ATLANTIS

Newsletter #3 **May 2024**

In our first newsletter, we outlined the ATLANTIS project's commitment to securing Europe's Critical Infrastructures (CIs) — those essential services that underpin our society's well-being and economic health.

In the second newsletter, we presented the progress demonstrated within the project, through the advancements in its three Large-Scale Pilots (LSPs). These pilots aim to set a precedent for the future of European CI protection, in order to elevate the systemic resilience of Europe's critical infrastructures and safeguard European security, well-being, and economic prosperity against the risks of the modern era.

This newsletter demonstrates the progress of the technologies that are being developed for the needs of ATLANTIS project, under each Task of WP3 and WP4.

- **WP3** focuses on the development of Protective Technologies to reduce systemic risks through innovation. A key development is the Network Monitoring Intrusion Detection System (NMIDS), which employs unsupervised machine learning to swiftly detect network anomalies, enhancing early threat detection. Alerts are standardized into a Unified Alert Format (UAF), improving ATLANTIS's situational awareness capabilities. We're advancing an Awareness & Comprehension Framework (ACF) for early risk detection and a novel Risk Reduction & Incident Mitigation Framework (RRIM) for proactive threat management. Additionally, we are addressing the challenge of disinformation with new tools designed to combat misinformation and verify data authenticity, enhancing the accuracy and reliability of information crucial for decision-making in critical infrastructure protection.

- **WP4** targets on Cooperative Prevention, Anticipation, and Mitigation of Systemic Risks. WP4 is enhancing our capacity to handle systemic threats through the development of the CCI-SAAM framework and human-explainable AI (XAI). These tools improve the analysis and understanding of threat intelligence, making complex data actionable and comprehensible for operators. Additionally, WP4 integrates these advancements into the ATLANTIS platform with a robust Continuous Integration/Deployment/Pilot (CI/CD/CP) framework, ensuring rapid deployment and high standards of reliability and security.

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

After this brief introduction of the current situation of the project, the third ATLANTIS newsletter will offer updates on the progress made by each partner within WP3 and WP4, highlighting the technologies under development and their impact to the project.

WP3– Protective Technologies to reduce systemic risks by innovation

Interfacing existing CI security systems & patterns extraction

Technology: Network Monitoring Intrusion Detection System (NMIDS)

Description, Key Features and Benefits:

NMIDS is a network monitoring tool that uses machine learning to detect anomalies and potential cyber threats within critical infrastructure networks in real-time. Unsupervised machine learning establishes a baseline of normal network behavior and detects deviations. The adaptation layer converts detected alerts into the Unified Alert Format (UAF) based on IDMEFv2.

Early detection of cyber incidents enables timely response, reducing impact on network operations and dependency systems. Data will improve ATLANTIS' situational awareness capabilities.

Current Status and Next Steps:

Detection algorithms are developed and tested on local data. Adaptation layer in final testing stage.

- Immediate Next Steps: Evaluate detection performance on open-source datasets to prepare for integration with ATLANTIS platform.
- Anticipated Milestones: Complete development and integrate NMIDS as part of the ATLANTIS situational awareness and threat detection framework (D3.2, available in November 2024).

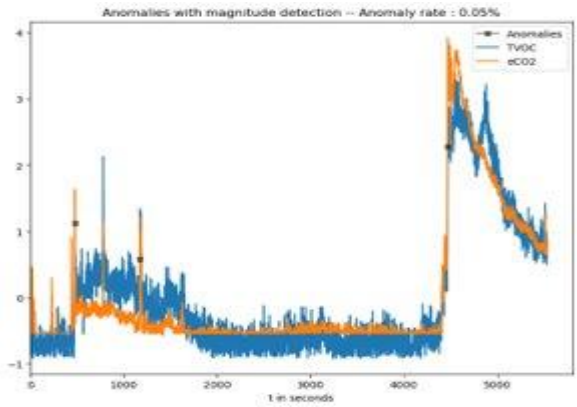
Technology: Ai Quality Sensor – SNIFFERLIKE

Description, Key Features and Benefits:

SNIFFERLIKE is a sensor that can monitor level of gazes at given positions of interest and transmit the data in real time on the chosen network. It is able to measure the level of different gazes that could be dangerous for human or the environment. It has low consumption, is made with low price sensors. It is also very compact and easily transportable and can be connected via Ethernet to send its data every second.

This sensor is a simple solution to have localized information on the presence or absence and the level of dangerous gazes following an accident or an attack. Since it is easily transportable and connected, it can be deployed on key location in a short time and generate data almost immediately.

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909



Time series anomaly detection on SNIFFER data

Current Status and Next Steps:
At the moment, the sensor exists in a simple prototype model, which was designed to prove the simplicity of measurement methods and deployability of the system. It measures CO2 and small particles levels. The measure is functional, as well as the communication and can be deployed for testing but the levels are not metrological.

- Immediate Next Steps: Extend the possibility of the sensor to detect specific gazes of interest, like hazardous material. Once the possibilities have been listed, it is important to choose the most pertinent ones and integrate them in the sensor.
- Anticipated Milestones: The selection of the pertinent gazes of interest, the nit is the integration in the sensors. After that, there will be a phase of testing and experimental validation.

Tools to fight disinformation

Technology: Content-Based Similarity search (CBS)

Description, Key Features and Benefits:
The Content Based Similarity Search (CBS) module is an advanced tool designed for retrieving images by analyzing their content. Utilizing powerful machine learning algorithms, it extracts feature vectors from images, enabling it to search for and identify similar content within a database efficiently. The module integrates a MongoDB for storing feature vectors, a Feature Extraction module for analyzing images, a Feature Search module for retrieving similar images, and a MinIO storage making it a comprehensive solution for content-based image retrieval.
Main Features:

- **Efficient Feature Extraction:** Employs state-of-the-art machine learning models to extract detailed feature vectors from images, enabling precise content analysis.
- **Flexible Input Methods:** Supports image input through direct uploads or specifying image directories, accommodating diverse user needs.
- **Modular Design:** Comprises distinct but integrated modules for feature extraction and search, alongside MongoDB for data management and MinIO for storage, ensuring scalability and ease of maintenance.

Implementing the CBS module can significantly enhance the capabilities of systems requiring efficient and accurate image retrieval. By facilitating quick access to similar images, it improves user experience, aids in content discovery, and supports effective digital asset management.

