EGI2024

Report of Contributions

Contribution ID: **1** Type: **Long Talk**

iMagine: an AI platform supporting aquatic science use cases

Thursday, 3 October 2024 09:10 (20 minutes)

Aquatic ecosystems are vital in regulating climate and providing resources, but they face threats from global change and local stressors. Understanding their dynamics is crucial for sustainable use and conservation. The iMagine AI Platform offers a suite of AI-powered image analysis tools for researchers in aquatic sciences, facilitating a better understanding of scientific phenomena and applying AI and ML for processing image data.
The platform supports the entire machine learning cycle, from model development to deployment, leveraging data from underwater platforms, webcams, microscopes, drones, and satellites, and utilising distributed resources across Europe. With a serverless architecture and DevOps approach, it enables easy sharing and deployment of AI models. Four providers within the pan-European EGI federation power the platform, offering substantial computational resources for image processing. Five use cases focus on image analytics services, which will be available to external researchers through Virtual Access. Additionally, three new use cases are developing AI-based image processing services, and two external use cases are kickstarting through recent Open Calls. The iMagine Competence Centre aids use case teams in model development and deployment. The talk will provide an overview of the development status of the use cases and offer insights on the platform.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary author: SIPOS, Gergely (EGI.eu)

Co-authors: LOPEZ GARCIA, Alvaro (CSIC); SCHAAP, Dick (Mariene Informatie Service MARIS BV); FAVA, Ilaria (EGI.eu); Dr KOZLOV, Valentin (Karlsruhe Institute of Technology)

Presenter: SIPOS, Gergely (EGI.eu)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **3** Type: **Demonstrations & Tutorials**

Autopoietic Cognitive Edge-cloud Services

Tuesday, 1 October 2024 18:00 (30 minutes)

Currently, data processing and analysis predominantly occur within data centers and centralized computing environments, with 80% of this activity centralized and only 20% happening through intelligent, connected devices. Additionally, merely one out of four European companies leverages cloud technologies, and non-EU entities control 75% of Europe's cloud market.

To leverage the shift towards edge computing, which aligns with Europe's strategies on data, the environment, and industry, it's crucial for Europe to consolidate significant investments. The emphasis should be on the creation and implementation of advanced computing components, systems, and platforms. These technologies are essential for facilitating the move to a computing continuum that boasts potent edge and far-edge capabilities, all while being energy-efficient and reliable.

The development of the edge to cloud continuum faces a number of technological and conceptual challenges. First, seamless, transparent and trustworthy integration of diverse computing and data environments spanning from core cloud to edge, in an AI-enabled computing continuum. Secondly, automatic adaptation to the growing complexity of requirements and the exponential increase of data driven by IoT deployment across sectors, users and contexts while achieving optimal use of resources, holistic security and data privacy and credibility. Finally, interoperability challenges among computing and data platform providers and cloud federation approaches based on open standards, interoperability models and open platforms.

To cope with those challenges, ACES will provide an edge-services cloud with hierarchical intelligence, specifically autopoiesis and cognitive behaviors to manage and automate the platform.

These solutions include:

- Autopoiesis-based edge-services cloud
- Awareness tools, AI/ML agents for workload placement, service and resource management, data and policy management, telemetry and monitoring
- Autopoiesis agents to safeguard stability in situations of extreme load and complexity
- Swarm technology-based methodology and implementation for orchestration of resources
- Edge-wide workload placement and optimization
- App store for classification, storage, sharing and rating of AI models used in ACES.

Such new solutions are tested through three use cases:

- The energy marketplace case study in Greece demonstrates how distributed edge services can autonomously match energy supply and demand across regions, promoting renewable energy use and optimizing resource distribution.
- Distributed energy grid process management: utilizing an edge mesh for the Greek energy grid decentralizes management, enhancing the use of local energy resources and adapting to consumption needs, shifting from centralized to a resilient, adaptive infrastructure with a user interface that aids operators in decision-making and intervention.
- An IoT based asset monitoring and management: the innovation aims to show that integrating Advanced Metering Infrastructure data, grid-edge sensors, and GIS systems can enhance outage detection, improve prediction accuracy, and support reliable investment planning, including deferral, by analyzing diverse IoT and operational data for asset life assessment.

Extended list of the consortium members includes:

- RTOs and Universities: INESC-ID Lisboa, Technical University of Darmstadt, Polytechnic University of Madrid, University of Ljubljana, The University of Applied Sciences and Arts of Southern Switzerland, Lakeside Labs
- Small and Medium Enterprises: Hiro Microdatacenters, Datapower Consulting, Martel Innovate, SixSq
- Public Agency: Hellenic Independent Power Transmission Operator

Topic

Environmental informatics: Green Computing

Primary author: Dr BUINING, Fred (HIRO Microdatacenters)

Co-author: Mr REMOTTI, Luca (DataPower Consulting)

Presenters: MUREDDU, Francesco (The Lisbon Council); Mr REMOTTI, Luca (DataPower Consulting)

Session Classification: Demonstrations & Posters

Contribution ID: **4** Type: **Poster**

CEDAR - Common European Data Spaces & robust AI for transparent Public Governance

Tuesday, 1 October 2024 18:00 (1 hour)

CEDAR is a brand new Horizon Europe projects whose key goal is to develop methods, tools, and guidelines to digitise, protect, and integrate data to address significant issues like corruption, aligning with the European Strategy for Data and the development of Common European Data Spaces (CEDS), and the European Data Act. This will lead to improved transparency and accountability in public governance, promoting European values and rights in the digital world, and enriching the European data ecosystem and economy.

We would be pleased to join the Poster Session to effectively showcase the main objectives and challenges tackled by the initiative as well as the three pivotal Pilot Studies undertaken to validate these efforts, taking place in 3 different European countries and addressing important domains for the public governance (such as monitoring national RRP (Recovery and Resilience Plan funds in Italy, transparent management of Slovenian public healthcare funds, and transparent management of foreign aid for rebuilding Ukraine).

Contacts
LinkedIn - https://www.linkedin.com/company/cedar-eu/
X - https://twitter.com/cedar_eu

Topic

Data innovations: Data Spaces

Primary author: OSIMANTI, Osimanti (The Lisbon Council)

Presenter: OSIMANTI, Osimanti (The Lisbon Council)

Session Classification: Demonstrations & Posters

Contribution ID: **5** Type: **Short Talk**

A path to future-compatibility to navigate the complexity of integrating AI-powered virtual sensing in Digital Twin

Tuesday, 1 October 2024 17:45 (15 minutes)

Digital Twin technology isn't a single monolithic software solution. It is a complex system that must adapt to varying and potentially unpredictable user needs. This adaptability is crucial in environments where data, models, and objectives are shared across different domains, sectors, organisations, and expertise groups and roles across the organisations. Striving to ensure Digital Twin applications and models are future-compatible and can cater to diverse requirements, a holistic approach to design and management is essential. We outline a strategy using a Platform-as-a-Service model for Digital Twinning Infrastructure Components. This approach enables AI-powered virtual sensing to support multiple Digital Twin applications and models, illustrated through a case study where groundwater level measurements are integrated into a digital twin of The Netherlands. We share the insights gained from developing this operational platform service and their implications for future services.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: Dr PILEGGI, Paolo (TNO)

Co-authors: Mr KESKIN, Serkan (TNO); Ms SHARMA, Shreshtha (TNO); Ms BEN AZIZA, Syrine (TNO); Mr TRANTAS, Thanasis (TNO)

Presenter: Dr PILEGGI, Paolo (TNO)

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **6** Type: **Short Talk**

CEDAR - Common European Data Spaces & Robust AI for transparent Public Governance

Tuesday, 1 October 2024 15:55 (15 minutes)

CEDAR is a brand new Horizon Europe projects whose key goal is to develop methods, tools, and guidelines to digitise, protect, and integrate data to address significant issues like corruption, aligning with the European Strategy for Data and the development of Common European Data Spaces (CEDS), and the European Data Act. This will lead to improved transparency and accountability in public governance, promoting European values and rights in the digital world, and enriching the European data ecosystem and economy.

The Consortium boasts nine top research institutions and universities, twelve technology and business developing companies, seven public sector end users, and three relevant NGOs. By sharing high-quality datasets, developing secure connectors for European data repositories, and employing innovative technologies for efficient big data management and analysis, CEDAR aims to promote better evidence-based decision-making, combat corruption, and reduce fraud in public administration.

In this short talk (10 minutes) we would like to present the key objectives of the project and, most prominently, the three Pilot Studies (co-located in three different EU member states) to effectively co-create and test the projects' outcomes in a relevant setting with the end users, as well as to validate the key CEDAR benefits, which are:

- Efficient, scalable, and trustworthy data technologies.
- Vast amounts of interoperable and analytics-ready data. -
- Collaboration of relevant public and private, local and regional stakeholders.
- Evidence based recommendations for legislative improvement in the public governance sphere.

Topic

Data innovations: Data Spaces

Primary authors: Mr MUREDDU, Francesco; OSIMANTI, Osimanti (The Lisbon Council)

Presenters: Mr MUREDDU, Francesco; OSIMANTI, Osimanti (The Lisbon Council)

Session Classification: Inside Data Spaces: Enabling data sharing paradigms

Contribution ID: **8** Type: **Long Talk**

The HPC+AI Cloud: flexible and performant infrastructure for HPC and AI workloads

Tuesday, 1 October 2024 17:05 (20 minutes)

In recent years, in particular with the rise of AI, the diversity of workloads that need to be supported by research infrastructures has exploded. Many of these workloads take advantage of new technologies, such as Kubernetes, that need to be run alongside the traditional workhorse of the large batch cluster. Some require access to specialist hardware, such as GPUs or network accelerators. Others, such as Trusted Research Environments, have to be executed in a secure sandbox.

Here, we show how a flexible and dynamic research computing cloud infrastructure can be achieved, without sacrificing performance, using OpenStack. By having OpenStack manage the hardware, we get access to APIs for reconfiguring that hardware, allowing the deployment of platforms to be automated with full control over the levels of isolation. Optimisations like CPU-pinning, PCI passthrough and SR-IOV allow us to take advantage of the efficiency gains from virtualisation without sacrificing performance where it matters.

The HPC+AI Cloud becomes even more powerful when combined with Azimuth, an open-source self-service portal for HPC and AI workloads. Using the Azimuth interface, users can self-service from a curated set of optimised platforms from web desktops through to Kubernetes apps such as Jupyter notebooks. Those applications are accessed securely, with SSO, via the open-source Zenith application proxy. Self-service platforms provisioned via Azimuth can co-exist with large bare-metal batch clusters on the same OpenStack cloud, allowing users to pi the environments and tools that best suit their workflow.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: PRYOR, Matt (StackHPC)

Co-author: Mr GARBUTT, John (StackHPC)

Presenter: PRYOR, Matt (StackHPC)

Session Classification: Bridging the Gap: Integrating the HPC Ecosystem

Contribution ID: **9** Type: **Long Talk**

Establishing and Verifying Trust for Data Products and Processing

Thursday, 3 October 2024 09:40 (20 minutes)

Establishing and Verifying Trust for Data Products and Processing

Motivation and Challenge

In today's infrastructures, the collection, exchange and continues processing of geospatial data takes place at pre-defined network endpoints of a spatial data infrastructure. Each participating operator hosts a predefined static functionality at a network endpoint. Some network endpoints of an operator may provide data access, other endpoints may provide processing functionality or uploading capabilities. Security context constraints are fundamental for installing services in production environments. Several legislations from security technical implementations guides to information security policies apply. Recent legislation like EU Data Act entered into force on 11 January 2024, and will become applicable in September 2025. Because of this regulation, connected products will have to be designed and manufactured in a way that empowers users (businesses or consumers) to easily and securely access, use and share the generated data. The EU DSA Data Service Act states applicability for simple websites, Internet infrastructure services and online platforms.

Approach

Our novel approach introduces an agile decentralized eco-system that is concerned with trust and authenticity by introducing Smart Certificates which can be applied to data products, workflow processes and services. The Smart Certificates enable the flexible and trustworthy creation, distribution and verification of data products. The certification process can either take place manually or automatically if the data has appropriate Identity-Integrity Provenance and Trust I2PT-enabling metadata.

Our approach differs from well-known X-509 certificates by schema definition of the information contained in Smart Certificates. The schema - hence the structure of the certificate - is stored on a Blockchain to become immutable. Supporting Zero-Knowledge-Proof as well as requesting information from the certificates during the verification procedure supports a wide range of use cases.

Example implementation: The Satellite Imagery Reprojection

For illustrating the use of Smart Certificates, a process - Reprojection - allows a user or a workflow engine to reproject an image for which the user has a Smart Certificate. For the output image, the process creates a Smart Certificate. The process itself is verifiable because it also has a Smart Certificate associated. The bundling of image data and Smart Certificates allows the process to check for authentic input but also to verify the usage of the image. If the usage is not appropriate for the trusted process, the execution in refused.

The Trusted Reprojection process was implemented in Python and deployed using the OGC API Processes Standard and a modified version of pygeoapi. The deployed process validates the input image Smart Certificates and creates a Smart Certificate for the created output product - the reprojected image. The Hyperledger Indy Blockchain and Aries Cloud Agent are used as the backbone.

Conclusion

Our approach introduces trusted computing in a distributed environment by leveraging Hyper-

ledger Indy, Aries Cloud Agent and specific business logic. The solution can be used to verify and issue Smart Certificates for data products and trusted processes. The introduced eco-system is one example solution to support the EU Data Act.

Topic

Trust and Security: Trusted computing:

Primary author: Dr MATHEUS, Andreas (Secure Dimensions GmbH)

Co-author: Dr COLAIACOMO, Lucio (European Union Satellite Centre)

Presenter: Dr MATHEUS, Andreas (Secure Dimensions GmbH)

Session Classification: Trust & Security

Contribution ID: **11** Type: **Long Talk**

Notify me: Updates about mytoken

Thursday, 3 October 2024 09:00 (20 minutes)

We present updates about the mytoken service, giving a short overview of the mytoken service, its idea and concept, and then focusing on the newest developments and future work.

These include the new notification feature, which allows user to obtain email notifications for various things, e.g. to be notified before a mytoken expires to easily create a new one. Also mytoken expirations can be integrated into a user's calendar application.

The mytoken software offers a central service to obtain OpenID Connect Access Tokens in an easy but secure way for extended periods of time and across multiple devices. In particular, mytoken was developed to provide OIDC Access Tokens to long-running compute jobs.

Mytokens can be restricted through the concept of "restrictions" and "capabilities" which allow very fine-grained access rights - much more detailed and flexible as OIDC tokens would allow.

Public instances are available at https://mytoken.data.kit.edu and https://mytok.eu
The latter running in a secure credential store environment.

Topic

Trust and Security: Access control

Primary authors: ZACHMANN, Gabriel (Karlsruhe Institute of Technology); Dr HARDT, Marcus (KIT-G)

Presenters: ZACHMANN, Gabriel (Karlsruhe Institute of Technology); Dr HARDT, Marcus (KIT-G)

Session Classification: Trust & Security

Contribution ID: **12** Type: **Short Talk**

SimpleVM - a framework for federated research environments in the de.NBI Cloud

Wednesday, 2 October 2024 15:40 (10 minutes)

Modern life sciences research has undergone a rapid development driven mainly by the technical improvements in analytical areas leading to miniaturization, parallelization, and high throughput processing of biological samples. This has led to the generation of huge amounts of experimental data. To meet these rising demands, the German Network for Bioinformatics Infrastructure (de.NBI) was established in 2015 as a national bioinformatics consortium aiming to provide high quality bioinformatics services, comprehensive training, powerful computing capacities (de.NBI Cloud) as well as connections to the European Life Science Infrastructure ELIXIR, with the goal to assist researchers in exploring and exploiting data more effectively.

Our de.NBI Cloud project type SimpleVM enables users with little to no background knowledge in cloud computing or systems administration to employ cloud resources with few clicks. SimpleVM is an abstraction layer on top of OpenStack to manage virtual machines (VMs) or clusters thereof. It was designed to support the combination of resources from independent OpenStack installations, thus operating as a federated multi-cloud platform which is accessible from a single web-based control panel. The entire software stack only requires access to 1) the OpenStack API and 2) an AAI provider (via Keycloak, LifeScience AAI) and it can be deployed on any vanilla OpenStack installation using Ansible. In general, SimpleVM primarily eases the creation and management of individual pre-configured virtual machines and provides web-based, SSO-protected access to popular research and development environments such as Rstudio, Guacamole Remote Desktop, Theia IDE, JupyterLab and Visual Studio Code. However, custom recipes based on Packer can be added to provide specific VMs tailored to user requirements.

A single SimpleVM project can host multiple VMs with individual access permissions for users. On top of this functionality, a dedicated SimpleVM Workshop mode streamlines virtual machine provisioning for workshops. Organizers can define a custom VM image and possible access methods, optionally based on the research environments mentioned above. When the workshop starts, participants can instantly access individual VMs based on this predefined configuration via ssh or browser. Once the VMs are ready, the system allows the organizers to automatically inform participants on how to access the resource.

Further, with SimpleVM, de.NBI Cloud users can effortlessly configure and manage their own SLURM-based BiBiGrid clusters with just a few clicks. This feature addresses the needs of researchers who want to run their tools or entire workflows across multiple machines and provides a simple route for users to learn how to use grid-based scheduling systems.
In summary, SimpleVM provides a comprehensive solution to bring federated, multi-cloud resources to end-users and in addition, provides a simple to use basis for online training and as an entry to grid-based computing.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: Dr HOFFMANN, Nils (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany); Mr BELMANN, Peter (Institute of

Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany)

Co-authors: Prof. SCZYRBA, Alexander (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany); Mr WEINHOLZ, David (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany); Ms MOK, Qiqi (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany); Mr RUDKO, Viktor (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany)

Presenter: Dr HOFFMANN, Nils (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany)

Session Classification: National Perspectives: EGI Member Countries' Latest Developments and Future Initiatives

Contribution ID: **13** Type: **Poster**

The de.NBI Cloud Federation

Tuesday, 1 October 2024 18:00 (1 hour)

In recent years, modern life sciences research underwent a rapid development driven mainly by the technical improvements in analytical areas leading to miniaturization, parallelization, and high throughput processing of biological samples. This has driven the growth and number of experimental datasets immensely, requiring scalable platforms for large scale data analysis beyond the capabilities of individual labs and training to effectively use such platforms. The German Network for Bioinformatics Infrastructure (de.NBI) was established in 2015 as a national bioinformatics consortium aiming to provide high quality bioinformatics services, comprehensive training, and with the de.NBI Cloud, powerful cloud-based computing capacities to address these requirements. de.NBI further provides its portfolio as the designated German node of the European Life Science Infrastructure ELIXIR [3].

The de.NBI Cloud is one of the flagship services of the de.NBI network. It consists of eight federated cloud locations that implement a common governance and use the project application and management workflow provided by the de.NBI Cloud portal. Registration, project resource application and authentication are facilitated by the integration of the LifeScience AAI as an EduGAIN-compatible single sign-on provider, backed by institutional ID providers of universities and research institutes.

The de.NBI Cloud portfolio includes several project types designed to suit different use cases and users with varying levels of knowledge in cloud computing. Two project types, OpenStack and Kubernetes, offer maximum flexibility in terms of the configuration of cloud-specific components and allow the installation of any large-scale analysis, stream processing or orchestration framework available in the cloud ecosystem. Both project types are ideal for science gateway developers to offer bioinformatics services to the national and international life sciences communities, like the Competence Center Cloud Technologies for Data Management and Processing (de.KCD), the National Research Data Initiative (NFDI), and EOSC-Life on the European level.

The project type SimpleVM enables users to employ cloud resources with little to no background knowledge in cloud computing or systems administration. SimpleVM performs as an abstraction layer on top of OpenStack to manage virtual machines (VMs) or clusters thereof. It is designed to support the combination of resources from independent OpenStack installations, thus operating as a federated multi-cloud platform which is accessible from a single web-based control panel.

For users who aim for the ability to define data processing workflows from tools available in BioConda and the Galaxy ToolShed with a graphical user interface, the de.NBI Cloud infrastructure also hosts the Galaxy service available at usegalaxy.eu. Galaxy simplifies the discovery and adaptation of existing workflows, that were shared by other users, from multiple scientific domains and enables their execution at scale in the cloud.

In conclusion, the de.NBI Cloud provides the ability to unlock the full potential of research data and enables easier collaboration across different ecosystems and research areas, which in turn enables scientists to innovate and scale-up their data-driven research, not only in the life and computational biosciences, but across different science domains.

Topic

Needs and solutions in scientific computing: National and scientific perspectives

Primary author: Mr RUDKO, Viktor (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany)

Co-authors: BELMANN, Peter (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany); WEINHOLZ, David (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany); MOK, Qiqi (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany); Prof. GOESMANN, Alexander (Justus-Liebig-University Giessen, Giessen, Germany); Prof. EILS, Roland (BIH-Zentrum Digitale Gesundheit, Charité -Universitätsmedizin Berlin, Berlin, Germany); Dr BORK, Peer (European Molecular Biology Laboratory, Heidelberg, Germany); Prof. KOHLBACHER, Oliver (Eberhard Karls University Tübingen, Tübingen, Germany); Prof. KUMMER, Ursula (BioQuant & Heidelberg University, Heidelberg, Germany); Prof. BACKOFEN, Rolf (Albert-Ludwigs University Freiburg, Freiburg, Germany); Dr BUCHHALTER, Ivo (Deutsches Krebsforschungszentrum DKFZ Heidelberg, Heidelberg, Germany); Prof. STOYE, Jens (Bielefeld University, Bielefeld, Germany); Dr HOFFMANN, Nils (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany); SCZYRBA, Alexander (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany)

Presenter: Mr RUDKO, Viktor (Institute of Bio-and Geosciences, IBG-5, Computational Metagenomics, Forschungszentrum Jülich, Bielefeld, Germany)

Session Classification: Demonstrations & Posters

Contribution ID: **14** Type: **Long Talk**

Exploring Climate Data Analysis with MATLAB in the ENES Data Space

Tuesday, 1 October 2024 15:15 (20 minutes)

The escalating volume and complexity of Earth and environmental data necessitate an effective, interdisciplinary partnership among scientists and data providers. Achieving this requires the utilization of research infrastructures that offer sophisticated e-services. These services enhance data integration and interoperability, enable seamless machine-to-machine data exchanges, and leverage High-Performance Computing (HPC) along with cloud capabilities.
In this presentation, we will demonstrate a case study focused on the import, analysis, and visualization of geodata within the ENES Data Space (https://enesdataspace.vm.fedcloud.eu), a cutting-edge cloud-enabled data science environment designed for climate data analysis. This platform is ingeniously constructed atop the European Open Science Cloud (EOSC) Compute Platform. By integrating with either an institutional or social media account, users gain entry to the ENES Data Space. Here, they can initiate JupyterLab, accessing a personal workspace equipped with computational resources, analytical tools, and pre-prepared climate datasets. These datasets, which include historical data recording and future projections, are primarily sourced from the CMIP (Coupled Model Intercomparison Project).
Our case study will utilize global precipitation data derived from the Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC) experiments, analyzed within the ENES workspace through two distinct approaches:
1. Direct MATLAB Online Integration: Users can launch MATLAB Online directly from the ENES Data Space JupyterLab. Utilizing a Live Script (.mlx), the process involves importing, filtering, and manipulating data, creating visual maps, comparing results, and conducting hypothesis testing to ascertain the statistical significance of the project findings. Live Scripts serve as interactive notebooks that facilitate the clear articulation of research methodologies and goals by integrating data, hyperlinks, figures, text, and code. These scripts also incorporate UI tools for intuitive, point-and-click data analysis and visualization, eliminating the need for extensive programming expertise.
2. MATLAB Kernel within Jupyter Notebook: This method demonstrates the analysis process using a MATLAB kernel executed from a Jupyter notebook (.ipynb) within the same JupyterLab environment.
In both scenarios, the results can be exported in multiple formats (e.g., PDF, markdown, LaTeX, etc.), allowing for easy downloading and sharing with other researchers, educators, and students. This entire workflow is seamlessly executed in MATLAB within the ENES Data Space, without the need for software installation or data downloads on local (non-cloud) devices. This case study exemplifies the power of cloud-based platforms in enhancing the accessibility, efficiency, and collaborative potential of climate data analysis.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: LEPTOKAROPOULOS, Kostas (MathWorks); Dr CHAKRABARTI, Shubo (Math-

Works); ANTONIO, Fabrizio (CMCC)

Presenter: LEPTOKAROPOULOS, Kostas (MathWorks)

Session Classification: Unlocking the Potential of Environmental Data

Contribution ID: **15** Type: **Demonstrations & Tutorials**

Parallel Inference in the Edge-to-Cloud for Health Monitoring

Tuesday, 1 October 2024 18:00 (30 minutes)

In the context of Artificial Intelligence (AI), the evolution of computing paradigms from centralized data centers to the edge of the network heralds a transformative shift in how AI applications are developed, deployed, and operated. Specifically, the edge computing paradigm is characterized by processing data directly in the devices where it is collected, such as smartphones, wearables, and IoT. Edge computing significantly reduces latency, conserves bandwidth, and enhances data privacy and security, thereby catalyzing the realization of real-time, responsive AI solutions.
The AI-SPRINT (Artificial Intelligence in Secure PRIvacy-preserving computing coNTinuum) H2020 project has developed a comprehensive design and runtime framework dedicated to accelerating the development and deployment of AI applications across the computing continuum. This continuum spans from cloud data centers to edge devices and sensors, integrating AI components to operate seamlessly and efficiently. The project's core objective is to offer a suite of tools that enable an optimal balance among application performance, energy efficiency, and AI model accuracy, all while upholding rigorous security and privacy standards.

