The HIGHER architecture is structured into three main conceptual layers that deliver all-European open-source and open-standard-based software and hardware modules integrated into a complete system:

- Hardware Platform layer, encompassing the development of OCP hardware modules and the final modular integrated platform.
- System Software layer, encompassing the software infrastructure for booting and controlling hardware, accessing hardware peripherals, and providing the execution environment for cloud and edge applications and services.

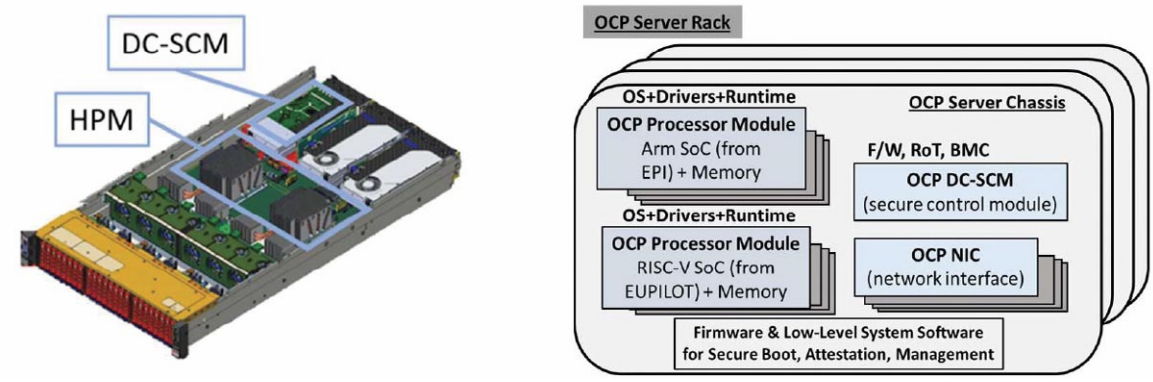


Figure 1: (left) OCP Server Chassis, with host processor modules and secure control module; (right) Hardware and Software Components in OCP Server Rack.

- Use Cases layer, encompassing four use cases for the evaluation of the capabilities of HIGHER platforms:
  1. Accelerated data processing and analysis;
  2. Infrastructure as a Service;
  3. Platform as a Service;
  4. Remote CXL-based disaggregated memory.

HIGHER receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101189612. The project is funded under the European Health and Digital Executive Agency (HaDEA) call on "Digital and emerging technologies for competitiveness and fit for the Green Deal (2023/24)"[L4].

Links:
[L1] https://www.higher-project.eu
[L2] https://www.european-processor-initiative.eu
[L3] https://eupilot.eu
[L4] https://kwz.me/hFH

Reference:
[1] EU strategic autonomy 2013-2023: From concept to capacity (European Parliament briefing, July 2022). https://kwz.me/hFE

Please contact:
Manolis Marazakis, ICS-FORTH, Greece
maraz@ics.forth.gr
Stelios Louloudakis, ICS-FORTH, Greece
slouloudak@ics.forth.gr

Multimedia Understanding through Semantics, Computation and Learning

Call for Participation

13th International Workshop on Computational Intelligence for Multimedia Understanding

in cconjunction with ISCAS'2025, London 25-28 May 25–28 2025

MUSCLE (formerly IVU, Image and Vision Understanding) is the ERCIM Working Group focused on multimedia understanding through semantics, computation, and learning. For over 15 years, this working group has brought together teams from ERCIM and non-ERCIM institutions, uniting expertise in machine learning, artificial intelligence, image/video/audio processing, and multimedia processing and management.

The International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM) is the group's flagship annual event. Following the 2024 edition, held in conjunction with ISCAS in Singapore, and the 2023 edition, co-located with ICASSP in Rhodes, Greece, the 2025 workshop will again be hosted in collaboration with ISCAS. This time, it will take place in London from May 25–28, 2025.

The primary aim of IWCIM 2025 is to present and discuss current trends and future directions in computational intelligence for multimedia understanding, while fostering the creation of a robust network of scientists and practitioners. This network aims to provide seamless access to expertise, data, and ideas.

Multimedia understanding is a cornerstone of many intelligent applications that influence our daily lives, from household devices to commercial, industrial, service, and scientific domains. Analyzing raw data to extract semantics is vital for unlocking their full potential and enabling us to manage everyday tasks efficiently. Today's raw data originate from a wide variety of sensors and sources, differing in nature, format, reliability, and information content. Multimodal and cross-modal analysis is indispensable to harnessing these data effectively. Beyond data analysis, these challenges extend to data description, supporting efficient storage and mining. The interoperability and exchangeability of heterogeneous and distributed data are crucial for practical applications. Semantics represents the highest level of information. Inferring semantics from raw data requires leveraging both the data itself and prior knowledge to extract structure and meaning. Computational methods, including machine learning, statistical modeling, and Bayesian approaches, are essential to achieve this goal at various levels.