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

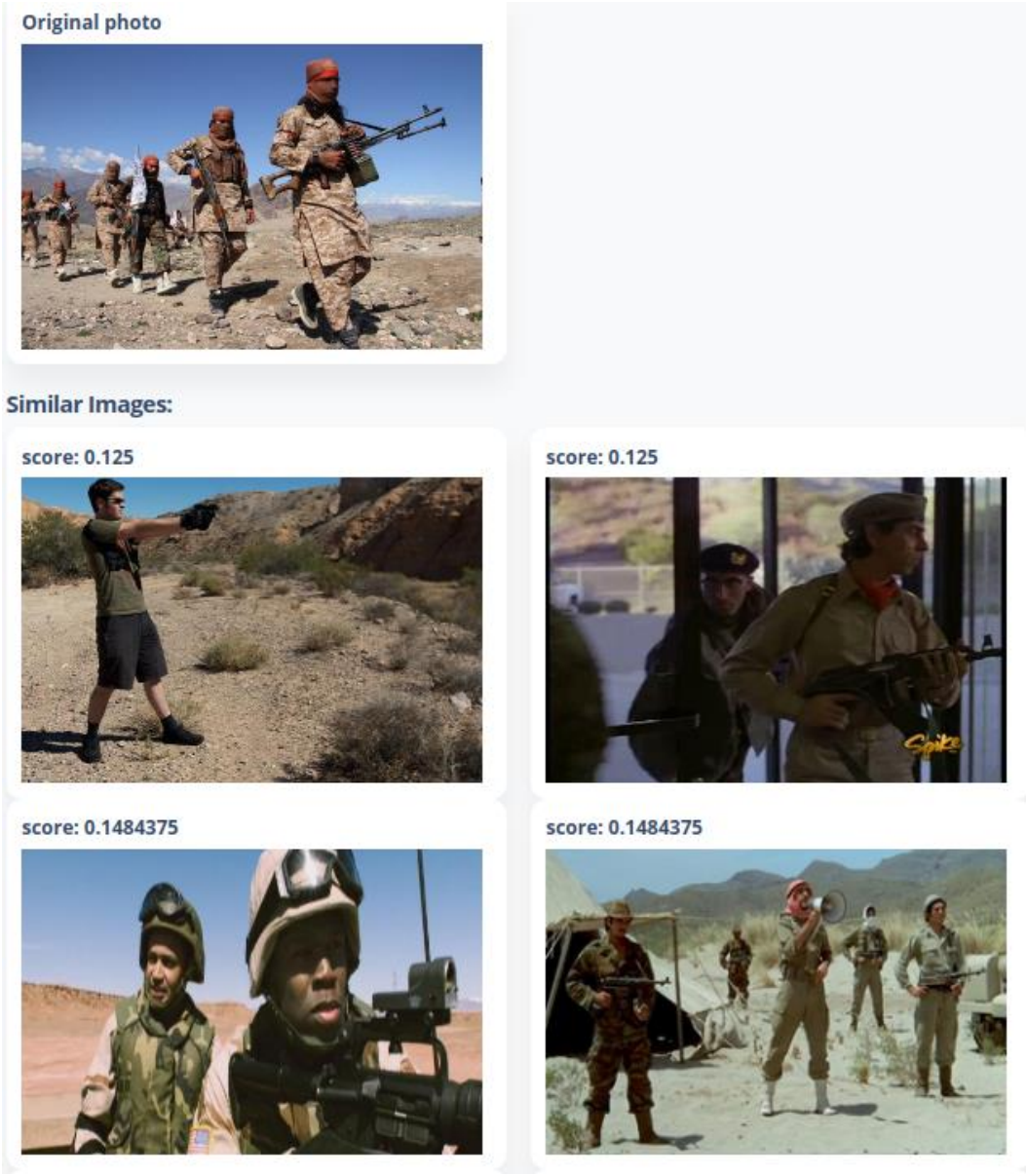


Image retrieval results with CBS module

Current Status and Next Steps:

The CBS module is currently in a late development stage, with its core components for feature extraction and search fully functional and integrated with MongoDB for data management.

- Immediate Next Steps:
  - **Additional Feature Extraction Models:** Investigate and incorporate more advanced feature extraction models to broaden the scope of similarity searches. This aims to enhance the module's ability to accurately identify similarities across a wider array of image types and contents, further improving the versatility and utility of the CBS module.
  - **Deployment to Project's Cloud Space with Kubernetes:** Deploy the containerized CBS module to the project's cloud infrastructure using Kubernetes to leverage its auto-scaling and management capabilities for handling varying loads.
- Anticipated Milestones: D3.2

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

Technology: Deepfakes - Detection

Description, Key Features and Benefits:

DeepFakes Detection Component is an AI tool that accurately identifies deepfake media, ensuring digital content integrity with fast, advanced machine learning algorithms.

- Main Features**:**
  - **API Integration:** Features a comprehensive FastAPI implementation, offering straightforward endpoints for submitting detection tasks and retrieving results.
  - **Scalable Processing:** Incorporates a queue management system (RQ) with Redis for efficient task management and scalability.
  - **Video and Image Support:** Capable of analyzing both video files and images to detect deepfakes.
  - **Flexible Input Methods**: Supports a range of input methods for media submission, including direct file uploads, URLs to media files, and YouTube URLs, catering to diverse user needs and scenarios.
  - **Secure Authentication:** Enforces API key validation to ensure secure access to the service.

By implementing the DeepFakes Detection Component, our project can significantly enhance the trustworthiness and authenticity of digital media content. It serves as a crucial tool in combating misinformation and preserving content integrity, which is vital in today's digital age. Moreover, the API's scalability and easy integration facilitate its adoption in various digital platforms, further broadening its positive impact on media credibility.

Deepfakes detection with gradcam analysis

Current Status and Next Steps:

The technology is currently in the development stage, with core functionalities for video and image deepfake detection fully implemented and operational through a FastAPI interface. Continuous research into the latest deepfake detection methodologies is a priority, ensuring that the detection algorithms are always at the cutting edge.

- Immediate Next Steps:
  - **Deployment to Project's Cloud Space with Kubernetes**: Plan and execute the deployment of the DeepFakes Detection API on the Atlantis cloud infrastructure using Kubernetes. This involves setting up the necessary Kubernetes configurations for scalability and high availability, and ensuring seamless integration with the cloud resources.
  - **Update User Documentation**: Update the existing user documentation and tutorials to include the latest features, deployment guidelines, and use cases to facilitate easier adoption of the API.