The Personalized Healthcare use case focuses on harnessing the power of AI and wearable technologies for health monitoring. By integrating quantitative data on heart functions from wearable device sensors with qualitative lifestyle information, the use case aims to develop a personalized stroke risk assessments model. This initiative is particularly critical given the prevalence of stroke among the aging population, marking it as a significant cause of death and disability globally. Leveraging the AI-SPRINT framework, this application enables efficient resource distribution and computation across the edge-to-cloud continuum, facilitating real-time, non-invasive and secure monitoring and risk assessment.
The use case has been implemented using the PyCOMPSs programming framework from BSC to develop a Machine Learning model for detecting atrial fibrillation (AF) in electrocardiogram data (ECG) implemented using the parallel dislib library built on top of PyCOMPSs.
In this demo we show how the prediction risk is calculated, using ECGs extracted from a wearable device, on a Cloud server deployed in the EGI Cloud, with resources provided on-demand by the OSCAR framework from UPV.
OSCAR is an open-source platform to support the serverless computing model for event-driven data-processing applications. It can be automatically deployed on multiple Cloud backends, thanks to the Infrastructure Manager, to create highly parallel event-driven data-processing serverless applications that execute on customized
runtime environments provided by Docker containers than run on an elastic Kubernetes cluster.
In the demo, the OSCAR cluster is deployed and configured with a MinIO storage. When new data (ECG files) is sent through an HTTP request by uploading the ECG file to MinIO, OSCAR triggers a PyCOMPSs/dislib container creation to serve the execution of the inference computation in a Function as a Service mode. This is implemented through a script that starts a PyCOMPSs instance and uploads the result data back in the storage. The number of resources used in the execution can be configured dynamically through an integration with OSCAR and the COMPSs runtime through environment variables.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary authors: ALARCON MARIN, Caterina (Universitat Politècnica de València); LEZZI, Daniele (Barcelona Supercomputing Center); Dr CIRILLO, Davide (Barcelona Supercomputing Center); Dr LORDAN-GOMIS, Francesc (Barcelona Supercomputing Center); MOLTO, German (Universitat Politècnica de València); CABALLER, Miguel (Universitat Politècnica de València)

Presenter: LEZZI, Daniele (Barcelona Supercomputing Center)

Session Classification: Demonstrations & Posters

Contribution ID: **16** Type: **Poster**

UHI-Stream: A User-Friendly, Cloud-Based Tool for Rapid Analysis of Urban Heat Island Effect Changes Anywhere On Earth

Tuesday, 1 October 2024 18:00 (1 hour)

The term "Urban Heat Island" (UHI) effect describes the phenomenon where urban environments exhibit higher air temperatures than their rural counterparts, a difference that is especially pronounced at night. This effect arises from the greater capacity of urban materials and man-made structures, such as buildings and pavements, to absorb, store, and then re-radiate heat compared to natural materials and landscapes.

First identified over two centuries ago, the UHI effect is still subject of research to understand, measure, and mitigate its impacts on society as a whole, and in particular on economic activities and public health. Although traditionally the prerogative of specialists, the UHI is also attracting increasing interest among citizens. However, not all have the necessary technical expertise or infrastructure access to source relevant data (from in-situ measurements, satellite remote sensing or numerical models), process it efficiently, synthesize it and interpret the changes over time or between different locations.

The UHI-Stream tool was specifically developed to bridge this gap and quickly analyze temperature differences between two points anywhere on Earth's by leveraging EGI compute and storage resources (owned by CESNET) and ERA5-Land reanalysis data (available from 1950, as part of the Copernicus Climate Change Service). The corresponding hourly 2m air temperatures are streamed from S3 buckets, processed on-the-fly and visualized as annual heat-maps or animations spanning user-defined time-frames.

Conveniently hosted on RoHub as a FAIR (Findable, Accessible, Interoperable, and Reusable) Executable Research Object, UHI-Stream is expected to be further converted into a Galaxy tool with a Graphical User Interface as part of the EuroScienceGateway project, potentially incorporating additional features to help users pinpoint representative urban and adjacent rural areas, or account for more grid cells.

In summary, UHI-Stream is poised to become a valuable asset in urban climatology studies, enabling easier identification of UHI patterns and estimating climate impacts on a regional scale. The tool's versatility in analyzing any two geographic points enhances its usefulness beyond the mere urban-rural context, allowing for comparative analyses of temperature changes across diverse locales, regardless of their relationship.

Topic

Environmental informatics: Climate Change/Environment

Primary authors: IAQUINTA, Jean; FOUILLOUX, Anne (Simula Research Laboratory)

Presenter: IAQUINTA, Jean

Session Classification: Demonstrations & Posters

Contribution ID: **17** Type: **Poster**

Climate Change Impact on Renewable Energy Output in Türkiye: Insights from Ensembling Global Climate Models with Extreme Gradient Boosting Regression Trees

Tuesday, 1 October 2024 18:00 (1 hour)

The heightened focus on global warming and climate change has prompted a substantial shift towards green energy technologies, which are crucial in shaping electricity generation capacity. Turkey has actively been investing in renewable energy sources, such as wind and solar, to reduce its dependency on imported fossil fuels and improve its energy security. This study investigates the future of electricity production in Turkey under a changing climate using climate model projections and a machine learning algorithm. The aim was to identify the most suitable Global Climate Models (GCMs) in simulating Turkey's climate conditions and evaluate how climate change, considering changing wind speeds, solar radiation, and temperature, will impact future electricity production in renewable energy output. Historical data from 13 CMIP6 Global Climate Models was acquired, focusing on temperature, wind speed, and solar radiation parameters. Model resolution was standardized, and daily data for 120 grids in Turkey were collected for the period 2010-2014. The performance of GCMs was assessed against ERA5/CRU-biased corrected datasets using metrics such as Kling-Gupta efficiency (KGE), modified index of agreement (md), and normalized root mean square error (nRMSE). A Multiple-criteria Decision Analysis (MCDA) method ranked the models based on performance, and Comprehensive rating metrics (MR) provided a unified score. The top-performing models (ACCESS-CM2, INM-CM5–0, INM-CM4–8, and ACCESS-ESM-1-5) were ensembled and utilized to predict Turkey's future climate using the Extreme Gradient Boosting Tree (XGBoost) algorithm. Daily data from 2010-2013 served as the train dataset, while 2014 daily data was set as the test dataset. Following the grid search for the optimization of XGBoost model parameters in each grid, projections of each climate variable were made for 2020-2064 under the SSP5-8.5, SSP3-7.0, and SSP2-4.5 scenarios. To evaluate the wind energy potential of each grid, the Wind Power Density method was utilized by recalculating the forecasted wind speed outputs from GCMs at 10m to 100m height using the wind power profile law. Additionally, the electricity production potential from solar PV systems in each grid was assessed using efficiency correlation coefficients from Evans-Florscheutz, which consider factors such as wind speed, ambient temperature, and solar radiation. The findings of this study provide valuable insights into Turkey's future electricity production landscape under the influence of climate change and the transition to green energy technologies. This information can aid the government in determining future energy policies more accurately and enable independent power producers to make investment decisions more precisely.

Topic

Environmental informatics: Climate Change/Environment

Primary author: Mr GUVEN, Denizhan (Istanbul Technical University)

Co-author: Prof. KAYALICA, M. Ozgur (Istanbul Technical University)

Presenter: Mr GUVEN, Denizhan (Istanbul Technical University)

Session Classification: Demonstrations & Posters

Contribution ID: **18** Type: **Short Talk**

Why does document structuring still matter and struggle in the era of GenAI and Large Language Models?

Wednesday, 2 October 2024 11:00 (15 minutes)

Document structuring is a fundamental aspect of information management, involving the categorization and organization of documents into logical and physical structures. This presentation explores the benefits and challenges associated with document structuring, focusing on the distinctions between physical and logical structures, metadata and content, as well as addressing the implications for businesses and research centers dealing with large volumes of data encompassed in *data warehouses* and *lakes* of textual documents.

In the task of document structuring, distinctions arise between physical and logical structures. Physical structures pertain to the layout and presentation of documents, encompassing elements such as tables, figures, and images. On the other hand, logical structures refer to the organization of content within documents, including metadata that describes document attributes and content that comprises the textual information.

Implementing structured document management systems brings several benefits for business and research bodies. Firstly, structured documents target search queries more effectively, yielding more relevant search results and reducing the volume of irrelevant hits. This not only enhances search efficiency but also saves time and resources, resulting in cost savings and eco-friendly practices. Additionally, structured documents facilitate comparisons between similar structures, enabling deeper analysis and insights. Moreover, the adoption of structured documents enables the extraction of statistics and the creation of dashboards, as it allows for the identification and analysis of document elements beyond mere text.

However, document structuring still faces great challenges. Legacy documents pose a significant hurdle, particularly those with poor scans or generated through low-quality optical character recognition (*OCR*). These documents may contain noise, artifacts, or degradation, compromising the accuracy of structure recognition algorithms. Furthermore, complex layouts, heterogeneous documents, handwritten content, tables, figures, images, multilingual text, and dynamic content all contribute to the complexity of document structuring. Moreover, the scarcity of labeled data exacerbates the challenge, hindering the development of accurate and robust structuring algorithms.

While *Generative Artificial Intelligence* (*GenAI*) has demonstrated remarkable capabilities in various domains, including natural language understanding and image recognition, it still struggles with document structuring due to the complexity of document layouts, ambiguity in content, and limited contextual understanding. It faces challenges in handling diverse document formats, noisy data, and legacy documents. Additionally, *GenAI*'s reliance on labeled data for training limits its generalization ability, hindering its performance on unseen document structures. Overcoming these challenges requires interdisciplinary collaboration and continued research to develop more robust *Artificial Intelligence* (*AI*) models capable of effectively managing the complexities of document organization and content extraction.

In conclusion, document structuring offers substantial benefits for businesses and research centers, enabling more efficient information retrieval, automated data extraction, enhanced searchability, standardization, and improved data analysis. However, overcoming these challenges requires innovative solutions and advancements in document structuring technology. By addressing these challenges, organizations can harness the full potential of structured documents to optimize workflows, facilitate knowledge management, and drive innovation.

Topic

Data innovations: Data Management/Integration/Exchange

Primary author: Dr KHEMAKHEM, Mohamed (MandaNetwork)

Presenter: Dr KHEMAKHEM, Mohamed (MandaNetwork)

Session Classification: Maximizing Data Efficiency and Availablity: Innovations in Data Management Platforms

Contribution ID: **19** Type: **Short Talk**

Clouds, competence centres and research infrastructures in the cultural heritage domain

Wednesday, 2 October 2024 16:50 (10 minutes)

The proliferation of terms used to characterize digital infrastructures providing services for cultural heritage reflects different perspectives on their role but may be confusing for their potential users. Deciding what is a research infrastructure (RI) and what is not may affect stakeholders'understanding of them and the related use, and impact on their funding, where support to their creation and management depends on the legal definition of what is eligible and what is not. Adding the qualification of "research"to such infrastructures further complicates the situation, as it also involves the understanding of what research-oriented means. Cultural heritage is a domain in which the term "research"has a broad connotation, including activities performed according to the methodology of social sciences and humanities and to the scientific method (within the so-called heritage and archaeological sciences) and requires a tailored approach to digitization. Important research results can be achieved both by professionals and by researchers, thus also the target user communities can have fuzzy delimitations.

Attempts to identify different approaches to research date back to the well-known study by Stokes, which distinguished among fundamental research, use-inspired research and solution-oriented applied research, with most of the research activities in cultural heritage belonging to the second and the third categories. This distinction was considered in a study carried out in 2018 by the RI-PATHS EU project, which included competence centres in the third grouping but also considered the case of multi-purpose RIs. Digital infrastructures are not considered separately in this RI-PATHS taxonomy, or just mentioned as data and service providers. This requires an update, since research activities now increasingly consist or rely on digital services. Notable examples of digital research infrastructures on cultural heritage are national initiatives such as DIGILAB.be, a Belgian data infrastructure for conservation, and HSDS in the UK, a similar one in the UK. Similarly, the ARIADNE RI provides a catalogue of 4,000,000 archaeological research datasets with services to process the data they contain. A European initiative is 4CH, implementing a European competence centre for heritage conservation relying on a knowledge base of heritage information. Thus the question moves to the digital backbones of the above and how they could be further developed to reap the benefits of an advanced digital transformation. The just started ECCCH project aims at developing cloud-based vertical services for the different heritage communities, and will deploy its results by 2029.

So the main question seems to be not just which services are required, but in which cloud environment they should be implemented. If answering to the question "is a cloud the place to develop a digital research infrastructure for cultural heritage"has an obvious positive answer, the features of such cloud still need clarification and further investigation. The present contribution will discuss this question, with a focus on the digital services to be provided by 4CH in its development as an international digital research infrastructure and competence centre.

Topic

Data innovations: Business models

Primary author: Dr PEZZATI, Luca (CNR)

Co-author: NICCOLUCCI, Franco (ARIADNE RI)

Presenter: Dr PEZZATI, Luca (CNR)

Session Classification: Powering Collaboration: Technical Computing and Data Continuum Requirements

Contribution ID: **20** Type: **Poster**

Docker container in DWD's Seamless INtegrated FOrecastiNg sYstem (SINFONY)

Tuesday, 1 October 2024 18:00 (1 hour)

At Deutscher Wetterdienst (DWD) the SINFONY project has been set up to develop a seamless ensemble prediction system for convective-scale forecasting with forecast ranges of up to 12 hours. It combines Nowcasting (NWC) techniques with numerical weather prediction (NWP) in a seamless way for a number of applications. Historically the operational NWC and NWP forecasts are generated on separate IT-Infrastructures, which in turn increases the number of potential error sources. To reduce data transfer between both infrastructures and to reduce the complexity of SINFONY those NWC components, which solely rely on NWP pre-products, are ported to the NWP infrastructure using software container.
With this aim in view a container image containing all relevant NWC components is created in a CICD oriented procedure. The respective containers are integrated into DWD's development and operational code bases and executed on DWD's HPC using apptainer. The integration into DWD's development code base is completed already and currently used for further development of the data assimilation procedure.
A major innovation of SINFONY is the rapid update cycle (RUC), an hourly refreshing NWP procedure with a maximum lead time of 12 hours. Currently RUC is in a preoperational stage and the subsequent generation of SINFONY products combining NWP and NWC forecasts is still executed on the NWC infrastructure. The RUC will be implemented to the operational forecasting system at the end of 2024 and together with that our aim is to implement the containers to the RUC for applications that rely on NWP data. At this step the container implementations have to meet even harder requirements in terms of performance, update reliability and support, since it will then be part of Germanys national critical infrastructure.

Topic

Topic not listed

Primary author: ZACHARUK, Matthias (Deutscher Wetterdienst)

Presenter: ZACHARUK, Matthias (Deutscher Wetterdienst)

Session Classification: Demonstrations & Posters

Contribution ID: **21** Type: **Long Talk**

Enlivening the Heritage Digital Twin

Tuesday, 1 October 2024 16:45 (10 minutes)

The term "digital twin"has been used to designate 3D models of physical cultural artefacts to which additional information might be added. If the 3D model consisted in a point cloud, as in the case of generating it via scanning, such information was attached to its points or regions as a sort of Post-it, thus creating so-called "augmented objects". When, instead, CAD systems are used to produce the 3D model, the extra data were incorporated in an extension of BIM (Building Information Modelling) called HBIM (Heritage BIM), which adds the heritage-related necessary classes to BIM, an ISO standard used in the construction industry to incorporate information about materials, services and processes of a building.
In 2023 we proposed a novel ontology for heritage information based on the Heritage Digital Twin (HDT) a holistic approach to heritage information where the 3D graphical component is just one element. It allows to document intangible heritage as well, where the visual documentation may consist in video or audio recordings or even be totally absent. Such ontology, named HDTO, is a compatible extension of CIDOC-CRM, the standard for heritage documentation, allowing a straightforward incorporation of existing data organized according to it. The HDTO has been used to set up the cloud-based Knowledge Base (KB) created in 4CH, an EU-funded project designing a Competence Centre for the Conservation of Cultural Heritage. Documentation in the 4CH KB includes the relevant information about heritage assets, from visual one to the results of scientific analyses, conservation activities and historical documents.
The HDT does not consider the dynamic and interactive aspects connecting a digital twin to reality. The proposed improved model, named Reactive HDT Ontology (RHDTO), includes the documentation of dynamic interactions with the real world. A first example of application concerns the Internet of Cultural Things (IoCT), i.e. the use of IoT in the cultural heritage domain, for example fire sensors based on smoke or heat and other environmental sensors, activating processes and reactions. But the connection with reality may also consist in data directly provided by external digital systems, such as those providing weather forecasts or monitoring landslide hazards. The "reactive"nature of the system consists in three steps: an *input/sensor*, receiving data from the real world and processing them; the resulting outcome is input into a *decider*, which then transmits orders to an *activator*: each of them is documented as a member of a digital process RHDTO class and the related process is described in a specific instance of it. Such instances vary according to the nature of the planned reaction and are programmed according to scientific or heuristic knowledge about the relevant phenomenon, which may also be stored in the KB. The system may be connected and receive inputs from larger models such as the Digital Twin of the Earth, the ECMWF or the CMCC. Finally, the system allows also "what-if"simulation to experiment risks and mitigating measures, by defining simulated deciders and activators and providing as outcomes the simulation results.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary authors: Mr FELICETTI, Achille (PIN); NICCOLUCCI, Franco (ARIADNE RI)

Presenter: NICCOLUCCI, Franco (ARIADNE RI)

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **22** Type: **Short Talk**

Virtual Data Center: a platform for enabling data scientists to access integrated and scalable online environments

Wednesday, 2 October 2024 17:05 (10 minutes)

The advancement of **EOSC** promotes a research paradigm more reproducible and verifiable in response to the growing complexity and interdisciplinarity of modern research, necessitating an unprecedented level of collaboration and data sharing. In line with this, federated data infrastructures, like the **Blue-Cloud project**, have been established, integrating marine data sources across Europe to catalyze advancements in marine science. Among these initiatives, the *NEw REsearch Infrastructure Datacenter for EMSO* (**NEREIDE**) developed by INGV and located near the Western Ionian Sea facility, is designed to drive data science forward through its Virtual Data Center (VDC) platform.

The core of NEREIDE innovation is the **Virtual Data Center**, a managed environment where Data Scientists (DSs) can control cloud resources, including Virtual Machines (VMs), behind a dedicated customizable **Gateway** with a public IP address. DSs have administrative privileges over their cloud segments, enabling them to **autonomously** create VMs and manage network components including firewalls, VPNs and advanced routing configurations. Meanwhile, the overarching management of cloud infrastructure, including the physical data center, remains in charge and under the control of the Infrastructure Administrators. This **dual-structure** ensures a balance between stringent infrastructural measures and DSs operational autonomy. VDCs provide a sophisticated, plug-and-play infrastructure that sidesteps the complexities of traditional data center management, allowing DSs to focus on their **data services**, bypassing the complexities of traditional physical data center management.
DSs want to work into dynamic and customizable environments where they manage substantial computational resources to tackle complex scientific questions. The main component of VDCs is **OpenStack**, an open-source "Infrastructure as a Service"platform that enables seamless scalability and flexibility in resource management, aided by workflow automation tools such as **MaaS** and **JuJu**. This setup allows DSs to optimize computing and storage capacities according to project needs, essential for processing extensive datasets and performing complex simulations.
VDCs rely also on **Ceph**, a distributed software defined storage engine, which offers flexible and scalable storage resources, in conjunction with data security and integrity. This solution allows DSs to face heavy scientific data loads and to efficiently manage multiple storage types.
Additionally, VDCs not only provide bare computational power and raw storage space to DSs, but also enable the integration of sophisticated arrays of tools, such as **JupyterHub** for interactive data analysis, **ERDDAP** for data distribution, **ElasticSearch** for data querying, etc. These tools underpin data science activities including data analysis, visualization, and collaborative research, thereby making complex data comprehensible and accessible across various scientific domains.

Essentially, with the introduction of a custom gateway, VDCs represent an evolution of the concept of virtualization, extending it beyond individual virtual machines to include an entire ready-to-use data center infrastructure.

The potential integration of VDC platforms within federated data infrastructures, like Blue-Cloud, suggests a future where seamless data and resource sharing could significantly boost the analytical and operational capacities within different scientific domains. These advancements foster new scientific applications and innovations, accelerating the achievement of open science goals and easing the work of data scientists.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: CACCIAGUERRA, Stefano (INGV); Dr CHIAPPINI, Stefano (INGV)

Presenters: CACCIAGUERRA, Stefano (INGV); Dr CHIAPPINI, Stefano (INGV)

Session Classification: Simplifying Data-Driven Science with User-Friendly Platforms and Gateways

Contribution ID: **23** Type: **Long Talk**

Beskar Cloud: status update

Tuesday, 1 October 2024 15:25 (15 minutes)

Last year, we introduced Beskar cloud - an open-source community around deploying and maintaining OpenStack cloud on top of Kubernetes cloud. Since then, we have successfully built two OpenStack sites and seamlessly transitioned users from our original OpenStack instance to the new environment built on Beskar Cloud.

In this presentation, we aim to provide an overview of our progress throughout the past year, detailing the advancements made within the project. Additionally, we will share insights gained from our experiences with migrations and day-to-day operations.

Topic

Needs and solutions in scientific computing: Federated operation

Primary authors: Mr ROSINEC, Adrian (CESNET); MORAVCOVA, Klara

Presenters: Mr ROSINEC, Adrian (CESNET); MORAVCOVA, Klara

Session Classification: Cloud Compute federation and national initiatives

Contribution ID: **24** Type: **Short Talk**

Updates on SSH with OpenId Connect

Thursday, 3 October 2024 10:00 (20 minutes)

The Secure Shell (SSH) Protocol is widely recognized as the de-facto standard for accessing remote servers on the command line, across a number of user cases, such as: remote system administration, git operations, system backups via rsync, and high-performance computing (HPC) access.

However, as federated infrastructures become more prevalent, there is a growing demand for SSH to operate seamlessly and securely in such environments. Managing SSH keys in federated setups poses a number of challenges, since SSH keys are trusted permanently, can be shared across devices and teams, and do not have a mechanism to enforce the use of passphrases. Unfortunately, there is currently no universally accepted usage pattern for globally federated usage.

The large variety of users with different backgrounds and usage profiles motivated us to develop a set of different tools for facilitating the integration with federated user identities. The main novelty that will be presented in this contribution is the integration of an SSH-certificate-based mechanism into the existing ecosystem for SSH with OpenId Connect, consisting of motley-cue and oidc-agent.