The scope of IWCIM 2025 encompasses, but is not limited to, the following topics:

- Multisensor systems
- Multimodal analysis
- Crossmodal data analysis and clustering
- Mixed-reality applications
- Activity and object detection and recognition
- Text and speech recognition
- Multimedia labeling, semantic annotation and metadata
- Multimodal indexing and searching in very large data-bases
- Big and Linked Data
- Search and mining Big Data
- Large-scale recommendation systems
- Multimedia and Multi-structured data
- Cloud Optimization
- Pervasive intelligence
- Machine learning in multimedia understanding
- Attention based approaches for multimedia understanding
- Diffusion models for multi-modal data analysis
- Multi-modal data analysis in compressed domain
- Multi-modal data analysis for remote sensing applications
- Semantic web and Linked data
- Case studies

Link: http://iwcim.itu.edu.tr

Please contact:
Behçet Uğur Töreyin, ITU, Istanbul,Turkey
toreyin@itu.edu.tr

Maria Trocan, Institut Supérieur d'Électronique de Paris (Isep), Paris, France
maria.trocan@isep.fr

Davide Moroni, ISTI-CNR, Pisa, Italy
davide.moroni@isti.cnr.it

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Call for Proposals

Dagstuhl Seminars and Perspectives Workshops

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is accepting proposals for scientific seminars/workshops in all areas of computer science, in particular also in connection with other fields.

If accepted, the event will be hosted in the seclusion of Dagstuhl's well known, own, dedicated facilities in Wadern on the western fringe of Germany. Moreover, the Dagstuhl office will assume most of the organisational/ administrative work, and the Dagstuhl scientific staff will support the organizers in preparing, running, and documenting the event. Thanks to subsidies the costs are very low for participants.

Dagstuhl events are typically proposed by a group of three to four outstanding researchers of different affiliations. This organizer team should represent a range of research communities and reflect Dagstuhl's international orientation. More information, in particular details about event form and setup, as well as the proposal form and the proposing process, can be found on

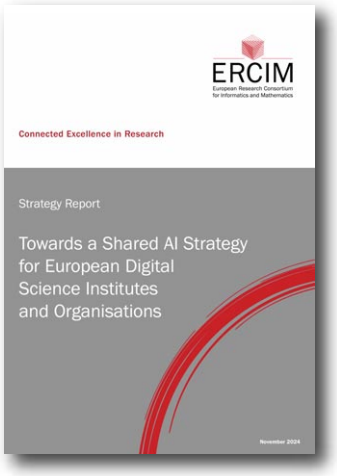
https://www.dagstuhl.de/dsproposal

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is funded by the German federal and state government. It pursues a mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

Important Dates

- *Next submission period:*
  April 1 to April 15, 2025
- *Seminar dates:*
  Between April 2026 and March 2027 (tentative).

ERCIM Published a Stragegy Report "Towards a Shared AI Strategy for European Digital Science Institutes and Organisations"



Generative AI (GenAI) and Large Language Models (LLMs) are transforming science and society, presenting significant opportunities alongside inherent challenges. European digital science institutes and organizations are well-positioned to contribute to responsible development and use of these technologies.

This report summarizes the discussions and findings from the 2024 ERCIM Visionary Event "Challenges and Opportunities of Foundational Models and Generative AI (GenAI) for Science and Society," held on 16 April 2024 in Brussels. Experts from across Europe convened to examine the rapidly evolving landscape of GenAI and LLMs, their impact on science and society, and the role of European institutions in this context.

This report first presents the findings of the event. Subsequently, it outlines several key highlights and high-level recommendations. For more details, see the section Recommendations for Institutional AI Strategies.

By promoting balanced discussions, implementing oversight mechanisms, and fostering multidisciplinary collaboration, institutions can harness the benefits of GenAI while mitigating its risks. These efforts are essential to ensure ethical alignment, maintain competitiveness, and maximize positive impacts on society and science.

We invite readers to explore the full report for a comprehensive understanding of the discussions, insights, and detailed analyses that underpin these key points and recommendations.

The full report is available for download at
https://www.ercim.eu/publication/reports/AI-Report.pdf

CWI Research Semester Programme

Truth is in the Eyes of the Machines

Amsterdam, 8-9 May 2025

How do misinformation and hate speech fuel and influence each other? How can sustainable and FAIR data (Findable, Accessible, Interoperable and Reusable) be developed to independently investigate misinformation and hate speech? How robust are generative AI models at detecting and actively countering information disorders? These research questions will guide the Research Semester Programme on misinformation detection and countering in the era of Large Language Models. The Program will be organized along three full-day workshops, with keynotes and breakout sessions.

The workshops will be held on
- 8-9 May 2025 in Amsterdam,
- 23 May 2025 in Groningen,
- 20 June 2025 in Amsterdam.

The organisers are Davide Ceolin (CWI, Human-Centered Data Analytics), Anastasia Giachanou (Utrecht University, Department of Methodology and Statistics) and Tommaso Caselli (University of Groningen, Jantina Tammes School).