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

- Anticipated Milestones: D3.2

Technology: Image - Manipulation

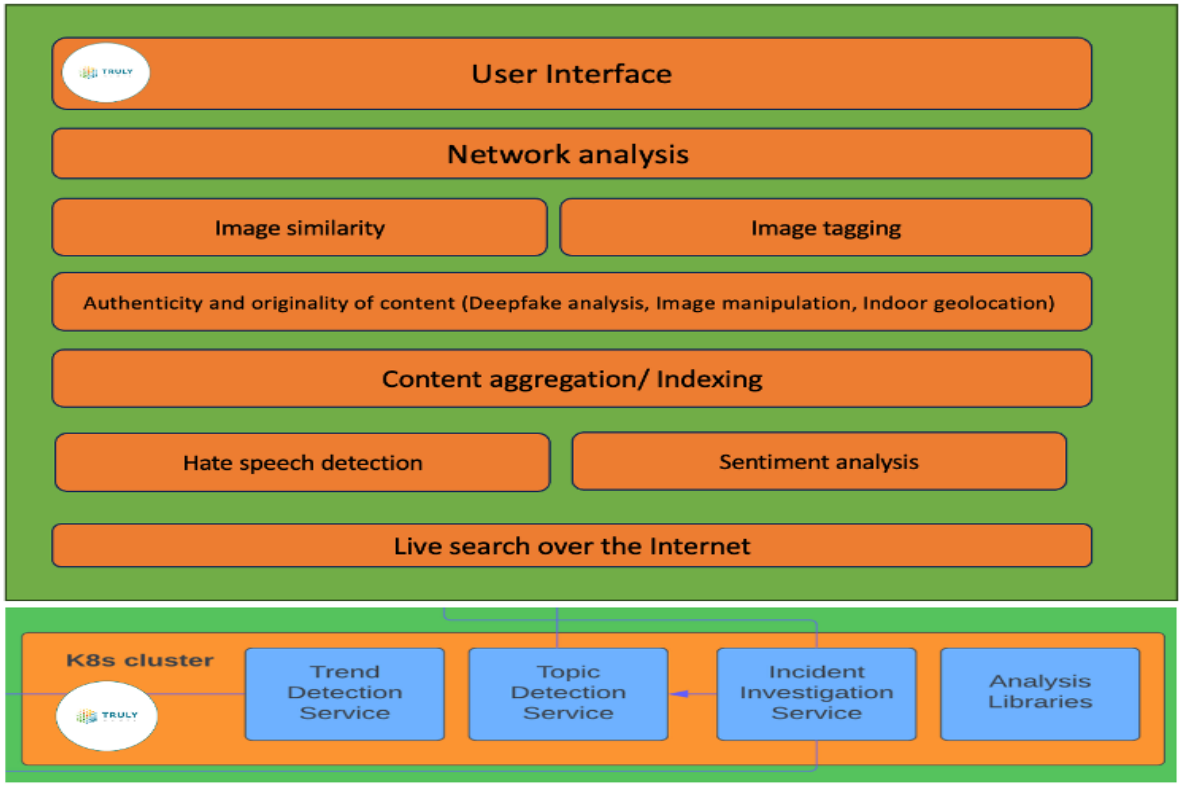
Description, Key Features and Benefits:

The Image Manipulation Component integrates four sophisticated methodologies: MantraNet, Exif, Busternet, and Quadratic, each targeting specific aspects of image manipulation. Built on FastAPI, it provides a secure and efficient way to detect and analyze manipulations through various endpoints, making it a versatile tool for digital forensics, media verification, and academic research.

Main Feature**s:**

- **Comprehensive Manipulation Detection:** Covers a wide range of image manipulation types, offering specialized analysis through ManTraNet for manipulation prediction, Exif for metadata analysis, Busternet for copy-move prediction, and Quadratic for complex manipulation detection.
- **Comprehensive Analysis:** Provides detailed responses including manipulated images, bounding boxes, detection areas, and manipulation scores.
- **Flexible Input Methods:** Supports a range of input methods for media submission, including direct file uploads and URLs
- **API Integration:** Features a comprehensive FastAPI implementation, offering straightforward endpoints for submitting detection tasks and retrieving results.
- **Secure Authentication:** Enforces API key validation to ensure secure access to the service.

This component significantly enhances the capabilities of platforms needing robust image manipulation detection. Its modular design and comprehensive analysis tools provide users with detailed insights into image manipulations, aiding in the fight against misinformation and ensuring digital content integrity.



Tools to fight Disinformation

Current Status and Next Steps:

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

- **Deployment to Project's Cloud Space with Kubernetes**: Plan and execute the deployment of the Image Manipulation API on the Atlantis cloud infrastructure using Kubernetes. This involves setting up the necessary Kubernetes configurations for scalability and high availability, and ensuring seamless integration with the cloud resources.
- Anticipated Milestones: D3.2

Situation Awareness & Comprehension Framework

Technology Name: Knowledge Graph

Description, Key Features and Benefits:
The Knowledge Graph serves as a centralized repository for storing critical infrastructure data and the Interdependency Graph. This repository allows various systems such as SAFER, DSS, and RRIM to access and consult crucial information related to interdependencies between different infrastructure elements. It allows efficient storage and retrieval and maintains interconnected maps of dependencies, enabling a holistic understanding of system vulnerabilities.
The Knowledge Graph optimizes decision support by providing comprehensive critical infrastructure data, enhancing resilience planning, and streamlining data access. It enhances interoperability between systems and is part of T3.3 and supports ATLANTIS SO.3, SO.6 and SO.7.

Current Status and Next Steps:
- It is developed and deployed.
- Immediate Next Steps: Populate with updated interdependency data.
- Anticipated Milestones: D3.2

Technology: SAFER

Description, Key Features and Benefits:
SAFER aims to enhance the awareness and understanding of critical infrastructure threats, both short-term and systemic, through advanced computational methods. It enhances CI situational awareness of cyber-physical threats through advanced ML analysis of cyber-physical data. It employs privacy-preserving methods like Differential Privacy during training, validated by zero knowledge proof. Trained models are aggregated for accuracy, utilizing integrated data from diverse sources for comprehensive threat assessment.
SAFER employs privacy-preserving techniques to safeguard sensitive CI data, analyses diverse data sources for better threat understanding, facilitates efficient information sharing among CI entities, incorporates advanced ML for precise threat assessment, and ensures confidentiality compliance, enhancing trust in system operations. It is part of T3.3 and supports ATLANTIS SO.3, SO.6 and SO.7.

Current Status and Next Steps:
- Development Stage: Under development.
- Immediate Next Steps: First integration steps with WP3 components.
- Anticipated Milestones: D3.2.

Technology: GreenTwin

The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

Description, Key Features and Benefits:

GreenTwin is a digital twin platform that aims to enhance the awareness and understanding of critical infrastructure threats through an advanced, dynamic, and live 3D interface. In the backend, ML and data from the site is processed and shown in 2D and 3D. Tool for communication between different parties enables quick, quality, and logged decisions in case of incidents or prevention of incidents.