This new mechanism consists of a set of programs collectively referred to as “oinit”. It aims to simplify the usage of SSH certificates by leveraging authorization information via established federation mechanisms. The main benefit is that, after an initial setup step, SSH may be used securely without interrupting existing flows, enabling the use of rsync, for example.

The core components of oinit include the following:

- oinit-ca: this component provides a REST interface to an SSH Certificate Authority (CA), allowing authorized users to obtain SSH certificates for specific hosts or host groups. Authorization decisions are handled by motley-cue, the component that enables federated use of SSH on the ssh-server side. User provisioning may also be triggered at this point, via motley-cue and feudal.
- oinit: client-side tool employed by users to add hosts to the oinit mechanism. Once configured, SSH certificates are automatically retrieved as needed and stored in the SSH agent.
- Server-side tools and configuration: these enable SSH usage without requiring knowledge of local usernames, a particularly useful feature in federated scenarios.

In addition to outlining the architecture and functionality of our solution, we provide an initial security assessment and offer a live demo of SSH with OpenID Connect, with oinit and selected components.

Topic

Trust and Security: Trusted computing:

Primary authors: GUDU, Diana (KIT); ZACHMANN, Gabriel (Karlsruhe Institute of Technology); BROCKE, Lukas (KIT); Dr HARDT, Marcus (KIT-G)

Presenter: GUDU, Diana (KIT)

Session Classification: Trust & Security

Contribution ID: **25** Type: **Demonstrations & Tutorials**

Rancher: the EGI container execution platform

Wednesday, 2 October 2024 14:30 (30 minutes)

The EGI Cloud Container Compute is a container orchestrator that facilitates the deployment of containerised applications. Containers offer the advantage of having the entire software and runtime environment in a single package, which can simplify the deployment process. However, deploying a containerised application can be challenging due to the need to install a container orchestrator. This often requires the user to manage the entire virtual machine with all the required system services, which can be time-consuming and complex. In recent years, the adoption of the Kubernetes container orchestrator has led to a notable improvement in the efficiency of both developers and system administrators. This trend is also being observed in scientific computing, where containers are being tailored for use in federated computing environments to facilitate the execution of scientific workloads. This demonstration will introduce the EGI Cloud Container Compute Service, a managed platform that provides seamless container execution.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: Mr ROSINEC, Adrian (CESNET); MORAVCOVA, Klara

Presenters: Mr ROSINEC, Adrian (CESNET); MORAVCOVA, Klara

Session Classification: Demonstrations & Posters

Contribution ID: **26** Type: **Demonstrations & Tutorials**

Leveraging MLflow for Efficient Evaluation and Deployment of Large Language Models

Thursday, 3 October 2024 12:30 (30 minutes)

In recent years, Large Language Models (LLMs) have become powerful tools in the machine learning (ML) field, including features of natural language processing (NLP) and code generation. The employment of these tools often faces complex processes, starting from interacting with a variety of providers to fine-tuning models of a certain degree of appropriateness to meet the project's needs.
This work explores in detail using MLflow 1 in deploying and evaluating two notable LLMs: Mixtral2 from MistralAI and Databricks Rex (DBRX) 3 from Databricks, both available as open-source models in the HuggingFace portal. The focus lies on enhancing inference efficiency, specifically emphasising the fact that DBRX has better throughput than traditional models of similar scale.
Hence, MLflow offers a unified interface for interacting with various LLM providers through the Deployments Server (previously known as "MLflow AI Gateway") [4], which streamlines the deployment process. Further, with standardised evaluation metrics, we present a comparative analysis between Mixtral and DBRX.
MLflow's LLM Evaluation tools are designed to address the unique challenges of evaluating LLMs. Unlike traditional models, LLMs often lack a single ground truth, making their evaluation more complex.
MLflow allows customers to use a bundle of tools and features that are specifically tailored to deal with difficulties arising from integrating LLMs in a comprehensive manner. The MLflow Deployments Server serves as the central location, eliminating the need to juggle multiple provider APIs and simplifying integration with self-hosted models.
We plan to implement this solution using the MLflow tracking server deployed in the AI4eosc project [5] as a showcase.
In conclusion, this contribution seeks to offer insights into the efficient deployment and evaluation of LLMs using MLflow, with a focus on optimising inference efficiency through a unified user interface. With MLflow capabilities, developers and data scientists can navigate through integrating LLMs into their applications easily and effectively, unlocking their maximum potential for revolutionary AI-driven solutions.

1 https://mlflow.org

2 https://huggingface.co/mistralai

3 https://huggingface.co/databricks

[4] https://mlflow.org/docs/latest/llms/index.html

[5] https://ai4eosc.eu

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: BERBERI, Lisana (KIT-G)

Co-authors: Mr ESTEBAN SANCHIS, Borja (Scientific Computing Centre, Karslruhe Institute of Technology); Dr ALIBABAEI, Khadijeh (Scientific Computing Centre, Karslruhe Institute of Technology); Dr KOZLOV, Valentin (Scientific Computing Centre, Karslruhe Institute of Technology)

Presenter: BERBERI, Lisana (KIT-G)

Session Classification: Demonstrations & Posters

Contribution ID: **27** Type: **Long Talk**

The interTwin Digital Twin Engine: a platform for building and managing scientific Digital Twins

Tuesday, 1 October 2024 15:15 (15 minutes)

The Horizon Europe interTwin project is developing a highly generic yet powerful Digital Twin Engine (DTE) to support interdisciplinary Digital Twins (DT). Comprising thirty-one high-profile scientific partner institutions, the project brings together infrastructure providers, technology providers, and DT use cases from Climate Research and Environmental Monitoring, High Energy and AstroParticle Physics, and Radio Astronomy. This group of experts enables the co-design of the DTE Blueprint Architecture and the prototype platform benefiting end users like scientists and policymakers but also DT developers. It achieves this by significantly simplifying the process of creating and managing complex Digital Twins workflows.

As part of our contribution, we'll share the latest updates on our project, including the DTE Blueprint Architecture, whose latest version will be under finalisation in Q4/2024. The interTwin components, thanks to the collaboration with ECMWF partner in the project, are designed to be aligned with what Destination Earth is designing and building. Therefore, we will show the activities carried out by the project to analyse DestinE architecture and the points of interoperability planned.

The contribution will also cover the status of the DT use cases we currently support and describe the software releases of the DTE.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: MANZI, Andrea (EGI.eu)

Presenter: MANZI, Andrea (EGI.eu)

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **28** Type: **Demonstrations & Tutorials**

Running Multi-Cloud Workloads on Distributed Datasets with Onedata

Tuesday, 1 October 2024 18:00 (30 minutes)

Onedata continues to evolve with subsequent releases within the 21.02 line, enhancing its capabilities and solidifying its position as a versatile distributed data management system. Key improvements include the rapid development of the automation workflow engine, the maturation of the S3 interface, and powerful enhancements to the web UI for a smoother user experience and greater control over the distributed data.

Apart from that, a significant focus has been put on enhancing the interoperability of the platform. Onedata can be easily integrated as a back-end storage solution for various scientific tools, data processing and analysis platforms, and domain-specific solutions, providing a unified logical view on otherwise highly distributed datasets. This is achieved thanks to the S3, POSIX, and Pythonic data interfaces and tools that enable effortless inclusion of Onedata as a 3rd party solution in CI/CD pipelines. For example, the "demo mode" makes it straightforward to develop and test arbitrary middleware against a fully functional, zero-configuration Onedata backend. With the ability to integrate with SSO and IAM services and reflect the fine-grained federated VO structures, Onedata can serve as a comprehensive data management solution in federated, multi-cloud, and cross-organizational environments. Currently, it's serving this purpose in the ongoing EuroScienceGateway, EUreka3D, and Dome EU-funded projects.

Automation workflows in Onedata can streamline data processing, transformation, and management tasks by automating repetitive actions and running user-defined logic fitted to their requirements. The integrated automation engine runs containerized jobs on a scalable cluster next to the data provider's storage systems. This allows seamless integration of data management and processing steps, allowing for efficient handling of large-scale datasets across distributed environments.

During our demonstration, we will present a comprehensive use case demonstrating Onedata's capabilities in managing and processing distributed data based on the EGI DataHub environment. It will showcase a pipeline that embraces the user's federated identity and VO entitlements, automated data processing workflows, the wide range of Onedata's tools for data management, and interoperability with scientific tools and middleware — with a special focus on the S3 interface.

Join us for the demo to see how Onedata empowers organizations to manage and process federated and multi-cloud data efficiently, driving collaboration and accelerating scientific discovery.

Topic

Data innovations: Data Management/Integration/Exchange

Primary authors: DUTKA, Lukasz (CYFRONET); OPIOLA, Lukasz (CYFRONET)

Co-authors: KRYZA, Bartosz (CYFRONET); ORZECHOWSKI, Michal (CYFRONET)

Presenter: OPIOLA, Lukasz (CYFRONET)

Session Classification: Demonstrations & Posters

Contribution ID: **29** Type: **Long Talk**

The Evolution of INFN's Cloud Platform: improvements in Orchestration and User Experience

Tuesday, 1 October 2024 15:40 (15 minutes)

Over the past years, the Italian National Institute for Nuclear Physics (INFN) has developed and refined its cloud platform, designed to facilitate access to distributed computing and storage resources for scientific research. This evolution in Platform-as-a-Service (PaaS) orchestration has focused on enabling seamless service deployment, improving user experience, and integrating innovative solutions to address changing demands and technological challenges.

INFN's journey toward a robust cloud platform began with the deployment of a national cloud system designed to streamline access to distributed resources. A key element of this initiative was a user-friendly web portal, the INFN Cloud Dashboard, allowing users to instantiate high-level services on-demand. This was achieved through TOSCA templates processed by an orchestration system that supported a lightweight federation of cloud sites and automated scheduling for optimal resource allocation.

The orchestration system used by INFN Cloud is based on the open-source INDIGO PaaS middleware, designed to federate heterogeneous computing environments. It plays a crucial role in orchestrating virtual infrastructure deployment, enabling high-level services like Jupyter Hub, Kubernetes, and Spark clusters. The core component, the Orchestrator, is supported by microservices, facilitating the optimal selection of cloud providers based on specific deployment requirements.

In the context of the internal INFN DataCloud project and some European projects like interTwin and AI4EOSC, INFN is undertaking a comprehensive revamp of its PaaS system to accommodate the changing technology landscape and replace old and legacy software components. A key example of this effort is the transition from the legacy Configuration Management Database (CMDB) to the Federation-Registry, a modern solution built on the FastAPI framework and using neo4j, a flexible graph database. This transition will ensure more robust and scalable management of federation-related information, supporting a diverse set of cloud providers and modern security protocols.

To further optimize the orchestration system, INFN is exploring the use of artificial intelligence to improve deployment scheduling. The Cloud Provider Ranker, which provides the list of providers based on various metrics and Service Level Agreements (SLAs), is going to be enhanced with AI techniques. This improvement will allow for the identification of meaningful metrics, creation of predictive models for deployment success/failure, and regression models for deployment times. These models will enable a more dynamic and accurate ranking of cloud providers, leading to more efficient resource usage and a reduction in deployment failures.

Finally, the PaaS dashboard, which serves as a gateway for user interaction with the orchestration and service deployment system, recently underwent a major renovation to improve usability and security. The dashboard redesign aimed to offer a more secure, efficient, and user-friendly interface while providing a visually appealing design.

This contribution will outline the key advancements in the PaaS orchestration system aimed at supporting scientific communities with a reliable, scalable, and user-friendly environment for their computational needs.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: ANTONACCI, Marica (INFN)

Co-authors: COSTANTINI, Alessandro (INFN); DONVITO, Giacinto (INFN); GRANDI, Claudio (INFN); GIOMMI, Luca; MARTELLI, Barbara; SAVARESE, Giovanni; SERRA, Ettore; SPIGA, Daniele; VIANELLO, Enrico (INFN)

Presenter: ANTONACCI, Marica (INFN)

Session Classification: Cloud Compute federation and national initiatives

Contribution ID: **30** Type: **Short Talk**

Recent developments in Account Linking - ALISE

Thursday, 3 October 2024 10:20 (20 minutes)

Account linking may be usefule at different places in the AAI Architecture. Over the past years, we have seen account linking at the Community-AAI, where multiple Home-Organisation logins may be used to log in to a single account at the Community. This typically allows linking attributes of services such as ORCID or Google. More recently this type of account linking is being integrated in an additional proxy above the Community-AAI. These additional proxies are known as “national Edu-ID”. They aim to support researcher mobility by allowing links to several different, sometimes international, Home Organistions.

To complement these early (or northbound) linkings, we have designed and implemented a system for late (or southbound) linking of accounts. Our use case are users, that authenticate with their federated identity to a modern service inside a particular computer centre. Computer centres are often reluctant to invest early into new AAI systems. Their Unix-based infrastructure (HPC Clusters, Filesystems) therefore do not support federated identities. To allow our modern service to use this infrastructure for federated users, we need to know to which Unix account the federated user will be mapped, when logging in with an account local to the computer centre.

ALISE, the Account LIkning SErvice, does exactly that. The web interface asks the user to log in with the computer centre account. Once authenticated, federated identities my be linked to the computer centre account. The linkage information may be accessed via a REST API, so that our modern service may use this information.

The initial setup is working for the VEGA HPC Centre in Slovenia, where an instance of dCache needs to utilise local storage to read or write data, that a VEGA HPC user owns.

Topic

Trust and Security: Access control

Primary author: Dr HARDT, Marcus (KIT-G)

Co-authors: GUDU, Diana (KIT); ZACHMANN, Gabriel (Karlsruhe Institute of Technology); MILLAR, Paul (DESY)

Presenter: Dr HARDT, Marcus (KIT-G)

Session Classification: Trust & Security

Contribution ID: **31** Type: **Demonstrations & Tutorials**

Come to Play: Federated, Interoperable AI-Cubes at Your Fingertips

Wednesday, 2 October 2024 12:30 (30 minutes)

Datacubes form an acknowledged cornerstone for analysis-ready data â€“the multi-dimensional paradigm is natural for humans and easier to handle than zillions of scenes, for both humans and programs. Today, datacubes are common in many places â€“powerful management and analytics tools exist, with both datacube servers and clients ranging from simple mapping over virtual globes and Web GIS to high-end analytics through python and R. This ecosystem is backed by established standards in OGC, ISO, and further bodies.

In the EarthServer federation, institutions from the US, Europe, and Asia contribute spatio-temporal datacubes through OGC compliant services, including the CoperniCUBE datacube ecosystem. Weather and climate data, satellite data timeseries, and further data are provided, altogether multi-Petabyte. A unique feature is the location transparency: users see the federation offerings as a single, integrated pool. The federation member nodes orchestrate incoming queries automatically, including distributed data fusion.

Further, a tight integration of AI and datacubes is provided through the novel concept of AI-Cubes.

We briefly introduce the concepts of datacubes and then explore hands-on together how to access, extract, analyze, and reformat data from datacubes. Particular emphasis is on federation aspects.

Most of the examples can be recapitulated and modified by participants with online access. Ample room will be left for discussion.

The contributor is editor of the datacube standards in OGC and ISO and member, EOSC.

Topic

Data innovations: Data Spaces

Primary author: Prof. BAUMANN, Peter (rasdaman GmbH)

Presenter: Prof. BAUMANN, Peter (rasdaman GmbH)

Session Classification: Demonstrations & Posters

Contribution ID: **32** Type: **Long Talk**

Management of Open Data Lifecycle with Onedata

Tuesday, 1 October 2024 15:35 (15 minutes)

Onedata1 is a high-performance data management system with a distributed, global infrastructure that enables users to access heterogeneous storage resources worldwide. It supports various use cases ranging from personal data management to data-intensive scientific computations. Onedata has a fully distributed architecture that facilitates the creation of a hybrid cloud infrastructure with private and commercial cloud resources. Users can collaborate, share, and publish data, as well as perform high-performance computations on distributed data using different interfaces: POSIX-compliant native mounts, pyfs (python filesystem) plugins, REST/CDMI API, and S3 protocol (currently in beta).

The latest Onedata release line, 21.02, introduces several new features and improvements that enhance its capabilities in managing distributed datasets throughout their lifecycle. The software allows users to establish a hierarchical structure of datasets, control multi-site replication and distribution using Quality-of-Service rules, and keep track of the dataset size statistics over time. In addition, it also supports the annotation of datasets with metadata, which is crucial for organising and searching for specific data. The platform also includes robust protection mechanisms that prevent data and metadata modification, ensuring the integrity of the dataset in its final stage of preparation. Another key feature of Onedata is its ability to archive datasets for long-term preservation, enabling organisations to retain critical data for future use. This is especially useful in fields such as scientific research, where datasets are often used for extended periods or cited in academic papers. Finally, Onedata supports data-sharing mechanisms aligned with the idea of Open Data, such as the OAI-PMH protocol and the newly introduced Space Marketplace. These features enable users to easily share their datasets with others, either openly or through controlled access.

Currently, Onedata is used in European projects: EUreka3D2, EuroScienceGateway3, DOME[4], and InterTwin[5], where it provides a data transparency layer for managing large, distributed datasets on dynamic hybrid cloud containerised environments.

Acknowledgements: This work is co-financed by the Polish Ministry of Education and Science under the program entitled International Co-financed Projects (projects no. 5398/DIGITAL/2023/2 and 5399/DIGITAL/2023/2).

REFERENCES:
- Onedata project website. https://onedata.org. EUreka3D: European Union's REKonstructed in 3D. https://eureka3d.eu. EuroScienceGateway project: open infrastructure for data-driven research. https://galaxyproject.org/projects/esg/. DOME: A Distributed Open Marketplace for Europe Cloud and Edge Services. https://dome-marketplace.eu. InterTwin: Interdisciplinary Digital Twin Engine for Science. https://intertwin.eu.

Topic

Data innovations: Data Management/Integration/Exchange

Primary authors: KRYZA, Bartosz (CYFRONET); DUTKA, Lukasz (CYFRONET); OPIOLA, Lukasz

(CYFRONET); ORZECHOWSKI, Michal (CYFRONET)

Presenter: ORZECHOWSKI, Michal (CYFRONET)

Session Classification: Inside Data Spaces: Enabling data sharing paradigms

Contribution ID: **33** Type: **Poster**

Analysis of Transitioning from Centralized Federated Learning to Decentralized Federated Learning: A Case Study on Thermal Anomalies Detection using UAV-Based Imaging

Tuesday, 1 October 2024 18:00 (1 hour)

Most machine learning models require a large amount of data for efficient model training. This data is usually expected to be placed in one centralized spot. When enough data is available but not located in one spot, such as data collected by edge devices, sharing data with a central server is necessary. Sharing a large amount of data introduces several issues: data might not be feasible to share because of privacy concerns or data restrictions. In other cases, sharing data is not even possible due to the lack of resources and communication overhead.
Federated Learning (FL) comes into play to solve these problems. It is a machine learning paradigm, which allows distributing a machine learning workflow onto multiple clients. Clients participating within the workflow are able to collaboratively train a machine learning model by training it locally on their own data and just share the updated state of the model after training. The data located on the client itself is not shared with other clients, which leads to a privacy strengthened and more resource saving training process. Based on the architecture, FL can be categorized into centralized and decentralized FL approaches. In the centralized approach, a server for administration and communication purposes is involved. In the decentralized approach, the clients themselves are responsible for the communication, administration and aggregation of a model. Examples of Decentralized FL are Swarm Learning, where in each round an aggregator client is chosen solely for the aggregation of the updated model states of all the clients, or Cyclic Learning, where the results are transferred from client to client in a sequential order, rather than aggregating all together.
A framework for transforming an existing machine learning workflow into a FL workflow is provided open-source with the NVIDIA Federated Learning Application Runtime Environment (NVFlare) library. It can be used for various machine learning workflows and provides plenty of functionality based on the use-case and also offers implementations of Swarm and Cyclic Learning.
We apply FL and NVFlare for the real-world application of UAV-based thermal imaging in urban environments from the third use case of AI4EOSC project. In particular, for detecting thermal anomalies caused by features like cars, manholes and streetlamps. By automatically filtering false alarms, we can improve the efficiency of energy-related systems. Since the data originates from different locations and cities, sharing the data causes data protection and communication costs. Therefore, each client is selected according to the geographical location where the images were taken.
A U-net model is trained to detect thermal anomalies automatically. This workflow is then transformed into a FL workflow using NVFlare. We investigate different centralized FL approaches such as FedAvg, FedOpt, FedProx, Scaffold and Ditto, as well as decentralized FL approaches such as Swarm and Cyclic Learning in terms of scalability, communication, accuracy and performance. Our research demonstrates challenges and opportunities associated with FL and highlights the effectiveness of different approaches in a real-world scenario.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary authors: DUDA, Leonhard (Karlsruhe Institute of Technology - KIT, Computer Science, Karlsruhe, Baden-Württemberg, Germany); Dr ALIBABAEI, Khadijeh (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany); Ms VOLLMER, Elena (Karlsruhe Institute of Technology - KIT, IIP, Karlsruhe, Baden-Württemberg, Germany); Mr KLUG, Leon (Karlsruhe Institute of Technology - KIT, IIP, Karlsruhe, Baden-Württemberg, Germany); Dr BENZ, Mishal (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany); Dr KOZLOV, Valentin (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany); Dr VOLK, Rebekka (Karlsruhe Institute of Technology - KIT, IIP, Karlsruhe, Baden-Württemberg, Germany); Dr GÖTZ, Markus (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany); Prof. SCHULTMANN, Frank (Karlsruhe Institute of Technology - KIT, IIP, Karlsruhe, Baden-Württemberg, Germany); Prof. STREIT, Achim (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany)

Presenter: DUDA, Leonhard (Karlsruhe Institute of Technology - KIT, Computer Science, Karlsruhe, Baden-Württemberg, Germany)

Session Classification: Demonstrations & Posters

Contribution ID: **34** Type: **Long Talk**

Comparative Study of Federated Learning Frameworks NVFlare and Flower for Detecting Thermal Anomalies in Urban Environments

Thursday, 3 October 2024 11:40 (20 minutes)

With the expansion of applications and services based on machine learning (ML), the obligation to ensure data privacy and security has become increasingly important in recent times. Federated Learning (FL) is a privacy-preserving machine learning paradigm introduced to address concerns related to data sharing in centralized model training. In this approach, multiple parties collaborate to jointly train a model without disclosing their individual data.
There are various aggregation algorithms for aggregating the local model updates in Federated Learning, e.g. FedAVG, FedProx, Scafflod, and Ditto to overcome the challenges posed by the fact that data can be unbalanced, non-independent, or non Identically Distributed (non-IID) in FL environments. There exist as well various workflows like Scatter and Gather, Cyclic Learning and Swarm Learning for communication strategies. In addition, various security enhancements including Differential Privacy (DP), Homomorphic Encryption (HE), and secure model aggregation have been developed to address privacy concerns. Key considerations when setting up an FL process involve selecting the best framework that meets the specific requirements of the task in terms of the best aggregation algorithm, workflow, and security enhancements.
To help researchers make informed decisions, within the AI4EOSC project, we provide a comprehensive evaluation and comparison of the two most widely used frameworks for federated learning NVFlare and Flower, which have also recently announced a collaboration between the two.
To compare the frameworks in terms of the features they offer, we develop a deep learning solution for the Detection of thermal anomalies use case of AI4EOSC. We use this real-world case study to demonstrate the practical impact and performance of various FL aggregation algorithms, workflows, and security enhancements, and their implementation in each FL framework.
We highlight the different features and capabilities that these frameworks bring to FL settings to provide a better understanding of their respective strengths and applications. The Flower server was seamlessly integrated into the AI4EOSC dashboard, which simplified our experimentation process. All experiments are monitored and tracked using the MLflow instance provided by the AI4EOSC project. Our evaluation included analyses of the convergence speed of various aggregation methods offered by these frameworks, global model accuracy, communication overhead in various workflows, and privacy-preserving functionalities of both frameworks such as HE and DP. Furthermore, we explore the novel collaboration between these two frameworks to explore synergies and potential improvements in federated learning methods for thermal bridge detection.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary authors: Mr DUDA, Leonhard (Karlsruhe Institute of Technology - KIT, Computer Science, Karlsruhe, Baden-Württemberg, Germany); Dr ALIBABAEI, Khadijeh (Karlsruhe Institute of Tech-

nology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany); Ms VOLLMER, Elena (Karlsruhe Institute of Technology - KIT, IIP, Karlsruhe, Baden-Württemberg, Germany); Mr KLUG, Leon (Karlsruhe Institute of Technology - KIT, IIP, Karlsruhe, Baden-Württemberg, Germany); Dr BENZ, Mishal (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany); Dr KOZLOV, Valentin (Karlsruhe Institute of Technology); Dr VOLK, Rebekka (Karlsruhe Institute of Technology - KIT, IIP, Karlsruhe, Baden-Württemberg, Germany); Dr GÖTZ, Markus (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany); Prof. SCHULTMANN, Frank (Karlsruhe Institute of Technology - KIT, IIP, Karlsruhe, Baden-Württemberg, Germany); Prof. STREIT, Achim (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany)

Presenter: Dr ALIBABAEI, Khadijeh (Karlsruhe Institute of Technology - KIT, SCC, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **35** Type: **Short Talk**

Exploring token authorization enforcements on Grid middleware with Open Policy Agent

Thursday, 3 October 2024 10:40 (20 minutes)

Open Policy Agent (OPA) is an open-source, general-purpose authorization engine that provides a high-level declarative language, called Rego, which allows the expression of policies as code, using a combination of data manipulation and logical operators. OPA takes policy decisions by evaluating the query input against policies and data. The OPA RESTful APIs allow the service to be integrated into any application, making it a versatile tool for authorization and access control.