More information:
https://kwz.me/hF1

AIVD, CWI, and TNO Published Renewed Handbook for Quantum-safe Cryptography

To prepare organizations for Q-Day, the day when quantum computers will be able to break certain widely used cryptography, the General Intelligence and Security Service (AIVD), Centrum Wiskunde & Informatica (CWI), and TNO have published a renewed handbook for quantum-safe cryptography. This extended second edition contains the latest developments and advice for transitioning to a quantum-safe environment, including more concrete advice on finding cryptographic assets, assessing quantum risks, and setting up cryptographic agility. It was presented on 3 December 2024 to the State Secretary for Digital Affairs and Kingdom Relations, Zsolt Szabó, during the Post-Quantum Cryptography Symposium in The Hague.

Q-Day

Cryptography is used to protect data that should not be accessible by others. However, not every form of cryptography is safe against attacks by quantum computers. This Q-Day could occur within the next five to fifteen years, according to some experts. Malicious actors, such as hostile state actors, could then largely bypass certain contemporary cryptography.

Second Edition and PQChoiceAssistant

Since the publication of the first edition, more knowledge has been gained in the field of post-quantum cryptography (PQC). PQC is a collection of encryption methods that, unlike certain current methods, should be safe against attacks with quantum computers. This revised and extended second edition includes the latest developments and advice in the field of PQC. Additionally, several essential actions for companies and organizations in the PQC migration have been examined in more detail. Furthermore, more concrete advice is included for inventorying cryptographic components in software used by organizations, assessing quantum risks, and cryptographic agility. It also provides a list of steps that are useful for any organization, regardless of the quantum threat ("no-regret moves"), and a detailed overview of PQC methods and international legislation.

European cooperation

Since 2021, the CWI Cryptology research group and TNO have been organizing a series of symposia on post-quantum cryptography with the theme "Act now, not later." The aim is to bring government, business, and science together. The event on 3 December in The Hague, the 7th episode of this series, focused on internationalization and was organized with the help of the Ministry of the Interior and Kingdom Relations. One of the main topics was the development of the European Roadmap to make the European digital infrastructure quantum-safe. This roadmap should lead to a coordinated transition, with attention to interoperability, standards, and knowledge sharing within Europe. The Netherlands plays a leading role in this, together with Germany and France. These three countries jointly coordinate the EU working group.

More information: https://kwz.me/hF4

From left to right: Maarten Tossings (COO, TNO), Zsolt Szabó (Minister of Digitalisation and Kingdom Relations), Bas Dunnebier (CSTO, AIVD), and Ton de Kok (director CWI).

Marcin Żukowski Receives CWI Dijkstra Fellowship

CWI awarded the Dijkstra Fellowship on 21 November to Marcin Żukowski, former CWI database researcher and co-founder of the globally successful tech company Snowflake. Żukowski's pioneering work produced techniques that are still essential for efficiently processing huge amounts of data, leading to faster analysis and new opportunities for companies working with big data.

With the Dijkstra Fellowship, named after Dutch computer pioneer Edsger W. Dijkstra, CWI honours Żukowski's exceptional scientific and technological contributions. For example, Żukowski's innovations have now been integrated into technologies that drive Snowflake's global success and play an important role in database management combined with cloud solutions.

Marcin Żukowski at the Dijkstra Fellowship & Lectures.

"Marcin is an excellent example of how CWI's mission can be put into practice. He used his PhD research at CWI to create versatile fundamental software products ", said CWI director Ton de Kok.

Marcin Żukowski began his career at CWI, where he introduced the award-winning concept of vectorised execution and led to the spin-off VectorWise (now Actian), an analytical database system. In 2012, he became Snowflake's third founder, bringing his vectorised execution and lightweight compression innovations to the first database system designed for external cloud storage. Having left Snowflake earlier this year, he now serves as an investor, consultant, and supporter of technology development and innovation in Poland.

More information: https://kwz.me/hF0

ERCIM
European Research Consortium for Informatics and Mathematics

ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.

ERCIM is the European Partner of the World Wide Web Consortium.

CNR
Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
www.iit.cnr.it

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

CWI
Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
www.cwi.nl

RISE
RISE SICS
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

Fonds National de la Recherche Luxembourg
Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
www.fnr.lu

SBA Research
SBA Research gGmbH
Floragasse 7, 1040 Wien, Austria
www.sba-research.org/

FORTH
INSTITUTE OF COMPUTER SCIENCE
Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
www.ics.forth.gr

SZTAKI
Eötvös Loránd Research Network
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
www.sztaki.hu/

Fraunhofer
IUK-TECHNOLOGIE
Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
www.iuk.fraunhofer.de

University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
www.cs.ucy.ac.cy/

inesc
INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal
www.inesc.pt

UNIVERSIDAD DE MÁLAGA
Institute for Software Engineering and Software Technology
"Jose María Troya Linero", University of Malaga
Calle Arquitecto Francisco Peñalosa, 18, 29010 Málaga
https://gp.uma.es/itis

Institut National de Recherche en Informatique et en Automatique
B.P. 105, F-78153 Le Chesnay, France
www.inria.fr

i.S.I.
Industrial Systems Institute
I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
www.isi.gr

UNIVERSITAS VARSOVIENSIS
Universty of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
www.mimuw.edu.pl/