GreenTwin enhances CI situational awareness of cyber-physical threats through 3D interface, datadriven and advanced ML analysis of cyber-physical data. It employs privacy-preserving methods like differential views depending on end user. Various data sets and services can be connected to digital twin (databased, software data, IoT data…).

GreenTwin offers quick and user-friendly situation visualisation that is alive, meaning that viewer can see situation as it is, with all real-time data and historic data. Access control module ensures privacy preserving techniques to safeguard sensitive CI data and assets, for safe log-in purposes blockchain technology is implemented. Visual data sources offer better threat understanding, facilitates efficient information sharing among CI entities, incorporates advanced ML for precise threat assessment, and ensures confidentiality compliance, enhancing trust in system operations.

Current Status and Next Steps:

- Development Stage:
  - It is already developed, some parts that are unique for ATLANTIS, are being optimised
  - Double virtualisation of system is in progress, showing physical assets such as servers, HVAC, power consumption, temperatures with real-time data. Services, connections, data flow, DDOS and other alerts pop up in case of incidents, users and microservices shown and managed
- Immediate Next Steps: connecting real data to both digital twins: LSP1 and Double Virtualisation

Systemic Risks Foresight and Incidents Detection DSS

Technology: Dynamic Risk Analysis using Bayesian Belief Networks

Description, Key Features and Benefits:

Conventional risk models consider prior knowledge about the risks from the hazards and threats; thus, they are more or less static. The proposal is to develop a dynamic risk model that can consider the information on the occurrence of the hazards and threats related events and thus update the model. It is advocated that the Bayesian Belief Networks (BBNs) based software tools can be used to develop case CI risk models that can be updated in real-time based on the information from the specific sensing technologies, such as those proposed by the ATLANTIS project.

Dynamic BBN risk models can easily consider complex sequences of disaster events and allow them to be updated with new information (evidence) about the hazards/threats, quality of the safety/security measures, etc.

The dynamic risk models are always case-specific and allow simulation of the expected outcomes of the disaster events. The main benefit is in the ability to search and optimize for the risk reduction measures, as well as the hazards/threats "live" diagnosis to support the risk managers daily. This aligns with the T3.4 goals and supports ATLANTIS SO.1 and SO.2.

Current Status and Next Steps:

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

- Development Stage: Developed. Tested in the past InfraStress project and currently applied in ATLANTIS WP5 LSP#1 within the "Fire" scenario.
- Immediate Next Steps: The "Fire" BBN risk model is to be linked with the proposed risk KPIs (subject to data provided by the partners).

Anticipated Milestones: D5.3 available in May 2025.

Technology: DSS

Description, Key Features and Benefits:

The ATLANTIS Incidents Detection Decision Support System (DSS) is designed to analyse current Situational Picture Model provided by SAFER and relevant data coming from other ATLANTIS tools to detect potential risks within CIs. It assesses systemic threats, evaluates the likelihood and scope of incidents, and determines the necessity of preventive actions.

The main features of DSS are Situation Analysis and Risk Detection, Multi-Criteria Detection Support and Semantically Enhanced Threat Extraction.

The ATLANTIS Incidents Detection Decision Support System (DSS) facilitates proactive risk management by analysing current situations and identifying potential incidents, enabling timely preventive actions. It enhances collective situational awareness and collaboration among critical infrastructure entities. By considering multiple criteria and utilizing advanced frameworks, it supports informed decision-making and prioritizes risk reduction measures. Overall, it improves the resilience of critical infrastructure against various threats and disruptions. It is part of T3.4 and supports ATLANTIS SO.3, SO.6 and SO.7.

Current Status and Next Steps:

- Development Stage: Under development.
- Immediate Next Steps: First integration steps with WP3 components.
- Anticipated Milestones: D3.2

Risk Reduction & Incident Mitigation/Recovery DSS

Technology: Risk Reduction and Incident Mitigation (RRIM)

Description, Key Features and Benefits:

The Risk Reduction and Incident Mitigation (RRIM) is a framework designed to identify, evaluate and recommend effective countermeasures for a wide range of security incidents occurring in critical infrastructure environments.

Using advanced algorithms and data analytics, RRIM matches countermeasures to incidents using a semantically enhanced countermeasure repository for knowledge that is enriched by external information sharing systems.

The RRIM enables informed decision-making to improve security of critical infrastructure by using historical data, external intelligence, adapting continuously to emerging threats.

Current Status and Next Steps:

- Development Stage: Under development
- Immediate Next Steps: First integration steps with WP3 components
- Anticipated Milestones: D3.5 available in January 2025.

The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

Humans in Vicinity Sensing and Engagement

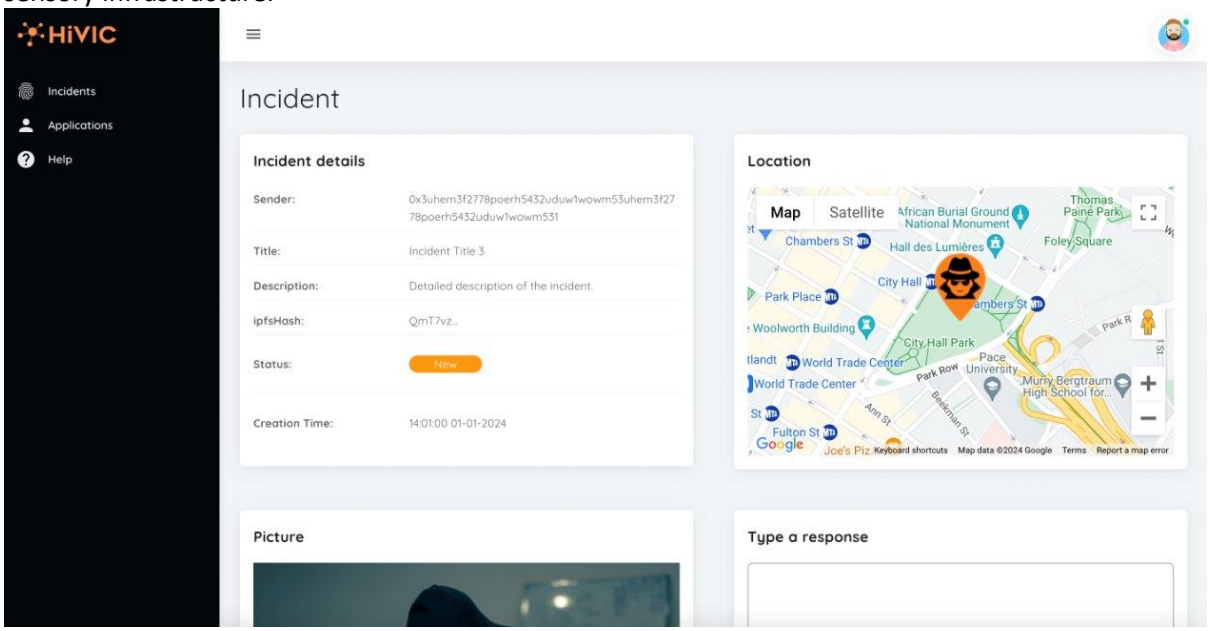
Technology: Humans in Vicinity Sensing and Engagement (HiViC)

Description, Key Features and Benefits:

HiViC enables crowdsensing and collaborative human engagement in the context of Critical Infrastructures (CI) protection, by allowing humans to act as dynamic virtual sensors, in the vicinity of CIs.