One of the main advantages of using OPA is its performance optimization capabilities. The OPA policy evaluation engine is designed to handle large volumes of requests, making it an ideal choice for the Grid middleware. Additionally, the OPA caching mechanism allows it to minimize the number of policy evaluations, further improving performance. Moreover, the OPA declarative approach to policy management allows for a more intuitive and straightforward policy development process.

With this contribution, we want to highlight the potential of this framework in the context of our Grid middleware and to illustrate how we are exploring the use of OPA in two use cases: to implement the authorization rules defined in the WLCG JWT profile for StoRM Tape and StoRM WebDAV, and to replace the home-made scope policy engine within INDIGO IAM. The appropriate comparison in terms of performance and compliance between the previous solutions and those based on OPA will also be illustrated.

Topic

Trust and Security: Access control

Primary author: Mrs AGOSTINI, Federica (INFN)

Co-authors: MARCATO, Davide (INFN); VIANELLO, Enrico (INFN); GIACOMINI, Francesco (INFN); GASPARETTO, Jacopo (INFN); BASSI, Luca; MICCOLI, Roberta (INFN); ZOTTI, Stefano Enrico

Presenter: Mrs AGOSTINI, Federica (INFN)

Session Classification: Trust & Security

Contribution ID: **36** Type: **Poster**

Geographic replication of the VOMS Attribute Authority service

Tuesday, 1 October 2024 18:00 (1 hour)

Virtual Organization Membership Service (VOMS) servers have long been used for authentication and authorization based on X509 Proxy Certificates within scientific collaborations. However, the trend is shifting towards token-based Identity and Access Management (IAM) systems. The VOMS Attribute Authority (VOMS-AA) service seamlessly integrates with existing VOMS clients by emulating the traditional VOMS server interface while retrieving authorization data from an IAM database. This approach guarantees continued support for Virtual Organizations still reliant on VOMS after the scheduled decommissioning of the legacy VOMS and VOMS-Admin servers.

To match the reliability, fault tolerance, and ability to handle heavy workloads of the previous servers, VOMS-AA needs a geographically distributed deployment option. This contribution explores strategies for implementing VOMS-AA with geographic replication, ensuring an uninterrupted and scalable service, and presents the results of preliminary tests.

Topic

Trust and Security: Interoperability

Primary authors: MARCATO, Davide (INFN); VIANELLO, Enrico (INFN); Mrs AGOSTINI, Federica (INFN); GIACOMINI, Francesco (INFN); GASPARETTO, Jacopo (INFN); MICCOLI, Roberta (INFN CNAF); ZOTTI, Stefano Enrico

Presenter: MARCATO, Davide (INFN)

Session Classification: Demonstrations & Posters

Contribution ID: **37** Type: **Long Talk**

From HPC to EOSC to DestinE: Leveraging Pangeo in the Global Fish Tracking System for Impactful Marine Conservation

Tuesday, 1 October 2024 15:35 (20 minutes)

The Global Fish Tracking System (GFTS) is a use case from the European Space Agency's DestinE Platform. It leverages the Pangeo software stack to enhance our understanding of fish habitats, and in particular Seabass and Pollack. By addressing a data gap highlighted by the International Council for the Exploration of the Sea (ICES), the project combines various data sources, including data from DestinE Climate Adaptation Digital Twin, data from Copernicus marine services, and biologging data from sea bass tracking.
The 'Pangeo-fish' software, a key part of GFTS, improves data access and usage efficiency. Initially developed for HPC, it was ported for cloud infrastructure because of the versatility of the Pangeo ecosystem. This system's model and approach can be adapted for wider marine ecosystem conservation efforts across different scales, species and regions.
The GFTS system was also tested on Pangeo@EOSC. This Pangeo platform, deployed in collaboration with the EGI-ACE and C-SCALE projects, offers Pangeo notebooks with a Dask gateway for comprehensive data analysis at scale. An equivalent system was implemented on the OVH cloud, to prepare for future porting on the DestinE Platform.
Reflecting its original Pangeo ecosystem, GFTS follows open science guidelines. It includes a Decision Support Tool (DST), which enables users to understand complex results and make informed decisions. Accessibility, usability, and data sharing compliance with FAIR principles are prioritised.
In conclusion, GFTS represents a perfect blend of careful management of data and computational resources, a strong commitment to improving ocean conservation, and their habitats, and the efficient use of advanced technology for data analysis and modelling. The presentation will delve into the project's achievements and challenges, providing valuable insights into the practical benefits of incorporating Open Science practices for marine ecosystem preservation.

Topic

Environmental informatics: Climate Change/Environment

Primary authors: Dr FOUILLOUX, Anne (Simula Research Laboratory); Dr ODAKA, Tina (LOPS); WIESMANN, Daniel (Development Seed); AUTRET, Emmanuelle (LOPS); WOILLEZ, Mathieu (DECOD, IFREMER); Dr RAGAN-KELLEY, Benjamin (Simula Research Laboratory)

Presenters: Dr FOUILLOUX, Anne (Simula Research Laboratory); Dr ODAKA, Tina (LOPS)

Session Classification: Unlocking the Potential of Environmental Data

Contribution ID: **38** Type: **Demonstrations & Tutorials**

Pangeo@EOSC: A EOSC service to enable big data geoscience scientific discoveries and collaborations

Thursday, 3 October 2024 12:30 (30 minutes)

The Pangeo community is eager to demonstrate the Pangeo@EOSC service, derived from a collaboration between Pangeo and the EGI-ACE and C-SCALE projects. Offering Pangeo notebooks as well as Machine Learning (both Pytorch and TensorFlow) and Data Science notebooks (R & Julia), Pangeo@EOSC provides an integrative platform within the EOSC for scientific data analysis. Our demonstration will effectively showcase the functionality, convenience, and far-reaching impact of this service.

Pangeo@EOSC, a powerful, scalable, and open-source platform, enables Big Data analysis across an array of disciplines using vast multi-dimensional data, such as geoscience and environmental science, among others. The platform serves as a bridge between data storage, computation, and the scientist, creating a seamless, integrated working environment that stimulates more efficient research and collaborations.

During our 30-minute demonstration, we will delve into Pangeo@EOSC's functionalities. Starting from data access, we will navigate through data exploration, visualisation, and analysis, and further explore its collaborative features. The demonstration will further illuminate how Pangeo@EOSC facilitates end-to-end reproducibility.

We look forward to engaging with fellow researchers, scientists, and data enthusiasts during the dedicated networking session. This will not only provide valuable insight into practical requirements and evolving expectations in the scientific world, but also offer us a great opportunity to receive feedback on Pangeo@EOSC.

With Pangeo@EOSC, the future of scalable, collaborative, and reproducible scientific research is not just a possibility, but a reality within our reach.

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary authors: FOUILLOUX, Anne (Simula Research Laboratory); ODAKA, Tina (IFREMER)

Presenters: FOUILLOUX, Anne (Simula Research Laboratory); ODAKA, Tina (IFREMER)

Session Classification: Demonstrations & Posters

Contribution ID: **39** Type: **Long Talk**

iMagine: Best practices for suppliers of image collections and analysis tools in aquatic sciences

Thursday, 3 October 2024 09:30 (20 minutes)

The iMagine platform utilizes AI-driven tools to enhance the processing and analysis of imaging data in marine and freshwater research, supporting the study of crucial processes for ocean, sea, coastal, and inland water health. Leveraging the European Open Science Cloud (EOSC), the project provides a framework for developing, training, and deploying AI models. To effectively achieve the objectives of the project, about twelve use cases in different areas of aquatic science are collaborating with the providers of the iMagine AI platform. This collaboration has yielded valuable insights and practical knowledge. Thoroughly revising the existing solutions from data acquisition and preprocessing to the final stage, provides a trained model as a service to the users.

Within the framework of iMagine, we outline various tools, techniques, and methodologies appropriate for aquatic science image processing and analysis. In this work, we delve into the best AI-based solutions for image processing, drawing on the extensive experience and knowledge we have gained over the course of the iMagine project. Clear guidelines for annotating images, coupled with comprehensive training and accessible tools, ensure consistency and accuracy in labeling. Thus, We verify annotation tools such as BIIGLE, Roboflow, LabelStudio, CVAT, and LabelBox based on the different features, along with real-time video streaming tools.

Preprocessing techniques and quality control measures are discussed to enhance data quality in aquatic datasets, aiming to identify and address issues such as blurriness, glare, or artifacts. Preparation of training datasets and their publishing in a data repository with the relevant metadata is assessed. Following this, an overview of deep learning models, including convolutional neural networks, and their applications in classification, object detection, localization, and segmentation methods is provided.

Performance metrics and evaluation methods, along with experiment tracking tools such as Tensorboard, MLflow, Weight and Biases, and Data Version Control are discussed for the purpose of reproducibility and transparency. Ground truth data is utilized to validate and calibrate image analysis algorithms, ensuring accuracy and reliability. Furthermore, AI model drift tools, data biases, and fairness considerations in aquatic science models are discussed, concluding with case studies, discussions on challenges and limitations in AI applications for aquatic sciences.

By embracing these best practices, providers of image collections and image analysis applications in aquatic sciences can enhance data quality, promote reproducibility, and facilitate scientific progress in this field. A collaboration of research infrastructures and IT experts within the iMagine framework results in the development of best practices for delivering image processing services. The project establishes common solutions in data management, quality control, performance, integration, and FAIRness across research infrastructures, thereby promoting harmonization and providing input for best practice guidelines.

Finally, iMagine shares its developments with other leading projects such as AI4EOSC and BlueCloud 2026 to achieve optimal synergy and wider uptake of the iMagine platform and best practices by the larger aquatic and AI research communities.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary authors: Dr AZMI, Elnaz (Karlsruhe Institute of Technology (KIT)); Dr ALIBABAEI, Khadijeh (Karlsruhe Institute of Technology (KIT)); Dr KOZLOV, Valentin (Karlsruhe Institute of Technology (KIT)); Dr LOPEZ GARCIA, Alvaro (CSIC); Dr SCHAAP, Dick (Mariene Informatie Service MARIS BV); Dr SIPOS, Gergely (EGI.eu)

Presenter: Dr AZMI, Elnaz (Karlsruhe Institute of Technology (KIT))

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **40** Type: **Long Talk**

Interoperability between Digital Twins in the Green Deal initiative

Tuesday, 1 October 2024 15:30 (15 minutes)

In this session we will discuss and report on the progress, how Earth System digital twins and digital twins that are part of the wider Green Deal initiative could operate together in a digital twin platform.
For this purpose we explain in detail the processes, technical implementation, and ontology alignment that needs to be put in place in order to allow for interoperability of digital twin systems stemming from different communities and initiatives. We do not intend to provide a generic interoperability framework but work from the assumption that the most value can be derived from providing specific solutions, driven by use cases, that are generic by design but not designed to be generic.
We are not aiming for integration through aggregation but integration through federation where each system focuses on the integration functions or services that allow interoperability between digital twin system components when required.
The level of integration requirement between digital twins can span a wire range of functions and services, from full integration on the physics level (tightly coupled digital twins) to integration through DT outputs (loosely coupled). In order to capture these requirements we defined a so-called integration continuum where we can map integration requirements between digital twins and digital twin systems.
From these exercises we developed a shared high level architectural view and also a common glossary that can describe the implementation for each participating project.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: GEENEN, Thomas (ECMWF)

Co-authors: MANZI, Andrea (EGI.eu); BIRO, Timea (Trust-It Services)

Presenter: GEENEN, Thomas (ECMWF)

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **41** Type: **Demonstrations & Tutorials**

Secure personalized federated learning within the AI4EOSC platform

Wednesday, 2 October 2024 14:30 (30 minutes)

Federated learning aims to revolutionize the scene when it comes to training artificial intelligence models, in particular deep learning and machine learning with distributed data. Emerging as a privacy preserving technique, it allows to train models without centralizing or sharing data, preserving their integrity and privacy. Moreover, different studies show that in some cases it also offers advantages from the point of view of accuracy and robustness of the developed models, but also regarding savings in energy consumption, computational cost, latency reduction, etc.
In this demonstration, we will showcase how to carry out the implementation of a complete federated learning system in the AI4EOSC platform. Specifically, during the session we will perform the live training of an AI model under a personalized federated learning approach. This federated training will be done with multiple clients using distributed data in different locations (including resources from the platform itself, but also from the EGI Federated Cloud), simulating a real world application, including participation from the audience in the overall training process.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary author: SAINZ-PARDO DIAZ, Judith (CSIC)

Co-author: LOPEZ GARCIA, Alvaro (CSIC)

Presenter: SAINZ-PARDO DIAZ, Judith (CSIC)

Session Classification: Demonstrations & Posters

Contribution ID: **42** Type: **Short Talk**

AI assisted user support

Thursday, 3 October 2024 12:20 (10 minutes)

Users may have difficulties to find the needed information in the documentation for products, when many pages of documentation are available on multiple web pages or in email forums. We have developed and tested an AI based tool, which can help users to find answers to their questions. The Docu-bot uses Retrieval Augmentation Generation solution to generate answers to various questions. It uses github or open gitlab repositories with documentation as a source of information. Zip files with documentation in a plain text or markdown format can also be used for input. Sentence transformer model and Large Language Model generate answers.
Different LLM models can be used. For performance reasons, in most tests we use the model Mistral-7B-Instruct-v0.2, which fits into the memory of the Nvidia T4 GPU. We have also tested a larger model Mixtral-8x7B-Instruct-v0.1, which requires more GPU memory, available for example on Nvidia A100, A40 or H100 GPU cards. Another possibility is to use the API of OpenAI models like gpt-3.5-turbo, but the user has to provide his/her own API access key to cover expenses.

Topic

Topic not listed

Primary authors: CHUDOBA, Jiri (CESNET); Mr CHUDOBA, Michal (Faculty of Mathematics and Physics, Charles University)

Co-author: Dr HEJTMÁNEK, Lukáš (Institute of Computer Science, Masaryk University)

Presenter: CHUDOBA, Jiri (CESNET)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **43** Type: **Long Talk**

EO4EU: An Integrated and Scalable Platform for Accessing and Processing Earth Observation and Earth Modeling data

Tuesday, 1 October 2024 15:55 (15 minutes)

The increase in the volume of Earth Observation (EO) data in the past decade has led to the emergence of cloud-based services in recent years. **Copernicus** data and services have provided several EO and Earth Modeling data to European Citizens. Data acquired from Sentinel satellites is made available to the end users through the Copernicus Data Space Ecosystem, providing free access to a wide range of data and services from the Copernicus Sentinel missions and other land, ocean, and atmosphere EO data. Moreover, there are six Copernicus services providing data for the atmosphere, marine, land, climate change, security, and emergency related services. As these services, which are not directly integrated, require different data access methods, Copernicus Data and Information Access Services (DIAS) are providing centralised access to Copernicus data and information, in addition to cloud infrastructure and processing tools. The Copernicus Data Access Service (C-DAS), builds on DIAS-es existing distribution services, ensuring their continuity, and bringing significant improvements like advanced search functions, virtualisations, APIs etc.

Destination Earth (DestinE) develops a high precision digital model of the Earth (a digital twin) to monitor and simulate natural and human activity, with the first two digital twins focusing on weather-induced and geophysical extremes, and on climate change adaptation. DestinE will deliver enormous new Earth modelling data and access to Copernicus data. Finally, there are several existing European Data Spaces providing data from various domains (agriculture, food security, health, energy, natural resources, environmental monitoring, insurances, tourism, security). This data opens new opportunities for the creation of beyond state-of-the-art solutions which can provide new products and services to the public.

Despite the significant volume and plethora of EO and Earth Modeling data offered, their access has not been yet extended beyond experts and scientists to the wider industry to deliver tangible applications that improve our health and lives and protect the planet. Unfortunately, a small part of the market has that kind of expertise and, as follows, **high value EO information remains unexploited**, it is often fragmented, complex, diverse, difficult to find, retrieve, download and process, while users must have some kind of domain expertise to find, access, understand how to pre-process data, find storage solutions and transform data into useful formats for analytics and Geographic Information Systems (GIS).

The **EO4EU project** is providing an integrated and scalable platform to make the above-mentioned EO data easily findable and accessible, relying on machine learning and advanced user interfaces supported by a highly automated multi-cloud computing platform and a pre-exascale high-performance computing infrastructure. EO4EU introduces an ecosystem for the holistic management of EO data, improving its FAIRness by delivering dynamic data mapping and labelling based on AI, while bridging the gap between domain experts and end users, and while bringing in the foreground technological advances to address the market straightness towards a wider usage of EO data.

In this session, the key innovative features of the EO4EU Platform will be presented, and architectural insights will be provided.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: KARATOSUN, Armagan; PISA, Claudio (ECMWF); Dr ALBUGHDADI, Mohanad (ECMWF); KAPROL, Tolga (ECMWF); Dr BAOUSIS, Vasileios (ECMWF)

Presenters: PISA, Claudio (ECMWF); Dr BAOUSIS, Vasileios (ECMWF)

Session Classification: Unlocking the Potential of Environmental Data

Contribution ID: **44** Type: **Demonstrations & Tutorials**

AI Inference Pipeline Composition with AI4Compose and OSCAR

Wednesday, 2 October 2024 14:00 (30 minutes)

OSCAR is an open-source serverless framework to support the event-driven serverless computing model for data-processing applications. It can connect to an object storage solution where users upload files to trigger the execution of parallel invocations to a service responsible for processing each file. It also supports other flexible execution approaches such as programmatic synchronous invocations and exposing user-defined APIs for fast AI inference.

Serverless computing is very appropriate for the inference phase of the AI model lifecycle, as it offers several advantages such as automatic scalability and resource optimization, both at the level of costs and energy consumption. This model, in combination with the composition of workflows using visual environments, can significantly benefit AI scientists. With this objective, we have designed, in the context of the AI4EOSC project, AI4Compose, a framework responsible for supporting composite AI by allowing the workflow composition of multiple inference requests to different AI models. This solution relies on Node-RED and Elyra, two widely adopted open-source tools for graphical pipeline composition, employing a user-friendly drag-and-drop approach. Node-RED, in combination with Flowfuse to support multitenancy, serves as a powerful graphical tool for rapid communication between different services; meanwhile, Elyra provides a visual Notebook Pipeline editor extension for JupyterLab Notebooks to build notebook-based AI pipelines, simplifying the conversion of multiple notebooks into batch jobs or workflows. The integration with OSCAR is made through flow and node implementations offered as reusable components inside both Node-RED and Elyra visual pipeline compositors.

During the session, we want to demonstrate how AI4Compose works, for both Node-RED and Elyra environments, making use of the Flowfuse instance of AI4EOSC and the EGI Notebooks service, empowered by the Elyra extension. We will present how to trigger the inference of AI models available in the AI4EOSC marketplace and compose the workflow graphically, demonstrating that, with AI4Compose, AI scientists can easily design, deploy, and manage workflows using an intuitive visual environment. This reduces the time and effort required for pipeline composition, while the AI model inference can be executed on remote OSCAR clusters running in the EGI Cloud.

This work was supported by the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud"that has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant 101058593. Also, Project PDC2021-120844-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR and Grant PID2020-113126RB-I00 funded by MCIN/AEI/10.13039/501100011033.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary authors: Dr CALATRAVA, Amanda (Universitat Politècnica de València); AGUIRRE, Diego Alejandro (Universitat Politècnica de València); Mr RODRÍGUEZ BENITEZ, Vicente (Universitat Politècnica de València); ALARCON MARIN, Caterina (Universitat Politècnica de València); CABALLER, Miguel (Universitat Politècnica de València); MOLTO, German (Universitat Politècnica de València)

Presenter: Dr CALATRAVA, Amanda (Universitat Politècnica de València)

Session Classification: Demonstrations & Posters

Contribution ID: **45** Type: **Poster**

Assessment of Physical Vulnerability of Dakar, Senegal to Coastal Erosion

Tuesday, 1 October 2024 18:00 (1 hour)

Significant obstacles to the wellbeing of coastal populations arise from the fast growth of megacities along coastal regions, which is driven by urbanisation and made worse by the effects of climate change. Coastal erosion poses a particular threat to the Dakar region in West Africa, given its vast 133-kilometer coastline. The purpose of this study is to measure the level of physical vulnerability to coastal erosion in the Dakar region using the Coastal Vulnerability Index (CVI), which is strengthened with the proximity of settlements to the sea and the presence of protective infrastructure. Using a combination of technologies such as ArcGIS, DSAS, Google Earth Pro, and GPS Visualizer, significant variations in vulnerability were identified across the region. Specifically, the northern and southern coasts are more vulnerable, with average CVIs of roughly 94 and 23, respectively, whereas the western coast has a lower average CVI of around 10, indicating considerably less vulnerability. These findings highlight the importance of taking into account a variety of criteria when assessing coastal vulnerability, as well as providing significant insights for personalised interventions. This research contributes to our understanding of the Dakar region's vulnerability, informing discussions on coastal resilience and adaptation planning in the face of ongoing global environmental changes, thereby increasing our ability to mitigate the negative impacts on coastal populations and infrastructure.

Topic

Environmental informatics: Climate Change/Environment

Primary author: Dr POUYE, Ibrahima (West African Science Service Center on Climate Change and Adapted Land Use (WASCAL))

Co-author: Prof. NDIONE, Jacques André (ECOWAS)

Presenter: Dr POUYE, Ibrahima (West African Science Service Center on Climate Change and Adapted Land Use (WASCAL))

Session Classification: Demonstrations & Posters

Contribution ID: **46** Type: **Short Talk**

yProv: a Cloud-enabled Service for Multi-level Provenance Management And Exploration in Climate Workflows

Thursday, 3 October 2024 09:40 (20 minutes)

Open Science plays an important role to fully support the whole research process, which also includes addressing provenance and reproducibility of scientific experiments. Indeed, handling provenance at different levels of granularity and during the entire analytics workflow lifecycle is key for managing lineage information related to large-scale experiments in a flexible way as well as enabling reproducibility scenarios, which in turn foster re-usability, one of the FAIR guiding data principles.

This contribution focuses on a multi-level approach applied to climate analytics experiments as a way to manage provenance information in a more structured and multi-faced fashion, thus allowing scientists to explore the provenance space across multiple dimensions and get coarse- or fine-grained information according to their needs. More specifically, the talk introduces the yProv multi-level provenance service, a new core component within an Open Science-enabled research data lifecycle, along with its design, main features and graph-based data model.

The service can be deployed on several platforms, including cloud infrastructures: indeed, thanks to the recent integration in the Infrastructure Manager Dashboard (https://im.egi.eu/im-dashboard), non advanced users can easily launch the deployment of a yProv service instance on top of a wide range of cloud providers.