It offers a secure and private communication channel between humans and the CI operators, enabling humans to offer security/safety reports on potential security incidents in the form of text, images or videos, at the same time allowing CI Operators to send security and safety guidelines to the humans in vicinity of the CI. This communication takes place with the help of a specialized application designed for Android and iOS smartphones.

HiViC allows the dynamic deployment of virtual (human) sensors, granting ATLANTIS CIs a multimodal sensory infrastructure.



HiViC framework development

Current Status and Next Steps:

- Development Stage: Developed, some features are getting re-engineered to better match the scope of ATLANTIS technology
- Immediate Next Steps: Real-life testing of the re-engineered version.
- Anticipated Milestones: Demonstration in the upcoming LSP2 workshop, expected 2024Q2.

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

WP4– Cooperative prevention, anticipation and mitigation of systemic risks

Threat Intelligence solutions for the anticipation of systemic risks

Technology: BlackBoxXAI-FL

Description, Key Features and Benefits:

BlackBoxXAI-FL provides explanations for individual predictions which helps users understand the model's decision-making process and gain insights into which features are most influential for a given prediction.

Main Features:

- **Model Agnostic:** BlackBoxXAI-FL is designed to work with any machine learning model, making it versatile and widely applicable across various types of models
- **Private data:** BlackBoxXAI-FL can be fed with private data in a csv format

BlackBoxXAI-FL provides intuitive explanations that help users interpret the model's decision-making process more easily. By understanding how each feature contributes to the model's output, users can build trust in the model and gain insights into its behavior.

Current Status and Next Steps:

- Development Stage: The technology is developed, with core functionalities fully implemented and operates through a FastAPI interface.
- Immediate Next Steps:
  - Include use cases from other partners
  - Deploy to Atlantis cloud infrastructure using Kubernetes
  - Update User Documentation to include the latest features, deployment guidelines, and use cases to facilitate easier adoption of the API.
- Anticipated Milestones: To be tested with available data from other partners.

Technology: THINT – Threat Intelligence Tool

Description, Key Features and Benefits:

The THINT framework is an advanced technological solution developed to enhance public safety by monitoring social media platforms. It aims to identify and analyze potential security threats through innovative use of technology, serving as a crucial tool in safeguarding against digital risks.

Main Features:

- Advanced Natural Language Processing (NLP): Employs sophisticated algorithms to interpret and analyze the language used across social media, detecting subtle cues and hidden messages.
- Real-Time Threat Assessment: Capable of continuously scanning social media traffic to instantly identify potential threats, allowing for prompt response.
- Large Scale Monitoring: continuous monitor and analysis of data, ensuring comprehensive coverage and enhanced security.

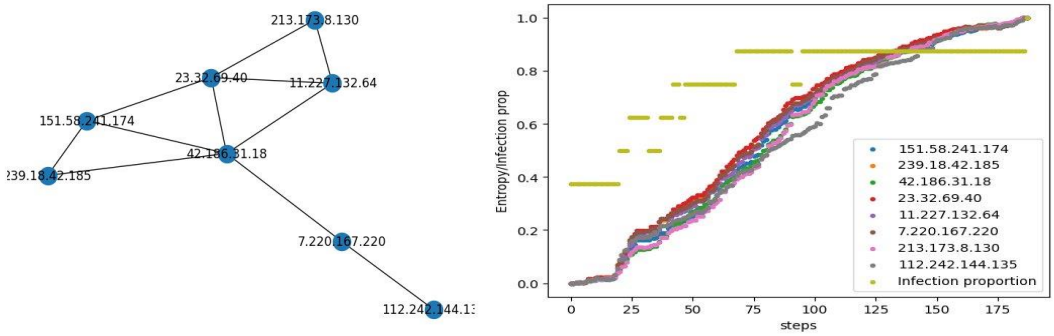
ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

Implementing the THINT framework significantly bolsters public safety by providing an effective means to preemptively detect and mitigate security threats found on social media. Its deployment could lead to a safer digital environment, reducing the risk of online terrorist activities, which currently exploit social media platforms extensively. For law enforcement agencies, security practitioners, and policymakers, THINT offers a cutting-edge tool that enhances strategic planning and response capabilities in a continuously evolving digital landscape. This proactive approach not only protects digital spaces but also has implications for safeguarding real-world environments, encouraging a collaborative effort to refine and expand the technology for future security challenges.

Current Status and Next Steps:
The THINT framework has completed the development phase with its architecture fully defined and the tool itself developed. It is now ready for deployment.
The immediate action involves extensive testing during the integration phase. This crucial step will ensure that THINT is properly connected with other tools and is receiving data seamlessly, which is vital for its functionality.
Key milestones expected soon, include the successful integration of THINT within the ATLANTIS framework, followed by use case scenarios operational testing to evaluate its effectiveness in live environments. Further milestones will focus on continuous improvements based on feedback and the expansion of its capabilities to encompass a broader range of threats and digital platforms.

Technology: EBRA

Description, Key Features and Benefits:
EBRA is a specialized tool for entropy-based risk assessment for device networks. Its main feature is the combination of ML algorithms and Graph entropy, that gives a risk score for every part in a network.
ATLANTIS will benefit from using EBRA, as it provides deep insights on the network state, endowing the operator with capability of situational anticipation.



Entropy based risk assessment – EBRA

Current Status and Next Steps:

- Development Stage: Development.
- Immediate Next Steps: Refinement of ML-based entropy computation.
- Anticipated Milestones: Testing in simulated spread environments.