This work is partially funded by the EU InterTwin project (Grant Agreement 101058386), the EU Climateurope2 project (Grant Agreement 101056933) and partially under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 1031 of 17/06/2022 of Italian Ministry for University and Research funded by the European Union – NextGenerationEU (proj. nr. CN_00000013).

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary author: ANTONIO, Fabrizio (CMCC Foundation)

Co-authors: Mr RAMPAZZO, Mattia; Mr CLOCCHIATTI, Jacopo; Mr TABARELLI DE FATIS, Gabriele; Mrs SACCO, Ludovica; Prof. FIORE, Sandro (University of Trento, Trento, Italy)

Presenter: ANTONIO, Fabrizio (CMCC Foundation)

Session Classification: Reproducible Open Science: making research reliable, transparent and credible

Contribution ID: **47** Type: **Long Talk**

AI4EOSC as a toolbox to develop and serve AI models in the EOSC

Thursday, 3 October 2024 11:00 (20 minutes)

Researchers exploiting artificial intelligence (AI) techniques like machine learning and deep learning require access to specialized computing and storage resources. Addressing this need, the AI4EOSC project is providing an easy to use suite of services and tools within the European Open Science Cloud (EOSC). This platform aims to facilitate the development of AI models, including federated learning, zero touch deployment of models, MLOps tools and composite AI pipelines among others.

In this presentation, we will provide an exploration of our platform's high-level architecture, with a particular emphasis on meeting the diverse needs of users. We will give an overview of the frameworks and technologies that lay the foundations of our implementation. Through real-world examples coming from active projects and communities (including the notable involvement of iMagine) we will illustrate how researchers are effectively leveraging the platform to advance their AI initiatives. This showcase serves not only to highlight the capabilities of the AI4EOSC project but also to underscore its practical utility and impact within the scientific community.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary author: LOPEZ GARCIA, Alvaro (CSIC)

Co-authors: COSTANTINI, Alessandro (INFN); Dr CALATRAVA, Amanda (Universitat Politècnica de València); MOLTO, German (Universitat Politècnica de València); HEREDIA CACHA, Ignacio (IFCA); SAINZ-PARDO DIAZ, Judith (CSIC); BERBERI, Lisana (KIT-G); PLOCIENNIK, Marcin (ICBP); Dr KOZLOV, Valentin (Karlsruhe Institute of Technology); TRAN, Viet (IISAS)

Presenter: LOPEZ GARCIA, Alvaro (CSIC)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **49** Type: **Long Talk**

Machine Learning Operations (MLOps): from global landscape to practice in AI4EOSC

Thursday, 3 October 2024 11:20 (20 minutes)

Managing and monitoring AI models in production, also known as machine learning operations (MLOps), has become essential in our days, resulting in the need for highly reliable MLOps platforms and frameworks. In the AI4EOSC project in order to provide our customers with the best available ones, we reviewed the field of open-source MLOps and examined the platforms that serve as the backbone of machine learning systems. Recognizing how tracking experiments may improve the process of organising and analysing the results of machine learning experiments as well as team collaboration and knowledge sharing should be noted. From workflow orchestration to drift detection, every aspect of the machine learning lifecycle was reviewed.
Based on that study and in order to aid scientists in their goal to achieve high model standards and implement MLOps practices, we deployed the MLflow platform for the AI4EOSC and iMagine users, are offering a Frouros drift detection python library, are developing a monitoring system for logging drift detection runs. The provided MLflow platform features a central remote tracking server so that every AI experimentation run either on the AI4EOSC platform or any other resources can be individually tracked and shared with other registered users if desired. Frouros library combines classical and more recent algorithms for both concept and data drift detection.
In this contribution, the global MLOps landscape of the continuously growing AI world will be presented together with our practical implementation in the AI4EOSC project and lessons learned from our users.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: BERBERI, Lisana (KIT-G); Dr KOZLOV, Valentin (Karlsruhe Institute of Technology); ESTEBAN SANCHIS, Borja (Karlsruhe Institute of Technology); NGUYEN, Giang (UI SAV); SAINZ--PARDO DIAZ, Judith (CSIC); Dr CALATRAVA, Amanda (Universitat Politècnica de València); MOLTO, German (Universitat Politècnica de València); TRAN, Viet (IISAS); LOPEZ GARCIA, Alvaro (CSIC)

Presenter: Dr KOZLOV, Valentin (Karlsruhe Institute of Technology)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **51** Type: **Demonstrations & Tutorials**

Templates for reproducible science

Thursday, 3 October 2024 12:30 (30 minutes)

Software engineering best practices favour the creation of better quality projects, where similar projects should originate from similar layout, also called software templates. This approach greatly enhances project comprehension and reduces developers'effort in the implementation of high-quality code. As an example, reproducibility and reusability are the key aspects of this software engineering process, the use of packaging tools and containers is a common practice to achieve robustness and portability for long-term software maintenance. However, these tools are not always easy to use and require a certain level of expertise to implement from scratch. Software templates are known to be an excellent way to reduce the complexity load on the use of such tools on the developer's side.

There exist various tools to create such templates and routinely generate projects from them. One such Open Source tool is cookiecutter 1, a cross-platform command-line utility where a new project is replicated according to a set of files and directories that are pre-configured to provide the base structure. These templates, or cookiecutters, can be re-used and freely hosted on software version control platforms e.g. GitHub, where customization is achieved by filling in placeholders in the template files using project-specific values.

In this contribution, we show you how to develop custom modules within the AI4OS dashboard and make use of the best software development practices for this scientific framework. We present a new service that provides a collection of best templates on a marketplace/hub and use them to generate new projects on-fly through a web interface without requiring the installation of the cookiecutter tool on the client side. The platform features a GitHub repository to collect metadata about templates, a python-based backend, and a javascript Web GUI with the authentication via EGI Check-In. From data preprocessing to model evaluation, this session covers the most critical steps in the process of module development, empowering participants to achieve the desired reproducibility and reusability in their projects.

1 https://github.com/cookiecutter/cookiecutter

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary authors: ESTEBAN SANCHIS, Borja (Karlsruhe Institute of Technology); LAURES, Christophe; Dr HARDT, Marcus (KIT-G); Dr KOZLOV, Valentin (Karlsruhe Institute of Technology)

Presenter: ESTEBAN SANCHIS, Borja (Karlsruhe Institute of Technology)

Session Classification: Demonstrations & Posters

Contribution ID: **52** Type: **Short Talk**

A data statistics service for data publication and usage metrics in the climate domain

Tuesday, 1 October 2024 16:10 (20 minutes)

In the climate domain, the Coupled Model Intercomparison Project (CMIP) represents a collaborative framework designed to improve knowledge of climate change with the important goal of collecting output from global coupled models and making them publically available in a standardized format. CMIP has led to the development of the Earth System Grid Federation (ESGF), one of the largest-ever collaborative data efforts in earth system science involving a large set of data providers and modelling centres around the globe.

ESGF manages a huge distributed and decentralized database for accessing multiple petabytes of science data at dozens of federated sites. In this context, providing an in-depth understanding about the data published and exploited across the federation is of paramount importance in order to get useful insights on the long tail of research.

To this end, the ESGF infrastructure includes a specific software component, named ESGF Data Statistics, deployed at the CMCC SuperComputing Center. More specifically, the service takes care of collecting, storing, and analyzing data usage logs (prior filtering out sensitive information) sent by the ESGF data nodes on a daily basis. A set of relevant usage metrics and data archive information are then visualized on an analytics user interface including a rich set of charts, maps and reports, allowing users and system managers to visualize the status of the infrastructure through smart and attractive web gadgets.

Further insights relevant to the research infrastructure managers could come through the application of a data-driven approach applied to download information in order to identify changes in the download patterns and predict possible issues at the infrastructural level.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: NUZZO, Alessandra (CMCC Foundation)

Co-authors: ANTONIO, Fabrizio (CMCC); NASSISI, Paola (CMCC); MIRTO, Maria (CMCC); FIORE, Sandro (University of Trento); ELIA, Donatello (CMCC); ALOISIO, Giovanni (CMCC)

Presenter: NUZZO, Alessandra (CMCC Foundation)

Session Classification: Unlocking the Potential of Environmental Data

Contribution ID: **53** Type: **Short Talk**

ENVRI-Hub-NEXT: an Ambitious Leap in Interdisciplinary Environmental Research Data Access

Wednesday, 2 October 2024 17:10 (10 minutes)

ENVRI-Hub NEXT is a 36-month project designed to address major environmental challenges such as climate change, natural hazards, and ecosystem loss by advancing multidisciplinary research and integration among European research infrastructures (RIs). This project aims to build on the current ENVRI-Hub platform to create a robust framework that brings together different environmental RIs, facilitating their collaboration and contribution to the European Open Science Cloud (EOSC).

The primary goal of ENVRI-Hub NEXT is to support the integration of environmental RIs across four major domains: atmospheric, marine, terrestrial, and Earth observation. This integration is crucial to unlock the potential of RI data for addressing complex research questions, in line with the European Green Deal and Digital Transition. It will also help establish a coherent, sustainable, and world-leading RI cluster.

To achieve this goal, ENVRI-Hub NEXT promotes operational synergies among environmental RIs and e-infrastructures, provides interdisciplinary science-based services, and enhances the integration with EOSC. The consortium leading this project includes key ESFRI Landmarks (ACTRIS, AnaEE ERIC, EPOS ERIC, Euro-Argo ERIC, IAGOS AISBL, ICOS ERIC, LifeWatch ERIC) and RIs from ESFRI Roadmap (eLTER), along with several technology providers and the EGI Foundation to support the project's technical operations and integration with EOSC Core.

The technical objectives of the project include advancing the ENVRI-Catalogue, -Knowledge Base and Federated Search Engine using AI-based dialogue techniques, semantic WEB technologies and metadata standards. These advancements will enable end-users to effectively discover research assets from multiple RIs

ENVRI-Hub NEXT aims to consolidate the conceptual and technical structure of the ENVRI-Hub platform. This consolidation will involve providing data-driven services including an analytical framework and training that facilitate interdisciplinary research, promoting the integrated use of data from different environmental RIs, and expanding the frontiers of multidisciplinary environmental sciences supporting virtual research environments.

The project officially began in February 2024. Its early focus is on summarising the status of the ENVRI-Hub and outlining the plans for ENVRI-Hub NEXT, with an emphasis on building a framework that supports the long-term sustainability and impact of the integrated environmental research infrastructure.

Topic

Environmental informatics: Climate Change/Environment

Primary authors: GUTIERREZ DAVID, Marta (EGI.eu); BUNDKE, Ulrich (JUELICH); LA ROCCA, Giuseppe (EGI.eu); PETZOLD, Andreas (JUELICH)

Co-authors: DRAGO, Federico (EGI.eu); BRUS, Magdalena (EGI Foundation)

Presenter: BUNDKE, Ulrich (JUELICH)

Session Classification: Powering Collaboration: Technical Computing and Data Continuum Requirements

Contribution ID: **54** Type: **Long Talk**

The PITHIA e-Science Centre

Wednesday, 2 October 2024 16:45 (20 minutes)

PITHIA-NRF (Plasmasphere Ionosphere Thermosphere Integrated Research Environment and Access services: a Network of Research Facilities) is a project funded by the European Commission's H2020 programme to build a distributed network of observing facilities, data processing tools and prediction models dedicated to ionosphere, thermosphere and plasmasphere research. One of the core components of PITHIA-NRF is the PITHIA e-Science Centre (PeSC) that supports access to distributed data resources and facilitates the execution of various prediction models on local infrastructures and remote cloud computing resources. As the project nears its completion in 2025, the e-Science Centre has now become a mature and widely utilised tool within the community. The PeSC facilitates the registration of Data Collections, that can either be datasets or prediction models. Registration utilises a rich set of metadata based on the ISO 19156 standard on Observations and Measurements (O&M) and a Space Physics Ontology to define the applicable keywords. While these standards are based on XML, a wizard is also available for resource providers that makes the creation of these XML files easier and more automated. Users can either browse the registered Data Collections, or search for them utilising free-text keywords or the Space Physics ontology. Once found, they can interact with the Data Collection by either navigating to its external site or accessing is through an Application Programming Interface (API) directly from the e-Science Centre. Data Collections can be deployed either at the providers premises or on EGI cloud computing resources. Data storage is facilitated by the EGI DataHub service. Besides Data Collections, the PeSC also supports the execution of workflows. Workflows can be composed of registered Data Collections and executed via APIs. User management in the PeSC is realised by the integration of the EGI Check-in federated identity management system and the Perun authorisation framework. While the PeSC is completely open to end users and anyone can access it without registration, the publication of Data Collections and workflows requires authentication and authorisation. Users belong to Institutions that own the Data Collections and only members of a certain Institution can manage the given resource. Handling of user tickets is managed by the GGUS ticketing system, also provided by EGI and fully integrated with the PeSC. Currently, there are 57 Data Collections and two workflows registered in the PeSC, represented by 790 XML files describing institutions, individuals, projects, platforms, instruments, etc., and made available to the wider PITHIA research community. The presentation and a live demonstration will explain the above functionalities of the e-Science Centre, give examples of PITHIA Data Collections and workflows, and outline the next steps in the development process. The PeSC can be accessed at https://esc.pithia.eu/.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: KISS, Tamas (University of Westminster, London, UK); Mr CHAN YOU FEE, David (University of Westminster); Mr KAGIALIS, Dimitris (University of Westminster); Dr CHEN, Huankai (University of Westminster); Dr BELEHAKI, Anna (National Observatory of Athens); Dr GALKIN, Ivan (Bolearis Global Design); CHEN, Yin (EGI.eu); FARKAS, Levente

Presenter: KISS, Tamas (University of Westminster, London, UK)

Session Classification: Simplifying Data-Driven Science with User-Friendly Platforms and Gateways

Contribution ID: **55** Type: **Long Talk**

EOSC Beyond: enhancing EOSC Core capabilities and piloting EOSC Nodes

Wednesday, 2 October 2024 11:00 (20 minutes)

The ambition of EOSC Beyond is to support the growth of the European Open Science Cloud (EOSC) in terms of integrated providers and active users by providing new EOSC Core technical solutions that allow developers of scientific application environments to easily compose a diverse portfolio of EOSC Resources, offering them as integrated capabilities to researchers.

EOSC Beyond introduces a novel concept of EOSC, establishing a federated network of **pilot Nodes** operating at various levels (national, regional, international, and thematic) to cater to specific scientific missions. Key objectives include accelerating the development of scientific applications, enabling Open Science through dynamic resource deployment, fostering innovation with testing environments, and aligning EOSC architecture with European data spaces.

The project advances EOSC Core through co-design methodologies, collaborating with diverse use cases from national and regional initiatives (e-Infra CZ, NFDI, NI4OS), and thematic research infrastructures (CESSDA, CNB-CSIC, Instruct-ERIC, ENES, LifeWatch, METROFood-RI).

At the heart of the EOSC Beyond project lies the development of new EOSC Core services to further elevate the platform's capabilities: **EOSC Integration Suite**, **EOSC Execution Framework**, **EOSC Core Innovation Sandbox**. EOSC Beyond is also dedicated to enhancing current **EOSC Core services** and framework.

Topic

EOSC Developments and Open Science: EOSC

Primary author: SCARDACI, Diego (EGI.eu)

Co-author: DRAGO, Federico (EGI.eu)

Presenter: SCARDACI, Diego (EGI.eu)

Session Classification: Empowering Open Science: EGI Community's Impact on EOSC

Contribution ID: **56** Type: **Short Talk**

ML4Fires: A Digital Twin Component for Wildfire Danger Analysis via Global Burned Areas Prediction on Climate Projection Data

Thursday, 3 October 2024 10:20 (10 minutes)

In recent years, the escalation of Extreme Weather Events (EWEs), including storms and wildfires, due to Climate Change has become a pressing concern. This exacerbation is characterised by increased intensity, frequency as well as the duration of such events.

Machine Learning (ML) presents a promising avenue for tackling the challenges associated with predicting global wildfire burned areas. It offers sophisticated modelling techniques capable of estimating EWEs in a cost-effective manner. ML-based algorithms not only assist in detection and prediction but also provide robust data-driven tools for scientists,policymakers, and the general public. Yet, the implementation of such solutions requires a comprehensive infrastructure including data acquisition systems, preprocessing modules, computing platforms, and visualisation tools.

A relevant aspect which the InterTwin project - funded by the EU - focuses on is the development of a Digital Twin for EWE analysis. This Digital Twin harnesses artificial neural networks to model the non-linear relationships between various climate, geomorphological and human factors and the occurrence of EWEs, thereby enabling insights from historical data and projections for future events.

In particular, within the interTwin project, our work is emphasising on modelling and predicting global wildfire burned areas, together with tropical cyclones detection and tracking. Our work aims to establish a resilient system for timely prediction and EWE assessment and analysis on projections scenarios.

The Digital Twin on wildfires prediction integrates data and ML models to provide a proactive approach to the fire danger assessment. These efforts underscore the importance of leveraging cutting-edge technologies to address the challenges posed by Climate Change-induced EWEs, ultimately fostering informed actions and resilient communities.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: DONNO, Emanuele (CMCC Foundation)

Co-authors: ELIA, Donatello (CMCC Foundation); ACCARINO, Gabriele (Euro-Mediterranean Center on Climate Change (CMCC Foundation)); DONNO, Davide (CMCC Foundation); Mr IMMORLANO, Francesco (Euro-Mediterranean Center on Climate Change (CMCC Foundation)); ALOISIO, Giovanni (CMCC Foundation)

Presenter: DONNO, Emanuele (CMCC Foundation)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **57** Type: **Short Talk**

GreenDIGIT: Project and Initiative to Lower Environmental Impact of Future Digital Research Infrastructures

Tuesday, 1 October 2024 11:00 (20 minutes)

In order to keep Research Infrastructures (RIs) at the highest level of excellence in science, new technologies and solutions must be developed to steer toward a reduced environmental footprint, as it is the case for all domains of our societies. Lowering the environmental impact of digital services and technologies has to become a priority for both the operation of existing digital services and the design of future digital infrastructures. GreenDIGIT brings together 4 major distributed Digital Infrastructures at different lifecycle stages, EGI, SLICES, SoBigData, EBRAINS, to tackle the challenge of environmental impact reduction with the ambition to provide solutions that are reusable across the whole spectrum of digital services on the ESFRI landscape, and play a role model. GreenDIGIT will capture good practices and existing solutions and will develop new technologies and solutions for all aspects of the digital continuum: from service provisioning to monitoring, job scheduling, resources allocation, architecture, workload and Open Science practices, task execution, storage, and use of green energy. GreenDIGIT will deliver these solutions as building blocks, with a reference architecture and guidelines for RIs to lower their environmental footprint. User-side tools and Virtual Research Environments will also be expanded with energy usage reporting and reproducibility capabilities to motivate users to apply low-energy practices. The new solutions will be validated through reference scientific use cases from diverse disciplines and will be promoted to providers and users to prepare the next generation of Digital RIs with a low environmental footprint.

Topic

Environmental informatics: Green Computing

Primary authors: Dr DEMCHENKO, Yuri (University of Amsterdam); Prof. FDIDA, Serge (Sorbonne Université); Prof. KORAKIS, Thanasis (University of Thessaly); Dr CHOUNOS, Kostas (University of Thessaly); Ms MAGLAVERA, Stavroula (University of Thessaly); SIPOS, Gergely (EGI.eu)

Presenter: Dr DEMCHENKO, Yuri (University of Amsterdam)

Session Classification: Green Computing: towards greener digital services

Contribution ID: **58** Type: **Short Talk**

A Bayesian Optimization workflow for improving oil spill numerical simulations

Thursday, 3 October 2024 09:50 (10 minutes)

The release of oil into marine environments can result in considerable harm to coastal ecosystems and marine life, while also disrupting various human activities. Despite advances in maritime safety, there has been a noticeable uptick in spill occurrences throughout the Mediterranean basin, as documented by the European Maritime Safety Agency's Cleanseanet program. Precisely predicting the movement and transformation of oil slicks is crucial for assessing their impact on coastal and marine regions. Numerical modeling of oil spills plays a pivotal role in understanding their unseen consequences and addressing observational gaps. However, these models often rely on manually selected simulation parameters, which can affect result accuracy. We propose an innovative approach integrating satellite observations, the Medslik-II oil spill model, and Machine Learning techniques to optimize model parameterization, thereby enhancing the accuracy of oil numerical simulations. Utilizing a Bayesian Optimization Framework, the study seeks the optimal configuration within the parameter space for which model simulations best represent actual oil spill observations.

Validation of the proposed approach is performed using a real case of an oil spill in the Baniyas area (Syria) in 2021. Preliminary evaluations of this framework show promising results, suggesting that combining physics-based and data-driven methodologies can lead to more precise risk assessment and planning for oil spill incidents. Furthermore, the resulting workflow represents an integrated solution for optimal and automated selection of model simulation parameters.

The work is being developed within the framework of the EGI coordinated iMagine project, which focuses on a portfolio of "free at the point of use" image datasets, high-performance image analysis tools empowered with Artificial Intelligence (AI), and best practice documents for scientific image analysis.

Topic

Environmental informatics: Climate Change/Environment

Primary authors: DE CARLO, Marco Mariano (CMCC Foundation); ACCARINO, Gabriele (Euro-Mediterranean Center on Climate Change (CMCC Foundation)); RUIZ ATAKE, Igor (CMCC); ELIA, Donatello (CMCC Foundation); COPPINI, Giovanni (CMCC Foundation); ALOISIO, Giovanni (CMCC Foundation)

Presenter: DE CARLO, Marco Mariano (CMCC Foundation)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **59** Type: **Long Talk**

Lessons learnt with ReproVIP

Thursday, 3 October 2024 09:00 (20 minutes)

The ReproVIP project aimed at evaluating and improving the reproducibility of scientific results obtained with the Virtual Imaging Platform (VIP) in the field of medical imaging. ReproVIP focused on a reproducibility level ensuring that the code produces the same result when executed with the same set of inputs and that an investigator is able to reobtain the published results. We investigated reproducibility at three levels: (i) the code itself, and in particular different versions of the same code [Lig2023], (ii) the execution environment, such as the operating system and code dependencies [Vila2024], parallel executions and the use of distributed infrastructures and (iii) the exploration process, from the beginning of the study and until the final published results [Vila2023].

Within this project, we conducted different studies corresponding to these three reproducibility levels. Some of them were conducted on the EGI infrastructure, in production conditions, others on the Grid'5000 research infrastructure. Grid'5000 is a large-scale testbed deployed in France (and member of the SLICES RI) for experiment-driven research in all areas of computer science. It provides access to a large amount of resources highly reconfigurable and controllable, which allowed us to adopt solutions available on EGI, such as CVMFS.

Within ReproVIP, we also enriched the ecosystem around VIP with tools facilitating the assessment of the reproducibility of scientific results: a reproducibility dashboard, a data management platform and a continuous integration tool. The tools are interconnected and linked to VIP, providing researchers with an integrated end-to-end solution to improve the reproducibilityproducibility of scientific results.

The talk will present the studies and tools produced within ReproVIP, highlighting the findings and lessons learnt during the project.

References:
[Lig2023] Morgane Des Ligneris, Axel Bonnet, Yohan Chatelain, et al., "Reproducibility of Tumor Segmentation Outcomes with a Deep Learning Model,"in International Symposium on Biomedical Imaging (ISBI), Cartagena de Indias, Colombia, Apr. 2023
[Vila2023] Gaël Vila, Axel Bonnet, Fabian Chauveau, et al., "Computational Reproducibility in Metabolite Quantification Applied to Short Echo Time in vivo MR Spectroscopy"in International Symposium
on Biomedical Imaging (ISBI), Cartagena de Indias, Colombia, Apr. 2023
[Vila2024] Gaël Vila, Emmanuel Medernach, Inés Gonzalez, et al., "The Impact of Hardware Variability on Applications Packaged with Docker and Guix: a Case Study in Neuroimaging,"Submitted at https://acm-rep.github.io/2024/, Feb. 2024.