XAI Tools for continuous system risk analysis and forecasting/foresight of emerging risks

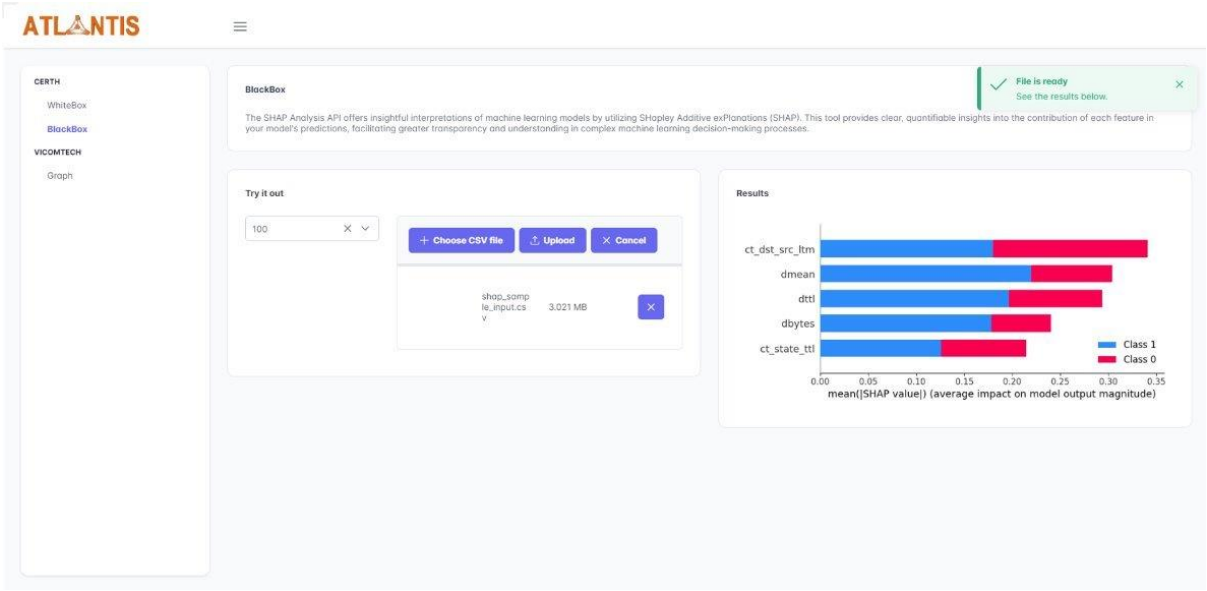
ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

Technology: xAI Dashboard

Description, Key Features and Benefits:

It aggregates and displays the xAI tools results into a consolidated view dashboard, where the CI owners will read the processed outcomes and decide and act accordingly.

- Main Features:
  - HTML5 Iframes -Visual and numerical results embed as 3rd party components or/and widgets into a web page.
  - Class Activation Maps (CAMs) heatmap plot per output class, i.e. highlight pixels in the input image that are relevant to the prediction.
  - Features plot: Most influential variables that best explain the model in local region observations and whether the variable causes an increase or decrease in the probability (supports/contradicts).
  - Explanations plot: A heatmap plot that helps to detect common features that influence all observations.
- Benefits/Impact: One single point of xAI visualisation for the following tools: Parsec Graph (VICOM), Evasion attack explainability in artificial neural networks (VICOM), BlackBox-XAI-NoFL (CERTH), WhiteBox-XAI-FL (CERTH), XAI4RecSys (SIEM), Digital Twin tool (ENG).



xAI Dashboard

Current Status and Next Steps:

- Development Stage: Under development- integration of all XAI outputs with the xAI Dashboard (CERTH, VICOM, SIEM, ENG).
- Immediate Next Steps:
  - Discuss with other partners the implementation of XAI tools in the LSP scenarios.
  - Engage LSP owners for data provision and define the synthetic data to be used temporarily.
- Anticipated Milestones: Completion of xAI tools visualization portlets and integration with T3.2 and/or other ATLANTIS components.

Technology: Parsec-graph

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

Description, Key Features and Benefits:

Parsec-graph provides a representation of networks on web. This tool is aimed to be a visualizer that can parse different models and convey relevant information through the representation of neurons with varied visual attributes (ex. Size, Colour). Parse-graph is currently developed to be the visual support of the interpretability and explainability module **evaxplainify**, and as such, it is the bridge between the end user and the API providing the results of processed image by the mentioned module. This is achieved by simple mouse interactions, that allow the navigation of the model structure with zooming and panning, neuron filtering by layer, or the exploration of a neuron detail by selection. The tool also shows the images that can be analysed by the explainability module, and that the end user can select.

- Main Features:
    - An image selection page, where the user can select among the available images for the analysis.
    - A synchronized navigation window and minimap for structure exploration. The navigation window is interactive and allows the zoom and panning action, and node selection. These user interactions are all mouse based.
    - A structured visualization of a neural network model by blocks and grouped by layers, and a comprehensive visual language that codifies the most relevant attributes of each neuron.
    - A detail view minimizable at any time that shows the explainability results of the selected neuron.
    - A layer filtering that allows the partial or total transparency of unwanted layers.
- Benefits/Impact:

A user-friendly visualization of a network model structure allows to convey complex analysis results on a more comprehensive manner. This visualization tool is designed to deliver the results obtained by the explainability module in a concise and intuitive fashion that is both informative and visually pleasant.

Current Status and Next Steps:

- Development Stage: Currently under development.
- Immediate Next Steps:
    - Integration with Xai dashboard (ATC) for deep fake images.

Technology: EvaExplainify

Description, Key Features and Benefits:

This tool combines interpretability and explainability techniques to understand artificial neural networks (ANN), converting them into a grey box. The tool generates a graph representation of the targeted ANN and tries to identify the critical or most important kernels in the model by coloring them. Moreover, in the pivotal kernels, the tool computes where they are focusing on the image, showing the regions in the images where the kernel is paying attention.

- Main Features: ANN graph representation, interpretability, explainability.
- Benefits/Impact: The ANNs are being implemented in more and more fields, which can be critical for humans. The correct functionality of those models is essential, and even the reason for their decision is crucial. This tool tries to understand the ANNs via interpretability and explainability techniques, allowing the owner to obtain more information about the behavior of the model. That brings benefits in terms of clarity and security of the model. Understanding how the ANN makes decisions is valuable information in several fields. Therefore, the

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

generated impacts are related to the trustworthiness of the model's decision-making and its security, improving the performance in terms of the viability of the model.

Current Status and Next Steps:

- Development Stage: At this moment, we are testing the tool, but it is not developed over the final artificial neural network environment.
- Immediate Next Steps: The next step in the final environment development and the tool deployment there.

Anticipated Milestones: Test finalization.