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary authors: POP, Sorina (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM--Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294); BLOT, Hippolyte (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294); BON-

NET, Axel (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294); CERVENANSKY, Frédéric (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294); CHATELAIN, Yohan (Department of Computer Science and Software Engineering, Concordia University); CORNIER, Alexandre (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294); FRINDEL, Carole (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, Lyon, France); MEDERNACH, Emmanuel (IPHC, CNRS/IN2P3, Université de Strasbourg); MOUTON, Claire (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294); PANSANEL, Jérôme (IPHC, CNRS/IN2P3, Université de Strasbourg); RATINEY, Hélène (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294); VILA, Gaël (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294); GLATARD, Tristan (Department of Computer Science and Software Engineering, Concordia University)

Presenter: POP, Sorina (Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294)

Session Classification: Reproducible Open Science: making research reliable, transparent and credible

Contribution ID: **60** Type: **Short Talk**

Towards an EOSC compliant research data repository in Hungary with ARP

Wednesday, 2 October 2024 15:50 (10 minutes)

The Hungarian Research Network's (HUN-REN) Data Repository Platform (ARP) is a national repository infrastructure that was opened to the public in March 2024. With ARP, we aim to create a federated research data repository system that supports the data management needs across its institutional network. Implementing ARP is our first step towards establishing an EOSC compliant research infrastructure.

Here we present the conceptualization, development, and deployment of this federated repository infrastructure, focusing on the ARP project's objectives, architecture, and functionalities.

The primary goal of the ARP project is to establish a FAIR focused, sustainable, continuously operational federated data repository infrastructure that not only supports the central storage and management of digital objects but also ensures the interoperability and accessibility of research data across various scientific domains.

The Hungarian Research Network (HUN-REN) currently comprises 11 research centers, 7 research institutes and 116 additional supported research groups, conducting research in the most varied disciplines of mathematics and natural sciences, life sciences, social sciences and the humanities.

ARP is built on a foundation of secure and scalable storage solutions, utilizing the HUN-REN's existing infrastructure to establish a resilient and redundant data storage environment. The system incorporates a hierarchical storage model with a capacity of 1.4 Petabytes, distributed across two sites for enhanced data security. This model supports triple replication of data, ensuring high availability and disaster recovery capabilities.

Central to the ARP's functionality is its suite of data management tools. The primary service of ARP is the data repository itself built on Harvard's Dataverse repository system. In ARP we addressed an important shortcoming of Dataverse, namely the difficulty to handle a diverse set of metadata schemas. As ARP's goal is to support the metadata annotation needs of researchers of various domains it was inevitable to provide a richer set of metadata schemas besides the ones built into Dataverse. To achieve this we added as a central component a Metadata Schema Registry, built on Stanford University's CEDAR framework and closely integrated with the ARP repository to manage diverse data types and standards, ensuring interoperability across different research disciplines.

Beside providing the possibility to author and use any domain specific metadata schema we also extended Dataverse with the import, export and authoring of datasets using RO-Crate via our custom AROMA tool. AROMA and RO-Crate facilitates the structured packaging and rich metadata annotation of research data, enhancing the granularity and usability of data curation. With RO-Crate it is possible to describe datasets or individual files in datasets in any detail that is not otherwise possible in Dataverse.

ARP as a federated service integrates disparate data management systems into a cohesive framework that supports a unified knowledge graph and query service for researchers nationwide. The ViVO based knowledge graph of ARP enables detailed, file-level search functionality and supports federated searches across a variety of national and international research databases, significantly improving data discoverability.

HUN-REN ARP project represents a significant advancement in the field of research data management for the Hungarian research community.

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary author: Mr PATAKI, Balázs (HUN-REN SZTAKI)

Co-author: Mr KOVÁCS, László (HUN-REN SZTAKI)

Presenter: Mr PATAKI, Balázs (HUN-REN SZTAKI)

Session Classification: National Perspectives: EGI Member Countries' Latest Developments and Future Initiatives

Contribution ID: **61** Type: **Short Talk**

EGI Software Vulnerability Group - evolving for the future

Wednesday, 2 October 2024 17:15 (10 minutes)

The EGI Software Vulnerability Group (SVG) has been handling software vulnerabilities since 2010, with the purpose 'To Minimize the risk of security incidents due to software vulnerabilities'.
This is important in order to help sites protect themselves against the most serious vulnerabilities and to give the communities using the services confidence that their credentials and data are secure and that sites patch in a consistent manner.

As the EGI is evolving, the EGI SVG is evolving to cope with the changing landscape. This includes increased inhomogeneity of the infrastructure, and increasing proliferation of services on the infrastructure.
This short talk aims to inform people of what changes we have made in recent times, what our plans are, and invite others to become involved. Whether reported vulnerabilities are deemed to be in scope depends on sufficient participation of people with expertise in the affected areas. We aim for service providers to effectively help one another stay secure via the sharing of their invaluable knowledge!

Topic

Trust and Security: Trusted computing:

Primary author: CORNWALL, Linda (STFC)

Presenter: CORNWALL, Linda (STFC)

Session Classification: Advancing Together: New Features and Roadmaps of the EGI Federation Core

Contribution ID: **62** Type: **Poster**

Enhancing Accuracy in Molecular Dynamics Simulations: Web service for Metal ions Force Field generation

Tuesday, 1 October 2024 18:00 (1 hour)

Molecular Dynamics (MD) simulations provide unique insight into the structural and dynamics of biological macromolecules, contingent upon their accuracy. Two primary determinants of accuracy include the precision of the MD model, particularly the molecular mechanics force field, and the depth of the sampling performed for the simulated system.

The purpose of the conventional force fields is to approximate the energetics of interatomic interactions in MD with relatively simple mathematical formulas. Such force fields are predominantly tailored for biomolecules due to their composition from recurring building blocks, namely amino acids and nucleotides. However, their accuracy diminishes notably for molecules that incorporate metal ions. In particular, metal ions typically form coordination bonds at catalytic sites and are crucial for recapitulating accurately the biological behavior of proteins.

In this contribution, we propose a service that provides a web user interface for the development of force fields for proteins containing metals. Within this service, we have implemented the automatic generation of generalised force field for proteins that contain zinc(II) ions 1. The service takes a pdb file as input and searches for the amino acids that bind these ions. As output, the user receives a zip file containing the input files necessary to initiate a Molecular Dynamics simulation using the Amber or Gromacs suite. Furthermore, we are actively working to implement additional metals as generalized force fields, such as Cu and Fe.

1 A Comparison of Bonded and Nonbonded Zinc(II) Force Fields with NMR Data. Bazayeva M, Giachetti A, Pagliai M, Rosato A. Int J Mol Sci. 2023. doi: 10.3390/ijms24065440.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: GIACHETTI, Andrea (CIRMMP)

Co-author: Prof. ROSATO, Antonio (University of Florence)

Presenter: GIACHETTI, Andrea (CIRMMP)

Session Classification: Demonstrations & Posters

Contribution ID: **63** Type: **Long Talk**

Towards a digital twin for flood risk management

Tuesday, 1 October 2024 16:55 (15 minutes)

Equitable flood risk management is contingent upon understanding the evolution of floods and their impacts on different groups in society. While rapid, open-source, physics-based flood and impact models offer valuable insights, their complexity often limits accessibility for decision-makers lacking technical expertise. Digital twins for flood risk management can address this issue by automating model pre-processing, execution, and post-processing, enabling end users to evaluate meaningful "what-if"scenarios, such as specific events, future conditions, or protective measures, regardless of their technical expertise. These digital twins employ automated workflows and model builders to configure and execute state-of-the-art flood and impact models across various contexts efficiently. However, orchestrating multiple models across disciplines poses challenges, including standardised data management and reproducibility. Our work focuses on developing a digital twin for flood risk management, building on the FloodAdapt desktop application. FloodAdapt integrates compound flood modeling and detailed impact assessment, providing an accessible platform for defining, simulating, and visualizing flood scenarios and their consequences. Users can explore diverse scenarios, including historical events, future projections, and adaptation strategies like green infrastructure, floodwalls, or elevating buildings. In our presentation, we will highlight the capabilities of the flood risk management digital twin that are under development. We'll describe how we leveraged Destination Earth and the interTwin Digital Twin Engine in the implementation of FloodAdapt as a digital twin web application, highlighting the benefits this presents to end-users.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: BACKEBERG, Bjorn (Deltares)

Co-authors: Dr WINTER, Gundula (Deltares); Dr ROSCOE, Kathryn (Deltares); Ms WRIGHT, Sarah (Deltares); Mr VAN 'T WESTENDE, Thierry (Deltares); Dr TROMP, Willem (Deltares)

Presenter: BACKEBERG, Bjorn (Deltares)

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **64** Type: **Short Talk**

VRE as an open scientific gateway to connect EOSC user and resource environments

Wednesday, 2 October 2024 17:25 (10 minutes)

The key task led by EOSC Focus is to coordinate 'Enabling an operational, open and FAIR EOSC ecosystem (INFRAEOSC)'projects under the Horizon Europe (HE) Programme. Technical coordination activities have evolved from annual coordination meetings with the European Commission and online meetings of HE Technology Working Group in EOSC Forum to the recent EOSC Winter School 2024, mainly represented by INFRAEOSC projects granted between 2021 and 2023. Among six Opportunity Areas (OA) identified for collaboration across INFRAEOSC projects, OA4 was dedicated to User & Resource Environments, aligned with the EOSC Strategic Research and Innovation Agenda (SRIA). The participants in OA4 had interactive and hands-on workshops focusing on Virtual Research Environments (VREs), supported by EGI and Galaxy initiatives and EOSC-A Task Force members, who are also the main stakeholders in INFRAEOSC projects highlighting the adoption of EGI resources in their scientific workflows. At the end of intensive Winter School, the roadmap of advancing VREs was discussed in contribution to the new normal to reproduce Open Science following the FAIR principles. In summary, short-term, mid-term and long-term objectives are updated to be contributed among participating projects towards the upcoming EOSC Symposium 2024 and the second edition of Winter School.

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary author: OTSU, Kaori (CREAF)

Co-author: SCARDACI, Diego (EGI.eu)

Presenter: OTSU, Kaori (CREAF)

Session Classification: Simplifying Data-Driven Science with User-Friendly Platforms and Gateways

Contribution ID: **65** Type: **Long Talk**

A digital twin for geophysical extremes: interim results from the DT-GEO project

Tuesday, 1 October 2024 15:45 (15 minutes)

The DT-GEO project (2022-2025), funded under the Horizon Europe topic call INFRA-2021-TECH-01-01, is implementing an interdisciplinary digital twin for modelling and simulating geophysical extremes at the service of research infrastructures and related communities. The digital twin consists of interrelated Digital Twin Components (DTCs) dealing with geohazards from earthquakes to volcanoes to tsunamis and that harness world-class computational (FENIX, EuroHPC) and data (EPOS) Research Infrastructures, operational monitoring networks, and leading-edge research and academic partnerships in various fields of geophysics. The project is merging and assembling latest developments from other European projects and EuroHPC Centers of Excellence to deploy 12 DTCs, intended as self-contained containerised entities embedding flagship simulation codes, artificial intelligence layers, large volumes of (real-time) data streams from and into data-lakes, data assimilation methodologies, and overarching workflows for deployment and execution of single or coupled DTCs in centralised HPC and virtual cloud computing Research Infrastructures (RIs). Each DTC addresses specific scientific questions and circumvents technical challenges related to hazard assessment, early warning, forecasts, urgent computing, or geo-resource prospection. This presentation summarises the results form the two first years of the project including the digital twin architecture and the (meta)data structures enabling (semi-)automatic discovery, contextualisation, and orchestration of software (services) and data assets. This is a preliminary step before verifying the DTCs at 13 Site Demonstrators and starts a long-term community effort towards a twin on Geophysical Extremes integrated in the Destination Earth (DestinE) initiative.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: FOLCH, ARNAU (Geociencias Barcelona (GEO3BCN-CSIC))

Presenter: FOLCH, ARNAU (Geociencias Barcelona (GEO3BCN-CSIC))

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **66** Type: **Short Talk**

Exploring the Potential of Graph Neural Networks to Predict the State of Seagrass Ecosystems in Italian Seas

Thursday, 3 October 2024 12:10 (10 minutes)

Marine and coastal ecosystems (MCEs) play a vital role in human well-being, contributing significantly to Earth's climate regulation and providing ecosystem services like carbon sequestration and coastal protection against sea level rise. However, they face serious threats, including one deriving from the interaction between multiple human stressors (e.g. pollution) and pressures more related to climate change (CC) (e.g. rising sea temperature, ocean acidification, etc.). The complex interplay of these pressures is escalating cumulative impacts on MCEs, jeopardizing their ability to provide ecosystem services and compromising their health and resilience. Machine Learning (ML), using different types of algorithms such as Random Forest (RF) or Support Vector Machine (SVM), can be effective tools to evaluate changes in environmental and ecological status against multiple pressures, but they often overlook the spatial dependence of pressure effects. The examination of spatial relationships among anthropogenic and CC-related pressures is facilitated by Graph Neural Networks (GNNs), which explicitly model the relationships between data points, hence offer potential solutions to the issue of neglecting spatial dependencies in the prediction. Based on these considerations, the main aims of this study are exploring the application of GNNs-based models to evaluate the impact of pressures on Seagrasses ecosystem in the Italian Seas and compare these methods with the models that usually are employed in this field (i.e., RF, SVM, Multi-Layer Perceptron (MLP)).The methodology involves compiling a comprehensive dataset encompassing key variables influencing Seagrass health, including several endogenic and exogenic pressures (e.g., nutrient concentrations, temperature, salinity). Geospatial data from open-source platforms (e.g., Copernicus, EMODnet) are processed and synthesized into a 4km raster grid. The study area was defined based on 2017 seagrass coverage, considering a bathymetry layer up to 50 meters. The seagrasses distribution in each pixel of the case study was considered, categorizing the latter as presence or absence pixels. Experiments include implementing and evaluating different GNN architectures, (i.e., Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs)), alongside traditional ML models. To construct the graph for GNNs, each pixel in the study area, identified by latitude and longitude, is a node. The feature vectors associated with each node represent the pressures. Nodes are connected to their nearest neighboring pixels, forming a spatially informed graph structure. Model performance is assessed using accuracy and F1-score metrics, with GNNs showing the highest F1-score in detecting presence of seagrasses. Qualitative analysis reveals that models lacking spatial context in their predictions tend to exhibit errors attributed to isolated consideration of individual pixels. For instance, these models incorrectly predict the absence of seagrass in regions surrounded by meadows or vice versa. In contrast, GNNs predominantly misclassify pixels along seagrass patch boundaries. While spatial context proves invaluable for prediction accuracy, challenges stemming from limited data availability of high-resolution datasets, impede comprehensive exploration of temporal dynamics within seagrass ecosystem. Future research aims to transition to a local scale, gathering high-resolution data. This facilitates the incorporation of temporal dimensions and the consideration of relevant physical processes, such as ocean currents or extreme events, influencing ecosystem dynamics within the graph.

Topic

Environmental informatics: Climate Change/Environment

Primary author: BIANCONI, Angelica (University School for Advanced Studies IUSS Pavia, Pavia, Italy - Ca' Foscary University of Venice, Department of Environmental Sciences, Informatics and Statistics, Via Torino 155, 30172, Venice, Italy - Centro Euro-Mediterraneo sui Cambiamenti Climatici, Risk Assessment and Adaptation Strategies Division, via Marco Biagi 5, 73100 Lecce, Italy)

Co-authors: Prof. CRITTO, Andrea (Centro Euro-Mediterraneo sui Cambiamenti Climatici, Risk Assessment and Adaptation Strategies Division, via Marco Biagi 5, 73100 Lecce, Italy - Ca' Foscary University of Venice, Department of Environmental Sciences, Informatics and Statistics, Via Torino 155, 30172, Venice, Italy); Dr FURLAN, Elisa (Centro Euro-Mediterraneo sui Cambiamenti Climatici, Risk Assessment and Adaptation Strategies Division, via Marco Biagi 5, 73100 Lecce, Italy - Ca' Foscary University of Venice, Department of Environmental Sciences, Informatics and Statistics, Via Torino 155, 30172, Venice, Italy); Prof. VASCON, Sebastiano (Ca' Foscary University of Venice, Department of Environmental Sciences, Informatics and Statistics, Via Torino 155, 30172, Venice, Italy - European Center for Living Technology, Ca'Foscari University of Venice, Venice, Italy)

Presenter: BIANCONI, Angelica (University School for Advanced Studies IUSS Pavia, Pavia, Italy - Ca' Foscary University of Venice, Department of Environmental Sciences, Informatics and Statistics, Via Torino 155, 30172, Venice, Italy - Centro Euro-Mediterraneo sui Cambiamenti Climatici, Risk Assessment and Adaptation Strategies Division, via Marco Biagi 5, 73100 Lecce, Italy)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **67** Type: **Short Talk**

EGI Helpdesk - Enhancing Support for European Open Science

Wednesday, 2 October 2024 16:55 (10 minutes)

This presentation unveils the enhanced EGI Helpdesk platform, designed to empower researchers and foster collaboration across Europe's open science initiatives. It provides an overview of the migration to the new EGI Helpdesk, detailing the optimization of existing workflows and the implementation of new ticketing processes to ensure a more efficient support experience for EGI users. The presentation will delve into the current status of the migration, highlighting accomplished integrations with other services and outlining the roadmap for a seamless transition. Several demonstrations will showcase the most interesting and complex ticket management workflows for the Worldwide LHC Computing Grid (WLCG) and broader EGI communities.

Topic

EOSC Developments and Open Science: EOSC

Primary authors: HRYNEVICH, Aliaksei (Karlsruhe Insitute of Technology); WEBER, Pavel (KIT-G)

Presenter: WEBER, Pavel (KIT-G)

Session Classification: Advancing Together: New Features and Roadmaps of the EGI Federation Core

Contribution ID: **68** Type: **Short Talk**

BEACON - High performance data access supporting marine data lakes

Wednesday, 2 October 2024 11:15 (15 minutes)

In many of the societal and scientific challenges, such as Digital Twins of the Oceans and virtual research environments, fast access to a large number of multidisciplinary data resources is key. However, achieving performance is a major challenge as original data is in many cases organised in millions of observation files which makes it hard to achieve fast responses. Next to this, data from different domains are stored in a large variety of data infrastructures, each with their own data-access mechanisms, which causes researchers to spend much time on trying to access relevant data. In a perfect world, users should be able to retrieve data in a uniform way from different data infrastructures following their selection criteria, including for example spatial or temporal boundaries, parameter types, depth ranges and other filters. Therefore, as part of the EOSC Future and Blue-Cloud 2026 projects, MARIS developd a software system called 'BEACON'with a unique indexing system that can, on the fly with incredible performance, extract specific data based on the user's request from millions of observational data files containing multiple parameters in diverse units.

The BEACON system has a core written in RUST (low-level coding language) and its indexed data can be accessed via a REST API that is exposed by BEACON itself meaning clients can query data via a simple JSON request. The system is built in a way that it returns one single harmonised file as output, regardless of whether the input contains many different data types or dimensions. It also allows for converting the units of the original data if parameters are measured in different types of units (for this it e.g. makes use of the NERC Vocabulary Server (NVS) and I-Adopt framework).

EOSC-FUTURE Marine Data Viewer
Showcasing the performance and usability of BEACON, the BEACON system is applied to the SeaDataNet CDI database, Euro-ARGO and the ERA5 dataset from the Climate Data Store. These are also connected to a Marine Data Viewer that was developed as part of the EOSC-FUTURE project to co-locate Copernicus Marine satellite derived data products for Temperature and Salinity with observed in-situ data, made available through BEACON instances for the Euro-Argo and SeaDataNet marine data services.

The user interface of the Marine Data Viewer (https://eosc-future.maris.nl/) is designed to allow (citizen) scientists to interact with the data collections and retrieve parameter values from observation data. Enabled by the performance of BEACON, the user can filter the data on-the-fly using sliders for date, time and depth. At present, the ocean variables concern temperature, oxygen, nutrients and pH measurements, from Euro-Argo and SeaDataNet. The in-situ values are overlayed at the same time and space with product layers from Copernicus Marine, based upon modelling and satellite data.

Presentation
During the presentation more details will be given about the BEACON software and its performances. Moreover, latest developments will be presented, which includes deploying BEACON instances for several leading marine and ocean data repositories as part of Blue-Cloud 2026 to provide data lakes to the VRE user community and DTO.

Topic

Data innovations: Data Management/Integration/Exchange

Primary author: THIJSSE, Peter (MARIS)

Co-authors: Mr KOOYMAN, Robin (MARIS); KRIJGER, Tjerk (MARIS)

Presenter: KRIJGER, Tjerk (MARIS)

Session Classification: Maximizing Data Efficiency and Availablity: Innovations in Data Management Platforms

Contribution ID: **70** Type: **Poster**

Prospective Geographies

Tuesday, 1 October 2024 18:00 (1 hour)

The landscape is a complex system characterized by multiple layers of interrelationships. Anthropic intervention, defined as deliberate actions to alter natural environments, is inherently tied to understanding the contextual state of the territory. In landscape projects, the soil acts as a fundamental interface, possessing specific spatial and environmental dimensions where interactions between different ecosystems occur.
Recognizing soil design as crucial for large-scale territorial management, this study explores the concept of "Prospective Geographies," which aims to outline preparatory information scenarios for planning and design interventions. The methodology involves reading and digitizing landscapes through the integration of Earth Observation (EO) data, Geographic Information Systems (GIS), Building Information Modelling (BIM), and computational tools. The primary objective is to develop a site-specific multi-criteria assessment model capable of mapping and classifying soil based on its potential for adaptation and change.
The resulting potential transformation scenarios serve several purposes: 1) supporting planning processes while respecting soil characteristics, 2) promoting effective soil management strategies to optimize available resources, and 3) guiding integrated, multidisciplinary landscape design efforts.
In the face of increasingly urgent environmental challenges due to climate change, adopting multidisciplinary approaches to generate Digital Twin models is essential. This ensures effective resource management and spatial planning based on the soil's transformative potential.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary authors: Mrs MAGAGNOLI, Beatrice (University of Ferrara, Sealine Research Centre); ROMIO, Francesco (Università degli Studi di Ferrara); Mr TINTI, Lorenzo (University of Ferrara, Sealine Research Centre)

Co-author: Dr GRANDO, Alberto (University of Ferrara, Sealine Research Centre)

Presenters: Mrs MAGAGNOLI, Beatrice (University of Ferrara, Sealine Research Centre); ROMIO, Francesco (Università degli Studi di Ferrara)

Session Classification: Demonstrations & Posters

Contribution ID: **72** Type: **Demonstrations & Tutorials**

Demonstrator of the European virtual human twin simulation platform

Wednesday, 2 October 2024 14:30 (30 minutes)

Building simulations for the Virtual Human Twin (VHT) is a challenging and complex task. In order to contemplate practical use of the VHT concept, we require an inclusive ecosystem of digital twins in healthcare, a federated cloud-based repository for gathering human digital twin resources such as models, data sets, algorithms and practices, along with a simulation platform to facilitate the transition towards personalised medicine.
These challenges are the focus of the EDITH EU-funded project 1, whose primary goal is to prepare the European roadmap for developing Virtual Human Twins. In the scope of preparing such a roadmap, we validated its key points by building a prototype implementation of the simulation platform. We began by analysing the internal structure and functional requirements of typical applications simulating human physiology, developed by EDITH partners. This formed the basis for a demonstrator of the execution subsystem of the VHT ecosystem: a software architecture that enables execution of computational models. An integrated versioning system enables collaborative editing and tagging of specific model versions that may be later selected to suit the researchers' needs. The platform also provides a straightforward way to display, download and analyse simulation results. The functionality of the demonstrator was successfully validated with a set of typical VHT modules on ACC Cyfronet HPC resources.
The demonstrator utilises standardised solutions to implement the simulation environment, such as Git repositories to store and version the simulation source code, S3 to store patient data, along with simulation outputs, Dataverse/Zenodo integration to utilise published datasets (or to create new ones), along with HPC to run complex and time-consuming workflows. The environment enables development of algorithms, models and simulations that can make personalised medicine easier, and, as a result, increase the effectiveness and timeliness of medical treatment. Our research has resulted in a demonstrator which can run VHT modules on HPC resources, and which may be integrated with model and data repositories. We consider this the first step towards elaborating the whole VHT ecosystem 2.
In the scope of this presentation we will show the main building blocks of the demonstrator, and discuss how they help build the VHTs and enact the corresponding methodology, ensuring that simulations follow the 3R principles (Repeatability, Replicability and Reproducibility). We will also address the obstacles and challenges posed by existing HPC infrastructures that need to be overcome to simplify the integration of platform similar to the demonstrator with large-scale computational resources.