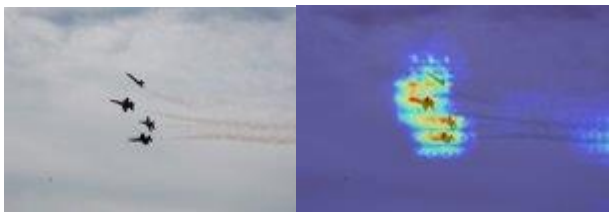
Technology: WhiteBoxXAI-FL

Description, Key Features and Benefits:

WhiteBoxXAI-FL is an explainable AI method used for visualizing the regions of an input image that had the most impact on the decision-making process with an overlay heatmap on the input image.

- Main Features:
  - **Semantic Segmentation**: WhiteBoxXAI-FL accurately localizes the important regions of the input image by generating a heatmap that highlights the relevant areas base on the model's decision
  - **Custom Target Selection**: Specific target class can be selected for better results
- Benefits/Impact:

WhiteBoxXAI-FL can be particularly useful in image-based tasks, providing interpretability and allowing a user to understand which parts of an image were crucial for the model's decision through visual representations.



WhiteBoxXAI-FL example

Current Status and Next Steps:

- Development Stage: The technology is developed, with core functionalities fully implemented and operates through a FastAPI interface.
- Immediate Next Steps:
  - Include use cases from other partners
  - Deploy to Atlantis cloud infrastructure using Kubernetes
  - Update User Documentation to include the latest features, deployment guidelines, and use cases to facilitate easier adoption of the API.
- Anticipated Milestones: This technology is planned to be integrated with technologies from WP3 and task 3.2

Technology: XAI-SP

Description, Key Features and Benefits:

XAI-SP focuses on providing Explainable Artificial Intelligence (XAI) tools to summarize incidents, offer recommendations, and identify emerging risks within critical infrastructure (CI). It aims to simplify

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

complex information to make it more accessible for cross-CI operators, thereby lowering barriers to acceptance and adoption.
Main features are Explainable AI Tools, Summarization Capabilities and Barrier Reduction.
XAI-SP improves decision-making by offering transparent summaries and recommendations, facilitating collaboration across CI sectors for a cohesive response to risks, ultimately enhancing overall resilience and security posture.

Current Status and Next Steps:

- Development Stage: Under development.
- Immediate Next Steps: Input/output definition and finalize prompt engineering.
- Anticipated Milestones: Retrofitting XAI-SP for the CIP task.

Technology: XAI-Enhanced Explainability for Risk Reduction and Incident Mitigation

Description, Key Features and Benefits:
This technology integrates Explainable AI (XAI) into the Risk Reduction and Incident Mitigation (RRIM) framework to provide clear, understandable insights into AI-driven recommendations for mitigating risks in Cyber-Physical Systems of Critical Infrastructure.
Main features:

- Integrated XAI with Large language Models (LLMs) for enhanced explanations.
- Uses advanced machine learning techniques for like LLMs.
- Incorporates LIME and SHAP for in-depth explanations.

Benefits/Impact:

- Enhances transparency, trust and reputation in automated systems.
- Supports humans-in-the-loop decision making.
- Improved effectiveness and efficiency in decision making.
- Facilitates continuous improvement and learning.

Current Status and Next Steps:

- Development Stage: Under development.
- Immediate Next Steps: First integration steps with WP3 and WP4 components
- Anticipated Milestones: D3.5.

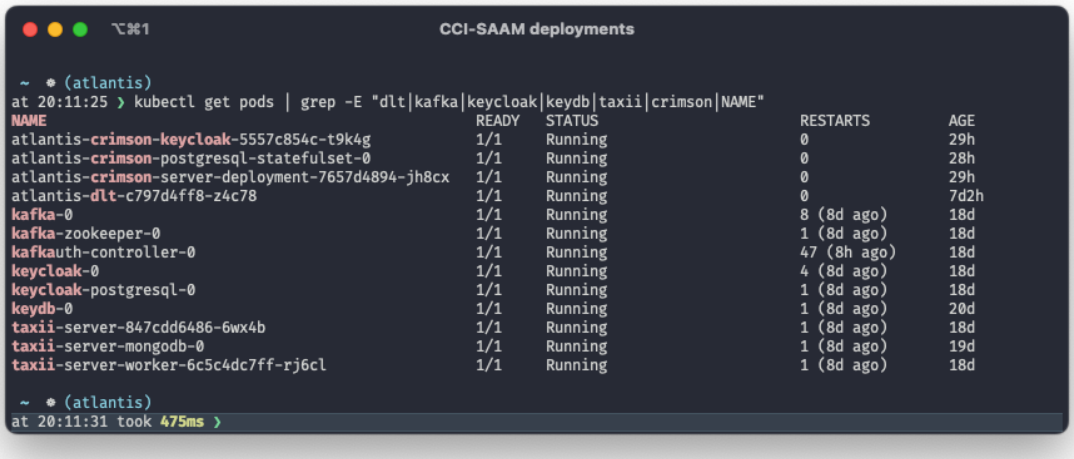
Strategies & Tools for cooperative remediation, mitigation, and response

Technology: CCI-SAAM

Description, Key Features and Benefits:
Cross-CI (CCI) Sharing Assessment Analysis Mitigation (CCI-SAAM) allows for the controlled data sharing among various, possibly cross-sector European critical infrastructures (CIs), also catering for the exposure of cascading systemic risks at pan-European level. It gathers data from the various ATLANTIS-enabled CIs and uncovers systemic risks at pan-European level. It features integration with standard protocols and frameworks such as MISP and MeliCERTes, offers a digital twin of the European CI at pan-European level and allows controlled data sharing among the various cross-connected CIs.

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

Its main benefit is that it offers a key regulatory platform towards managing risk at pan-European level and also unlocks pro-active risk-management at CI level due to timely security insights provisioning.



CCI-SAAM components already integrated and deployed

Current Status and Next Steps:

- Development Stage: Under development.
- Immediate Next Steps: Integration with the WP3 components
- Anticipated Milestones: First draft of the relevant group of components available at 2024Q2