Acknowledgements. We acknowledge the support of the EU, under grants EDITH No. 101083771, Teaming Sano No. 857533, and ACK Cyfronet AGH grant PLG/2023/016723.

References

1. EDITH –European Virtual Digital Twin, EU Project, Digital Europe, https://www.edith-csa.eu/
2. Peter Coveney, Roger Highfield: Virtual You. How Building Your Digital Twin will Revolutionize Medicine and Change Your Life, Princeton University Press, 2023

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: Mr KASZTELNIK, Marek (ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Kraków, Poland)

Co-authors: Mr BUBAK, Marian (ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Kraków, Poland; Sano Centre for Computational Medicine, Czarnowiejska 36, 30-054 Kraków, Poland); Mr NOWAKOWSKI, Piotr (ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Kraków, Poland; Sano Centre for Computational Medicine, Czarnowiejska 36, 30-054 Kraków, Poland); Mr POŁEĆ, Piotr (ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Kraków, Poland)

Presenter: Mr KASZTELNIK, Marek (ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Kraków, Poland)

Session Classification: Demonstrations & Posters

Contribution ID: **73** Type: **Demonstrations & Tutorials**

A complete workflow for finding, sharing and analysing sensitive data: the SSH use case

Tuesday, 1 October 2024 17:05 (15 minutes)

A large portion of datasets in the Social Science and Humanities (SSH) community is sensitive, for instance for privacy or copyright reasons. The Dutch national infrastructures for the social sciences and humanities, ODISSEI and CLARIAH, collaborate with the Dutch NREN SURF in the development of an integrated workflow to find, request and analyse sensitive data.

In the ODISSEI Portal, researchers can find datasets from a wide variety of data providers, through rich metadata. A service called the Data Access Broker enables researchers to submit a data access request that is processed semi-automatically based on the user's credentials and the data provider's access procedure. After approval, the sensitive data set is transferred directly to SANE: an off-the-shelve, data provider-agnostic Trusted Research Environment (TRE). It is a secure analysis environment that leaves the data provider in full control of the sensitive information.

In an interactive session, ODISSEI and SURF will illustrate how they facilitate a complete workflow: from finding, to requesting, and finally analysing sensitive SSH data.

Topic

Data innovations: Data Analytics, Sensitive Data/FAIR Data

Primary author: HESAM, Ahmad

Co-author: VAN DER MEER, Lucas (Erasmus University Rotterdam)

Presenter: VAN DER MEER, Lucas (Erasmus University Rotterdam)

Session Classification: Managing & Processing Sensitive Data

Contribution ID: **74** Type: **Long Talk**

Dynamic DNS for EGI Cloud federation

Tuesday, 1 October 2024 16:10 (20 minutes)

Nowadays, more and more services are dynamically deployed in Cloud environments. Usually, the services hosted on virtual machines in Cloud are accessible only via IP addresses or pre-configured hostnames given by the target Cloud providers, making it difficult to provide them with meaningful domain names. The Dynamic DNS service was developed by Institute of Informatics, Slovak Academy of Sciences (IISAS) to alleviate this problem.
The Dynamic DNS service provides a unified Dynamic DNS support for virtual machines across the EGI Cloud infrastructure. Users can register their chosen hostnames in predefined domains (e.g., my-server.vo.fedcloud.eu) and assign them to the public IPs of their servers.
The Dynamic DNS service significantly simplifies the deployment of services that are dynamically deployed in Cloud infrastructures. It removes the obstacles of changing IP addresses of services in Cloud at every deployment and enables obtaining SSL certificates for the hostnames. Service providers can migrate services from local servers to Cloud or from a Cloud site to another without noticing users from the change.
The service has been in operation since 2018 with more than one hundred active users. It is being upgraded for stability and security. There are several new and ongoing developments that may be interesting for the users of the Dynamic DNS service:

- Support for wildcards hostnames (already available): the wildcards are critical requirements for load balancers and Kubernetes ingresses
- Support for hostname registration via API (planned): the API for hostname registration would enable full automation of Dynamic DNS service
- Support for issuing SSL certificates (planned): this would overcome the quota limitation of LetsEncrypt, especially for large domain

Stay tuned!

Topic

Needs and solutions in scientific computing: Federated operation

Primary authors: ASTALOS, Jan (IISAS); TRAN, Viet (IISAS)

Presenter: TRAN, Viet (IISAS)

Session Classification: Cloud Compute federation and national initiatives

Contribution ID: **75** Type: **Short Talk**

Connecting Open Science Infrastructures: CSTCloud, GOSC Initiative, and Future Envisions

Wednesday, 2 October 2024 11:20 (20 minutes)

The China Science and Technology Cloud (CSTCloud) stands as one of the key national research e-infrastructures in China. Sponsored by the Chinese Academy of Sciences, CSTCloud aims to empower scientists with efficient and integrated cloud solutions across domains and disciplines. Through the integration of big data, cloud computing, and artificial intelligence, CSTCloud delivers robust data and cloud computing services to bolster scientific innovation and socioeconomic development. To break silos across domains and regions, the idea of co-designing and co-developing a Global Open Science Cloud was proposed during the CODATA Beijing 2019 Conference. This report introduces both CSTCloud and the GOSC initiative, focusing on CSTCloud cloud services and recent pilots within the CSTCloud-EGI and CSTCloud-AOSP EA cloud federations. It also highlights regional collaborations with Africa and Southeast Asia under the GOSC umbrella. Additionally, key cloud technologies and applications deployed in the newly established GOSC testbed will be shared, discussing the interoperability issues across diverse research domains and geographical regions. We aim to offer insights into constructing innovative open science infrastructures in the digital age, fostering robust alignment among stakeholders for interconnected and interoperable open science clouds, and bringing open dialogues in connecting research e-infrastructures for future-led Open Science and SDGs.

Topic

Needs and solutions in scientific computing: National and scientific perspectives

Primary author: Mr LI, Jianhui (CNIC, CAS)

Co-authors: Ms ZHANG, Lili (CNIC, CAS); Ms LI, Xueting (CNIC, CAS)

Presenter: Mr LI, Jianhui (CNIC, CAS)

Session Classification: Global perspectives on advancing Open Science with computational infrastructures

Contribution ID: **77** Type: **Short Talk**

The Global Open Science Cloud Initiative

Wednesday, 2 October 2024 11:40 (20 minutes)

The Global Open Science Cloud (GOSC) Initiative aims to connect worldwide research infrastructures and stakeholders to enable innovative scientific discovery in addressing global challenges. Since its inception, GOSC has embraced a diverse array of stakeholders, fostering partnerships with researchers, institutions, organizations, funding agencies, policymakers, governments, industry players, citizen scientists, and international collaborators. Through years of dedicated implementation, our initiative has made fruitful progress in policy and governance, technology validation, and disciplinary demonstrations, driving forward responsible open science practices towards the Sustainable Development Goals proposed by the United Nations. This poster offers a succinct overview of the GOSC initiative, highlighting current progress while acknowledging the challenges and opportunities that lie ahead. We aim to inspire continued collaboration and innovation within and beyond the GOSC community. Together, we can chart a course towards a more interconnected and inclusive future for both science and society.

Topic

Needs and solutions in scientific computing: Federated operation

Primary authors: Mr LI, Jianhui (CNIC, CAS); Ms ZHANG, Lili (CNIC, CAS); Ms LI, Xueting (CNIC, CAS)

Presenters: Ms ZHANG, Lili (CNIC, CAS); Ms LI, Xueting (CNIC, CAS)

Session Classification: Global perspectives on advancing Open Science with computational infrastructures

Contribution ID: **80** Type: **Long Talk**

Toward a compute continuum with interLink

Tuesday, 1 October 2024 17:40 (20 minutes)

he integration of High-Performance Computing (HPC), High-Throughput Computing (HTC), and Cloud computing is a key to enable convergent use of hybrid infrastructures.
We envision a model where multi stage workflows can move back and forth across multiple resource providers by offloading containerized payloads.
From a technical perspective the project aim is to use the Kubernetes API primitives to enable a transparent access to any number of external hardware machines and type of backends.
We created the interLink project, an open source extension to the concept of Virtual-Kubelet with the primary goal to have HPC centers exploitable with native Kubernetes APIs with an effort close to zero from all the stakeholders' standpoint.
interLink is developed by INFN in the context of interTwin, an EU funded project that aims to build a digital-twin platform (Digital Twin Engine) for sciences, and the ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing in Italy. In this talk we will walk through the key features and the early use cases of a Kubernetes-based computing platform capable of extending its computational capabilities over heterogeneous providers: among others, the integration of a world-class supercomputer such as EuroHPC Vega and Juelich will be showcased.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: SPIGA, Daniele; CIANGOTTINI, Diego (INFN)

Co-authors: MANZI, Andrea (EGI.eu); FILIPCIC, Andrej (JSI); SURACE, Giacomo; PRICA, Teo (IZUM); BOCCALI, Tommaso; MEMON, ahmed (JUELICH); Dr TEDESCHI, tommaso (INFN)

Presenter: SPIGA, Daniele

Session Classification: Bridging the Gap: Integrating the HPC Ecosystem

Contribution ID: **81** Type: **Long Talk**

An EGI Research Commons?

Wednesday, 2 October 2024 11:00 (20 minutes)

This presentation explores how the data, storage and compute Solutions and Services provided by EGI might be transformed into an EGI Research Commons.

The publication by the RDA Global Open Research Commons Working Group in October, 2023 of the Global Open Research Commons International Model (GORC Model) made available a well researched and fully featured template for a Research Commons. To borrow the definition by Scott Yockel, University Research Computing Officer at Harvard, a research commons "brings together data with cloud computing infrastructure and commonly used software, services and applications for managing, analyzing and sharing data to create an interoperable resource for a research community".

Since the publication of the GORC model, national organizations in Sweden, The Netherlands, Germany, and elsewhere, and ELIXIR, are using the GORC Model to explore establishing Research Commons.

A fully featured proposal to create a Research Commons for Norway (REASON) based on the GORC Model, was submitted to the Norwegian Infrastructure Fund in November, 2023. REASON is being used as a reference by many groups exploring the establishment of a Research Commons.

As Research Commons are coming into prominence, parallel initiatives, called Research Clouds, are also emerging. Examples include the ARDC Nectar Cloud in Australia, the Alliance Cloud in Canada, and the New England Research Cloud in the northeast US. 'Bringing Data to Compute' is a central objective of Research Clouds that is overlooked in most Research Commons. The New England Research Cloud (NERC) is arguably the most interesting Research Cloud, because it also incorporates as a foundational feature an important element of Research Commons, namely deployment of a series of researcher-facing research research data management tools that interoperate through the research data lifecycle, and are deployed in conjunction with storage and compute resources.

Three key features of both Research Commons and Research Clouds are: first, they offer researchers access to an integrated series of complementary services that are accessible from a single platform; second the researcher-facing data management services are integrated with the cloud and compute layer; the researcher-facing data management services facilitate passage of data and metadata between tools throughout the research lifecycle.

EGI provides most of the storage, cloud and compute services identified in the GORC model, but these do not present as an integrated platform, and it provides only a few, unconnected researcher facing data management services. How might EGI add these elements to present as a Research Cloud/Commons?

The presentation is divided into the following sections:

1. Introduction to the GORC Model
2. Overview of REASON and lessons learned in putting it together
3. REASON's applicability as a reference for other Research Commons
4. Introduction to the NERC with a particular focus on the integration of storage/compute with a series of interoperable researcher- facing tools
5. Comparison of the NERC with the Compute and Data services offered by EGI
6. Exploration of an 'EGI Research Commons' with reference to the NERC, and consideration of issues that would need to be addressed in its design and implementation

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: Mr MACNEIL, Rory (Research Space)

Presenter: Mr MACNEIL, Rory (Research Space)

Session Classification: Global perspectives on advancing Open Science with computational infrastructures

Contribution ID: **82** Type: **Short Talk**

Open Data for DESY and HIFIS

Thursday, 3 October 2024 10:00 (10 minutes)

DESY is one of the largest synchrotron facilities in Europe and as such is involved with a large amount of different scientific fields. Among these are High Energy and Astro particle Physics, Dark matter research, Physics with Photons and Structural Biology which generate huge amounts of data. This data is valuable and mostly handled in accordance with domain and community specific policies which take into account that embargo periods, ownership and license restrictions are respected. Nowadays there is a push towards opening the data up to the public as requested by funding agencies and scientific journals. In order to support this push, DESY IT is implementing and deploying solutions that support and enable the publishing of Open Data sets for the scientific community. These solutions will make the Open Data easily findable, browsable and reusable for further analyses by the long tail of science, especially when it's participants are not supported by large e-infrastructures.

With Open and FAIR data principles in mind, we will provide a metadata catalogue to make the data findable. The accessibility aspect is covered by making use of federated user accounts via eduGAIN, HelmholtzID, NFDI and later EOSC-AAI and will give community members access to the data with their institutional accounts. The interoperability of the data sets is ensured by establishing the use of commonly accepted data formats such as HDF5, specifically NeXuS and openPMD wherever possible. Providing the technical and scientific metadata will finally make the open data sets reusable for subsequent analyses and research. In order to address the spirit of sharing in Open Science, the blueprint for our Open Data solution will be shared with others through HIFIS first and upon successful evaluation also with the wider community.

Our protoype will initially consist of three connected solutions: the metadata catalogue SciCat, the storage system dCache and the VISA (Virtual Infrastructure for Scientific Analysis) portal. Scientific data is placed in a specific directory on dCache together with its metadata which is ingested into SciCat to be availabe for access and download options. Here, it is crucial to ensure that the scientific metadata stored in the catalog is harmonized among similar experiments. In order to achieve this, we are devising a method of creating experiment-specific metadata schemata against which metadata will be validated before ingestion. Simultaneously, a subset of the technical and scientific metadata will be integrated into the VISA portal such that scientists can access the dataset within it. VISA is a portal that allows creating virtual machines with pre-installed analysis tools, the selected data sets already mounted and accessible from a web browser forming a consistent environment allowing easy access to data and tools.

During the talk, we will present the architecture of the system, its individual components as well as their interplay. The focus will be the harmonization of the metadata schemata as well as the roadmap for the development of tooling and processes for ingestion and validation of the ingested metadata.

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary authors: FUHRMANN, Patrick (DESY); MILLAR, Paul (DESY); Dr WETZEL, Tim

Co-authors: Dr REPPIN, Johannes (Deutsches Elektronen-Synchrotron DESY); Dr PITHAN, Linus

(Deutsches Elektronen-Synchrotron DESY); Mr VAN DER REEST, Peter (Deutsches Elektronen-Synchrotron DESY); Dr HINZMANN, Regina (Deutsches Elektronen-Synchrotron DESY); JANDT, Uwe

Presenter: Dr WETZEL, Tim

Session Classification: Reproducible Open Science: making research reliable, transparent and credible

Contribution ID: **83** Type: **Demonstrations & Tutorials**

FAIR EVA (Evaluator, Validator & Advisor) and its Plugin System

Wednesday, 2 October 2024 12:30 (30 minutes)

FAIR EVA is a tool that allows checking the level of adoption of the FAIR principles for digital objects. It provides an API for querying via a persistent identifier and a web interface to interpret the offered results. These results assess, based on a series of indicators and automated technical tests, whether certain requirements are met. Additionally, FAIR EVA not only aims to evaluate and validate digital objects and their level of compliance with the FAIR principles, but it also intends to help data producers improve the characteristics of their published objects through a series of tips.

The diversity of repository systems and data portals means that, technically, the way data and metadata are accessed varies significantly. Although there are interoperability solutions like OAI-PMH or Signposting, certain indicators require a higher level of technical detail, such as those related to metadata standards or formats specific to scientific communities. Moreover, the FAIR principles mainly focus on metadata, and data quality is only superficially assessed.

To address this issue, FAIR EVA is designed modularly and, through its plugin system, can connect with various repositories or data portals with very different technical characteristics. In general, FAIR EVA implements the indicators of the RDA FAIR Maturity Working Group but allows them to be replaced with others or even extended to perform quality tests and metrics for a specific domain. For instance, a plugin has been developed for GBIF (Global Biodiversity Information Facility) that evaluates the adoption level of the FAIR principles and extends the tests to check certain specific quality indices for biodiversity data.

The proposed demo aims to showcase the fundamental features of FAIR EVA, particularly how a plugin can be created and adapted for a specific community, extending the list of tests to assess other aspects of data quality.

FAIR EVA started to be developed under the context of EOSC-Synergy project, and it has released the second version this year. There are different plugins being developed for diverse communities: DT-GEO project for geosciences, AI4EOSC, SIESTA, etc.

Topic

Data innovations: Data Analytics, Sensitive Data/FAIR Data

Primary author: Dr AGUILAR, Fernando (CSIC)

Presenter: Dr AGUILAR, Fernando (CSIC)

Session Classification: Demonstrations & Posters

Contribution ID: **84** Type: **Short Talk**

climdex-kit: an open software for climate index calculation, sharing and analysis towards tailored climate services

Tuesday, 1 October 2024 18:30 (30 minutes)

Climate change and transformation is urging scientific communities and decision makers around the world to better understand and handle such systemic shift and its consequences at different levels and to instill a gradual societal adaptation and change into the population.

The availability of tailored and robust information about current climate and climate change at local, regional or national scales is an increasing requirement in a wide range of end-user applications and as a decision-support basis in the fields of risk reduction and adaptation planning.

Numerous European and national portals have been recently developed to ease access to climate data, visualize precomputed information and promote climate change communication. At the same time, a wide range of ready-to-use packages written in popular programming languages and published in open-source code repositories, e.g., GitHub, have been released with the aim of enabling end users to derive customized data for specific applications, e.g., climate indices for sector-oriented analyses, or to further integrate the utilities into tailored climate services.

However, open-source packages completely integrating all steps composing a service-oriented application –from the calculation of climate information to the open-access publication in repositories, the metadata curation, and customizable analyses –are still missing.

In this framework, with the aim of answering to the increasing need of elaborating climate data for research activities as well as practice-oriented applications we developed an open-source tool called climdex-kit 1 and published in the official Python Package Index (PyPI, https://pypi.org/). The package is designed to support users with some programming skills carrying out research in the field of climate change and impact prediction, to support dissemination and educational activities through effective visualization or to develop more complex architectures for operational platforms addressing a broad audience. The tool is written in Python and integrates utilities from the well-established Climate Data Operators (CDO) and NetCDF Operators (NCO) libraries. climdex-kit provides utilities to implement the whole pipeline of calculation, orchestrate parallelized processing over multiple climate data, publish and analyze climate indices as well as to shape the visualization of results based on user needs. The current version offers the calculation of 37 climate indices, while the package can be easily extended to support other indices and more unforeseen operators, thanks to thorough documentation for developers available in the source repository.

We will present and discuss the climdex-kit functionalities as well as its potential integration into local applications by applying the software to a dataset of climate projections for the Italian region Trentino-South Tyrol, used as study case.

1 https://pypi.org/project/climdex-kit/

Topic

Environmental informatics: Climate Change/Environment

Primary authors: CAMPALANI, Piero (Eurac Research); CRESPI, Alice (Eurac Research)

Co-authors: PITTORE, Massimiliano (EURAC Research); Dr ZEBISCH, Marc (Eurac Research - Center for Climate Change and Transformation)

Presenters: CAMPALANI, Piero (Eurac Research); CRESPI, Alice (Eurac Research)

Session Classification: Demonstrations & Posters

Contribution ID: **85** Type: **Long Talk**

Sunet Drive - An Academic Toolbox for FAIR Data Storage, Analysis and Publication

Wednesday, 2 October 2024 12:15 (15 minutes)

Sunet Drive is a national file storage infrastructure for universities and research institutions in Sweden. It is based on a Nextcloud setup and is comprised of 54 nodes, one prepared and provisioned for each institution. The aim of Sunet Drive is to become an **Academic Toolbox** capable of collecting, storing, analyzing, and publishing research data, supporting FAIR principles. We present Sunet Drive as an integrated solution comprised of four essential building blocks:

- File sync and share based on Nextcloud and S3 as the underlying storage entities
- eduGAIN login using SeamlessAccess and added security through step-up authentication and MFA zones
- Scalable JupyterHub integration for flexible and reproducible data analysis
- Research Data Services for easy publication to public and curated repositories

Participating organizations co-manage their Sunet Drive node as part of a global scale setup, meaning that every node is governed by the operating organization, while being able to collaborate and share data with users within the federation, but also external partners through open cloud mesh protocol (OCM), such as the ScienceMesh. S3-compatible buckets are used as logical storage entities that can be assigned for different purposes: research projects, institutions, laboratories. They are technically independent from the EFSS layer and their life-cycle can be managed beyond the lifetime of the selected EFSS software, an important step towards long-term sustainability of FAIR data.

Collaboration is encouraged by allowing access through eduGAIN and subsequently accept documents, shares, and data from their collaboration partners. External collaboration is enabled via Eduid.se. Added security can be provided through step-up-authentication, adding a second authentication factor for identity providers that have not added support for 2FA yet. Further security can be added by activating MFA-zones, mandating the receiver of a file or folder to add a second authentication factor, such as TOTP or a WebAuthn device.

During the runtime of a research project, data can be processed and analyzed directly through a scalable JupyterHub integration, an open source application developed by Sunet and funded by the GÉANT Project Incubator. Compute resources are intelligently managed in a kubernetes environment and can be allocated on a per-project basis, which includes support for CPU and GPU flavours.

The integration of Research Data Services, RDS, enables the preparation and direct publication of datasets directly. This includes services like InvenioRDM (e.g., Zenodo), Harvard Dataverse, or Doris from the Swedish National Dataservice, SND. Research object crates (RO-Crate) are used as an intermediate lightweight package for the data, and respective metadata, connectors ensure compliance with each publication service. Domain-specific customizations include the integration of different publishing paradigms: While data is being actively pushed to repositories such as InvenioRDM or OSF, the SND Doris connector uses a more lightweight approach where the metadata is pushed to Doris, with the data storage remaining under the sovereignty of the publishing institution.

Providing researchers with an Academic Toolbox with streamlined support for authentication, data management, analysis, and publication helps to ensure compliance with local, national, and international guidelines for storing of research data, including FAIR principles.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: FREITAG, Richard (SUNET)

Co-authors: Mr ANDERSSON, Magnus (SUNET); Mr NORDIN, Micke (SUNET)

Presenter: FREITAG, Richard (SUNET)

Session Classification: Maximizing Data Efficiency and Availablity: Innovations in Data Management Platforms

Contribution ID: **86** Type: **Long Talk**

Bridging Cloud and HPC for Scalable Event-driven Processing of AI Workloads

Thursday, 3 October 2024 10:00 (20 minutes)

Cloud computing has revolutionized how we store, process, and access data, offering flexibility, scalability, and cost-effectiveness. On the other hand, High Performance Computing (HPC) provides unparalleled processing power and speed, making it an essential tool for complex computational tasks. However, leveraging these two powerful technologies together has been a challenge.

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have grown exponentially, with many software tools being developed for the Cloud. Despite this, the potential of integrating these tools with HPC resources has yet to be explored.

Containers have revolutionized application delivery due to their lightweight and versatility. They are standard in cloud-native applications, but new containerization technologies have emerged specifically for HPC environments.

Our team presents a solution for seamlessly integrating Cloud and HPC environments using two essential tools: OSCAR, a Kubernetes-based serverless event-driven platform where the user can easily create services for running jobs within a container, and interLink, a middleware that allows the offloading of tasks created in a Kubernetes cluster to an HPC cluster.

The OSCAR-interLink integration together with iTwinAI, a framework for advanced AI/ML workflows, allows for AI workloads, such as 3DGAN inference, to take advantage of the resources available in HPC, including GPU processing power.