Technology: Hypervision Tool

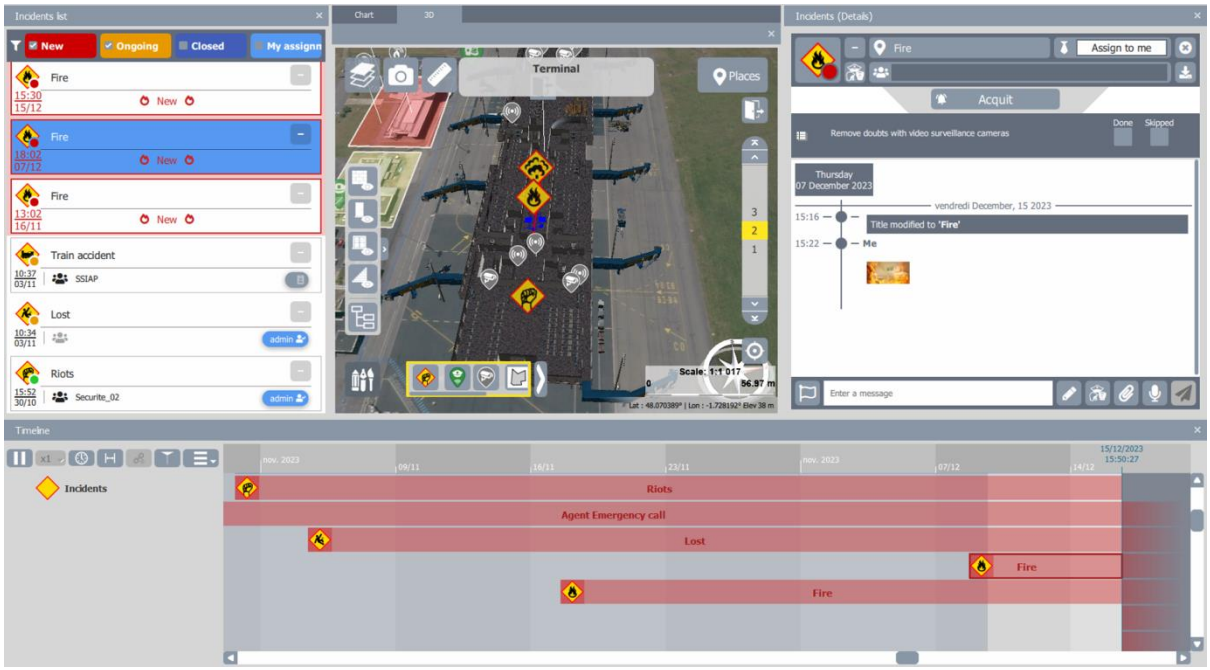
Description, Key Features and Benefits:

The Hypervision tool will serve as a common operational picture of the situation enabling critical infrastructure operators, stakeholders and decision-makers to have a complete situatial overview, collaboratively assess risks, share information, take large-scale decisions and implement mitigation strategies.

- Main Features:
  - Common Operational Picture
  - Complete and up-to date visualisation of the situation
  - Cartographic view aggregating different type of information from different sources
  - C4: Command, Control, Communication and Coordinate tool
  - Overview of ATLANTIS decision support tools outcomes to help decision makers
  - Overview of ATLANTIS mitigation strategies
  - Distrubuted solution
  - Resilient and self-healing server
- Benefits/Impact:
  - Controlled information sharing depending on the data sensitivity and user decision level,
  - Complete and up-to-date view of the current situation to facilitate large-scale decisions

The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

- Visualization of possible cascading effects and impacts of current threat on other CIs at a pan-European level
- Collaborative tool which reduce incident resolution time and impact



Hypervision tool overview

Current Status and Next Steps:

- Development Stage: Under development

- Immediate Next Steps: Management of STIX/TAXII standard format and ATLANTIS situational picture model
- Anticipated Milestones: Integration with the WP3 and WP4 components like Digital Twin, Decision Support System, XAI or Social Media analysis tools

DevSecOps CI/CD/CP framework

Technology: DevSecOps

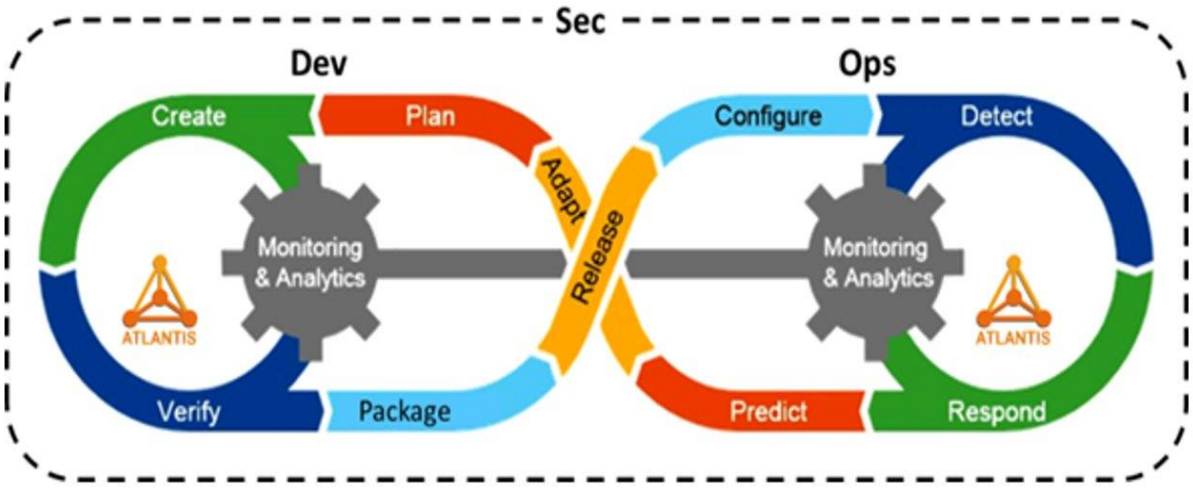
Description, Key Features and Benefits:

DevSecOps, short for development, security, and operations, automates the integration of security at every phase of the software development lifecycle, from initial design through integration, testing, deployment, and software delivery.

- Main Features:
  - security becomes an integral part of the DevOps practices in ATLANTIS.
  - monitoring and analytics tools which allow for proper logging in the whole software lifecycle.
  - Security is introduced within every process of this DevOps cycle.
- Benefits/Impact:
  - security issues are addressed as they emerge rather than relying on Quality Assurance (QA) testing at the end of the development cycle, on production.

The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

- software development, as well as release cycles are accelerated by automating the delivery of secure software without slowing the software development cycle.



DevSecOps CI/CD/CP approach

Current Status and Next Steps:

- Development Stage: The technology, based on GitLab's CI/CD pipelines and Kubernetes has been developed.
- Immediate Next Steps: Adding security stages in all components' build/deployment pipelines.
- Anticipated Milestones: To apply DevSecOps stages for all ATLANTIS Integrated Framework's components during their build and deployment phases.

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909

Project Coordinator
Mr. Gabriele Giunta
Engineering, Italy
gabriele.giunta@eng.it

Technical Manager
Dr. Artemis Voulkidis
Synelixis, Greece
voulkidis@synelix.com

Web

LinkedIn

https://www.atlantis-horizon.eu/

ATLANTIS The ATLANTIS project has received funding from the European Union's Horizon Europe framework programme under grant agreement No.101073909