Our results showcase a successful use case, integrating dCache, Apache NiFi, OSCAR, interLink, and iTwinAI, based on a 3DGAN for particle simulation, demonstrating the benefits of the approach by exploiting remote GPUs from an HPC facility from an OSCAR cluster running on a Cloud infrastructure.

This work was supported by the project "An interdisciplinary Digital Twin Engine for science"(interTwin) that has received funding from the European Union's Horizon Europe Programme under Grant 101058386. GM would like to thank Grant PID2020-113126RB-I00 funded by MCIU/AEI/10.13039/501100011033. GM and SL would like to thank project PDC2021-120844-I00 funded by MCIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary authors: PARCERO, Estibaliz (Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València); Mr LANGARITA, Sergio (Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València); CIANGOTTINI, Diego (INFN); BUNINO, Matteo; MOLTO, German (Universitat Politècnica de València); SPIGA, Daniele

Presenter: PARCERO, Estibaliz (Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **87** Type: **Short Talk**

Distributed computing platform on EGI Federated Cloud

Thursday, 3 October 2024 12:00 (10 minutes)

The AI4EOSC project will deliver an enhanced set of services for the development of Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) models and applications for the European Open Science Cloud (EOSC). One of the components of the platform is the workload management system that manages execution of compute requests on different sites on EGI Federated Cloud.

To be able manage the distributed compute resources in a simple and efficient way, a distributed computing platform must be created. We based this platform on the service mesh technology paradigm. The platform consists of three parts:

- The underlying network connection is based on the Hashicorp Consul that enables the managing of secure network connectivity with help of the Envoy proxy across different cloud environments (multi clouds and multi providers) and on premises as well. It offers different services like discovery, service mesh, traffic management, and automated updates to network infrastructure devices.
- To manage the workload on the computing resources we adopt the workload orchestrator Hashicorp Nomad, which enables deployments and managements of containers and non-containerized applications at scale. Nomad can run a diverse workload of Docker, non-containerized, microservices, and batch applications.
- The last but not least part is the AI4EOSC API for managing job execution on Nomad. The API enables advanced authentication/authorization mechanisms (OIDC authentication, VO-based authorization), jobs monitoring and also simplifies job management by attaching additional metadata to jobs.

This platform is a unified, reliable, distributed computing system on different sites on EGI Federated Cloud. It resembles the Kubernetes platform. On the other side the Hashicorp Consul and Nomad are more simpler, lighter and flexible compared to Kubernetes. And it is a completely distributed and fault tolerant platform for reliable job execution.

Topic

Needs and solutions in scientific computing: Federated operation

Primary author: SELENG, Martin (IISAS)

Co-authors: LOPEZ GARCIA, Alvaro (CSIC); HEREDIA CACHA, Ignacio (IFCA); Mr HABALA, Ondrej (IISAS); Mr FERNANDEZ, Saul (IFCA); TRAN, Viet (IISAS)

Presenters: SELENG, Martin (IISAS); TRAN, Viet (IISAS)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **88** Type: **Long Talk**

Empowering Reproducible Open Science through the Cloud Computing Platform (CCP) by D4Science

Thursday, 3 October 2024 09:20 (20 minutes)

Abstract
The Cloud Computing Platform (CCP), developed under the aegis of D4Science 1, an operational digital infrastructure initiated 18 years ago with funding from the European Commission, represents a significant advancement in supporting the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, open science, and reproducible data-intensive science. D4Science has evolved to harness the "as a Service" paradigm, offering web-accessible Virtual Laboratories 2 that have also been instrumental in facilitating science collaborations 3. These laboratories simplify access to datasets whilst concealing underlying complexities, and include functionalities such as a cloud-based workspace for file organisation, a platform for large-scale data analysis, a catalogue for publishing research results, and a communication system rooted in social networking practices.

At the core of the platform for large-scale data analysis, CCP promotes widespread adoption of microservice development patterns, significantly enhancing software interoperability and composability across varied scientific disciplines. CCP introduces several innovative features that streamline the scientific method lifecycle, including a method importer tool, lifecycle tracking, and an executions monitor with real-time output streaming. These features ensure that every step —from creation, through execution, to sharing and updating—is meticulously recorded and readily accessible, thus adhering to open science mandates. CCP supports a broad range of programming languages through automatic code generation, making it effortlessly adaptable to diverse scientific requirements. The robust support for containerisation, utilising Docker, simplifies the deployment of methods on scalable cloud infrastructures. This approach not only reduces the overhead of traditional virtualisation but also enhances the execution efficiency of complex scientific workflows. The platform's RESTful API design further facilitates seamless interactions between disparate software components, promoting a cohesive ecosystem for method execution and data analysis.

Significantly, CCP embodies the principles of Open Science by ensuring that all scientific outputs are transparent, repeatable, and reusable. Methods and their executions are documented and shared within the scientific community, enhancing collaborative research and enabling peers to verify and build upon each other's work. The platform's design also includes comprehensive provenance management, which meticulously tracks the origin and history of data, thus providing a record for scientific discoveries.

CCP serves as a platform for large-scale data analysis of the (i) EOSC Blue-Cloud2026 project VRE, which by leveraging digital technologies for ocean science, utilises CCP to perform large-scale collaborative data analytics, significantly benefiting from CCP's robust, scalable cloud infrastructure and tools designed for extensive data processing and collaboration, and of the (ii) SoBigData Research Infrastructure that, with its focus on social data mining and Big Data analytics, integrates CCP to facilitate an ecosystem for ethical, scientific discoveries across multiple dimensions of social life.

Keywords: Open Science, Cloud Computing, FAIR Principles, Reproducibility, Data-intensive Science, Containerisation, Microservices

Topic

EOSC Developments and Open Science: Reproducible Open Science

Primary authors: Dr ASSANTE, Massimiliano (CNR); Mr DELL'AMICO, Andrea (CNR); Mr LETTERE, Marco (Nubisware S.r.l.); PAGANO, Pasquale (CNR); Mr PANICHI, Giancarlo (CNR)

Presenters: Dr ASSANTE, Massimiliano (CNR); Mr LETTERE, Marco (Nubisware S.r.l.)

Session Classification: Reproducible Open Science: making research reliable, transparent and credible

Contribution ID: **89** Type: **Poster**

Using AI4EOSC platform for integrated plant protection use case

Tuesday, 1 October 2024 18:00 (1 hour)

The AI4EOSC project will deliver an enhanced set of services for the development of Artificial Intelligence (AI) models and applications for the European Open Science Cloud (EOSC). One of the scenarios making use and validating the platform is related to enhancement of the integrated plant protection (agriculture sector).

The experiment aims to enhance capabilities of currently used disease detection methods based on mathematical model calculations, with new possibilities of ML/DL-based models developed and scaled on the AI4EOSC platform.
The use case about plant protection aims to determine the risk of disease in agricultural crops and determine the phases of plant growth and the condition of crops. The developed AI models are going to be integrated into existing national advisory platforms, operated by WODR (Wielkopolska Agricultural Advisory Centre in Poznań) and PSNC.

WODR and PSNC are currently operating a national advisory platform for farmers (eDWIN), which includes a network of meteorological ground stations, the Farm Management System, and ground observations of the occurrence of diseases. The current solutions are based on predictive mathematical models. The goal was to add to the current mathematical prediction ML/DL-based models used for early detection of the plant diseases. The designed tools in its first release enables individual calculations of occurrence probability for common crops and related disease:
Sugar Beet Leaf Spot Disease
Rye brown rust

Mathematical model results are being displayed to the user in the time domain up to the present day. Users as farmers or professional advisors will gain additional opportunities to identify and react to the presence of predicted diseases.
Target users are farmers, public administration, local governments, scientific institutes and institutions responsible for monitoring hazards in agriculture in terms of plant protection.
The number of the eDWIN platform users (where the outputs will be integrated) is exceeding currently 20,000 in Poland (Farmers, Advisors).

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary authors: FOJUD, Adam (WODR); SMOK, Jędrzej (PSNC); PLOCIENNIK, Marcin (PSNC); KRZYZANEK, Mateusz (PSNC); BLASZCZAK, Michal (PSNC)

Presenter: PLOCIENNIK, Marcin (PSNC)

Session Classification: Demonstrations & Posters

Contribution ID: **90** Type: **Short Talk**

EGI Software Provisioning Infrascture

Wednesday, 2 October 2024 17:05 (10 minutes)

This presentation provides an overview of the architecture and implementation of the new artefacts repositories for EGI.

The EGI repositories are developed, maintained and operated by LIP and IFCA/CSIC. The new repositories will host RPMs for (RHEL and compatible distributions), DEBs (for Ubuntu and compatible distributions) and Docker images for container-based services and micro-services.

The presentation will describe the architecture of the new repositories, its several components and capabilities.

Topic

Needs and solutions in scientific computing: Federated operation

Primary author: PINA, Joao (LIP)

Co-authors: DAVID, Mario (LIP); ORVIZ, Pablo (CSIC); BERNARDO, Samuel (LIP)

Presenter: PINA, Joao (LIP)

Session Classification: Advancing Together: New Features and Roadmaps of the EGI Federation Core

Contribution ID: **91** Type: **Poster**

Enhancing Global Sea Level Anomaly Reconstruction Pre-Altimetry Using Tide Gauges and Scattering Covariance Analysis on the Pangeo-EOSC Platform

Tuesday, 1 October 2024 18:00 (1 hour)

Global sea level anomalies (SLA) are crucial for climate monitoring and have traditionally been studied using spatial altimetry for the past three decades. This research introduces a novel method to refine historical sea level reconstructions by integrating Scattering Covariance Analysis (SCA) with traditional tide gauge data, which spans over a century. This innovative approach allows for an enhanced understanding of past SLAs in the absence of altimetry data.

Our methodology utilizes SCA to capture and interpret scale correlations observed during the altimetry period, thereby improving the interpolation of global SLA data from historical coastal tide gauges. We validate our model using altimetry data and CMIP6 climate projections on the Pangeo-EOSC platform. This platform exemplifies the practical implementation of the 'compute continuum', enhancing the scalability and accessibility of computing resources through its cloud-based datasets, parallel processing infrastructures using Dask Kubernetes clusters, and GPU optimization.

The findings from our study, compared against traditional Empirical Orthogonal Function (EOF) decomposition methods, reveal that SCA provides a more nuanced understanding of the spatial and temporal dynamics of SLA. These insights uncover complex interactions and dependencies that were previously unobserved with the EOF approach. This enhanced methodological framework not only improves the accuracy of historical sea level reconstructions but also expands the potential for future climate impact assessments based on long-term sea level records. Utilizing the Pangeo-EOSC platform, our model leverages federated data from multiple sources and cloud technologies, ensuring efficient handling of large-scale datasets and advancing environmental research through integrated scientific computing environments.

Topic

Needs and solutions in scientific computing: Artificial Intelligence

Primary authors: Dr DELOUIS, Jean-Marc (UMR LOPS CNRS-IFREMER-IRD-Univ.Brest-IUEM); Dr ODAKA, Tina (UMR LOPS CNRS-IFREMER-IRD-Univ.Brest-IUEM)

Presenter: Dr DELOUIS, Jean-Marc (UMR LOPS CNRS-IFREMER-IRD-Univ.Brest-IUEM)

Session Classification: Demonstrations & Posters

Contribution ID: **93** Type: **Short Talk**

Detecting pulsar signals in vast real-time data streams with a machine learning / digital twin-based pipeline

Tuesday, 1 October 2024 17:10 (15 minutes)

One of the main benefits of modern radio astronomy, its ability to collect more higher-resolution and wider-bandwidth data from more and more antennas is now also starting to become one of its greatest problems. The advent of cutting-edge radio telescopes, such as MeerKAT, a precursor to the Square Kilometre Array (SKA), has made it impractical to rely on the traditional method of storing the raw data for extended periods and then manually processing it. Furthermore, the high data rates necessitate the use of High-Performance Computing (HPC), yet existing common radio astronomical data reduction tools, like Common Astronomy Software Applications (CASA), are not well-suited for parallel computing. We have addressed these challenges in developing the ML-PPA (Machine Learning-based Pipeline for Pulsar Analysis). It is an automated classification system capable of categorizing pulsar observation data and assigning labels, such as “pulse”, “pure noise”, or various types of Radio Frequency Interference (RFI), to each time fragment, represented as a 2D time-frequency image or “frame”. The analysis is performed by a Convolutional Neural Network (CNN). Given the highly imbalanced distribution of different frame types in real data (e.g. only 0.2% are “pulses”), it is essential to generate artificial data sequences with specific characteristics to effectively train such systems. To achieve this, “digital twins” were developed to replicate the signal path from the source to a pulsar-observing telescope. A corresponding pipeline was created and tested in Python, and then rewritten in C++, making it more suitable for HPC applications. The initial version of the ML-PPA framework has been released and successfully tested. This talk presents a comprehensive overview of the project, its current status and future prospects.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary authors: Mr KAZANTSEV, Andrei (MPIfR Bonn); Prof. BERTOLDI, Frank (University of Bonn); Dr DANGE, Gautam (FIAS Frankfurt); Mr TRATTNER, Marcel (HTW Berlin); Dr SAHA, Tanumoy (HTW Berlin); Mr OELKERS, Tim (HTW Berlin); PIDOPRYHORA, Yurii (MPG - Max-Planck--Gesellschaft); Prof. HESSLING, Hermann (HTW Berlin)

Presenter: PIDOPRYHORA, Yurii (MPG - Max-Planck-Gesellschaft)

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **94** Type: **Long Talk**

The oldest traditional Trawler of Cyprus on its way to the Data Cloud and MemoryTwin

Tuesday, 1 October 2024 15:15 (15 minutes)

Lambousa is a 25-meter long wooden boat the type of liberty, built in 1955 in Greece. It was registered in Cyprus in 1965 and was used as a fishing trawler until 2004, when it was withdrawn according to EU Fishing Policy (EU Directive 2008/56/EC). The boat was preserved in the sea, as a monument of the local cultural heritage by the Municipality of Limassol. In 2020, the boat was dry docked and a European fund of more than one million Euro, was acquired for its full restoration. The project began in January 2023, undertaken by a local marine maintenance company. More than 20 different traditional craftsmen were engaged in a combination of simultaneous works and completed the restoration in one year. The project was under the supervision of a municipal engineers'team and an archaeologist-consultant and superintendent, in order to record the restoration procedures and follow traditional shipbuilding technics during the restoration.

This, constitutes the largest, the most in detail renovation, the most expensive and complex multidisciplinary project of its type in Cyprus and most probably in the Eastern Mediterranean.

The UNESCO Chair on Digital Cultural Heritage at CUT team, in cooperation with the Municipality of Limassol and with the support of two EU projects H2020 ERA Chair Mnemosyne and the Digital Europe Eureka3D, undertook the detail 2D and 3D survey of the boat including its entire intangible/memory.

For the digital surveying a high-resolution photogrammetry and LIDAR was undertaken, which concluded with an accurate 3D model. The entire data acquisition and survey were based on the results of the newly published EU Study on quality in 3D digitisation of tangible cultural heritage.

In addition, an online platform for the holistic digital documentation of the boat including its entire biography/memory is under development to serve further research and the multidisciplinary community of users. The complex 3D reconstruction of the trawler and its related records such Paradata and Metadata will be harvested in Europeana and pesented during the Europeana's TwinIT-Event at the headquarters of the European Commission in Brussels on the 14th of May 2026.
This is the first time in the EU that a 3D object is harvested in Europeana using the Eureka3D methodology based on the latest requirements from the EU policy on the Data Cloud in Cultural Heritage by utilizing the full power of EGI Data Cloud Infrastructure.

This contribution discusses the boat's characteristics, its restoration procedures and the positive impact for the preservation of the local Maritime Cultural Heritage by creating the exact #MemoryTwin and make all information and data available under open-access to the entire world.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: Prof. IOANNIDES, Marinos (Cyprus University of Technology)

Presenter: Prof. IOANNIDES, Marinos (Cyprus University of Technology)

Session Classification: Inside Data Spaces: Enabling data sharing paradigms

Contribution ID: **95** Type: **Long Talk**

INFN-DataCloud a distributed infrastructure supporting multi purpose Scientific data analytics services

Wednesday, 2 October 2024 15:00 (20 minutes)

In Italy thanks to the investment coming from Italy's Recovery and Resilience Plan projects (mainly ICSC: https://www.supercomputing-icsc.it/, Terabit: https://www.terabit-project.it/, and others) INFN is implementing a distributed HW, SW infrastructure that will be used to support very heterogeneous scientific use cases, not only coming from the INFN Community, but also with the full Italian scientific community.
INFN DataCloud aims to create a next-generation integrated computing and network infrastructure by 2025. The primary goal is to enhance collaboration and information exchange among Italian scientific communities.
In particular both within ICSC and Terabit, INFN is collaborating with CINECA and GARR with the aim of building an integrated computing and network infrastructure to eliminate disparities in access to high-performance computing across Italy.
The ICSC project has a very large partnership of around 55 entities both public and private bodies that assure that the infrastructure we will put in operation by the end of the project has to be able to support requirements from about full Italian scientific communities.
The INFN DataCloud project must address technical and not technical challenges related to network architecture, data storage, and computational resources together with the distributed team of people working in all the projects activities from each of the INFN main sites (about 12 distributed data center in Italy).
One of the main challenges is Data Security and Privacy: Handling sensitive scientific data requires robust security measures. The project must address data encryption, access controls, and compliance with privacy regulations to protect researchers'work.
The computing and storage distributed facilities upgraded by Italy's Recovery and Resilience Plan projects that founded many fat nodes (we call those "HPC-Bubbles") with GPU, many cores CPU, large RAM Memory and SSD based storage increase the resources available to the researchers in order to support the modern requirements of AI algorithms and very large dataset needed for most of the science (Physics, bioinformatics, earth studies, climate etc).
In the contest of the Data access/management/transfer, the INFN DataCloud project is federating both posix and Object storage with a geographically distributed data lake.
In the talk we will show both technical and not technical solutions implemented to build a transparent high-level federation of both Compute and Data resources, and how the development, operation and user support activities are organized in such a heterogeneous environment.
In summary, the INFN DataCloud project faces a mix of technical, organizational, and logistical challenges. However, its potential impact on Italian scientific communities makes overcoming these hurdles worthwhile.
Moreover, the INFN DataCloud project has democratized access to high-performance computing and data resources, empowering Italian researchers to accelerate their scientific endeavors.

Topic

Needs and solutions in scientific computing: National and scientific perspectives

Primary author: DONVITO, Giacinto (INFN)

Co-authors: MARTELLI, Barbara; PELLEGRINO, Carmelo (INFN-Cnaf); GRANDI, Claudio (INFN); CESINI, Daniele (INFN); SPIGA, Daniele; MICHELOTTO, Diego (INFN); GIORGIO, Emidio (INFN); CARBONE, Luca (INFN-Milano); SGARAVATTO, Massimo (INFN); FOGGETTI, Nadina; CIASCHINI, Vincenzo (INFN); STALIO, stefano

Presenter: DONVITO, Giacinto (INFN)

Session Classification: National Perspectives: EGI Member Countries' Latest Developments and Future Initiatives

Contribution ID: **96** Type: **Short Talk**

Cross-disciplinary data infrastructures for research in South Africa: A case of DIRISA

Wednesday, 2 October 2024 12:00 (20 minutes)

Significant investments have been made by the South African government in efforts to support the e-research environments across multiple disciplines in the South African research landscape. This has given birth to the National Integrated Cyberinfrastructure Systems (NICIS) which currently supports communication networks, high performance computing (HPC), data storage and research data management services across the research landscape of South Africa.

The Data Intensive Research Initiative of South Africa (DIRISA) is tasked with dealing with the increased proliferation of data that is being generated from new technologies and scientific instruments. Large amounts of research data is created daily which introduces new challenges for DIRISA and requires increased efforts towards solving these challenges. This presentation discusses the primary objectives of DIRISA which are - providing a national research data infrastructure, providing coordination and advocacy, developing human capital skills, providing research data management services and providing thought leadership in local and international efforts.

DIRISA is critical for researchers that are engaged in data intensive research and international research collaboration as it is able to bridge the gaps of infrastructure limitations at various public institutions by providing dedicated access to data and high capacity data storage. The comprehensive suite of research data management services offered by DIRISA ensures that South African researchers derive value from their research data. DIRISA offers research data management services that span the entire research data lifecycle such as: single sign-on authentication and authorization mechanisms, tools for crafting data management plans, metadata catalogue and management, digital object identifier (DOI) issuance, and tools for data depositing, data sharing and long-term archival. As underscored by DIRISA's objectives, community training assumes paramount importance, enabling researchers to effectively harness the technologies and tools provided by the initiative. This presentation also deliberates on DIRISA's diverse human capital development and training endeavors that not only cover the researchers but also reach down to high school level students.

An impact assessment of how DIRISA services have contributed to the advancement of research in the country along with the challenges and gaps that currently exist at DIRISA are discussed. This presentation provides a framework that can be used by other African and developing countries towards creating cross-disciplinary data infrastructures through an analysis and evaluation of DIRISA by focusing on infrastructure, research data management services, policies and human capital development. DIRISA aims to provide a platform for supporting researchers through the provision of data infrastructure for South Africa and the lessons from DIRISA can have applicability for the African context. Finally, the future directions for addressing emerging challenges in data management and infrastructure development are discussed to provide a glimpse into how data infrastructure can adapt to the changing research data management landscape.

Topic

Needs and solutions in scientific computing: National and scientific perspectives

Primary author: Dr SHOZI, Nobubele (Council for Scientific and Industrial Research (CSIR))

Presenter: Dr SHOZI, Nobubele (Council for Scientific and Industrial Research (CSIR))

Session Classification: Global perspectives on advancing Open Science with computational infrastructures

Contribution ID: **97** Type: **Long Talk**

Incident Response (IR) on credentials provided through a global federated AAI service.

Access to EGI services is also provided through Federated AAI, as for example eduGAIN (https://edugain.org/)i. In particular egi-checkin, our IdP/SP proxy is part of this infrastructure and allows users from around the globe to authenticate with their home institutions IdP and receive credentials that could be used to access to EGI services.
Our IR procedure therefore also have to cover the aspect to be able to deal with compromised accounts provided through the eduGAIN service.

In this workshop we want to raise awareness of the complexity of IR where we have to coordinate our IT security activities with a global service managing the authentication of users.

The focus is on the inter federation aspect of IR, and what the key players in IR can do, to deal with an incident requiring the collaboration of the operators (Federation, IdP, SP) contributing to the eduGAIN service and the coordination with eduGAIN CSIRT and EGI CSIRT.

The participants will get an introduction to eduGAIN, the relevant security policies, the key security roles, and the IR supporting frameworks like SIRTFI.

After that, the participants will have to deal with an artificial incident and apply the IR concepts presented before in a Table Top Exercise (TTX) set-up. Although it's a "made up" scenario, it consists of real world incidents we had to deal with.

Each of the security roles will be taken by a group, in which the possible reaction to the developing incident response situation needs to be discussed and the found reaction fed back to the incident coordinator.
The goal here is to identify the organisational obstacles we may run into during IR, and check if the existing procedures are clear enough.

The enabled learning objectives (what the participants should learn) include:
* IdP/SP logfile analysis (check for/find a reported ID)
* know SIRTFI v2, and understand applying it
* Know how eduGAIN is organised, role of Federations, eduGAIN and eduGAIN CSIRT
* Name the risks of federated Identity Management.

Topic

Trust and Security: Access control

Primary authors: KOURIL, Daniel (CESNET); GROEP, David (Nikhef); KELSEY, David (STFC); DUSSA, Tobias (DFN-CERT)

Presenter: GABRIEL, Sven (NIKHEF)

Session Classification: Trust & Security

Contribution ID: **98** Type: **Short Talk**

STAC at CEDA - a scalable, standards-based search system

Wednesday, 2 October 2024 11:45 (15 minutes)

The Centre for Environmental Data Analysis (CEDA) stores over 20 Petabytes of atmospheric and Earth observation data. Sources for the CEDA Archive include aircraft campaigns, satellites, automatic weather stations and climate models, amongst many others. The data mainly consists of well-described formats such as netCDF files but we also hold historical data where the format cannot be easily discerned from the file name and extension.

CEDA are investigating the SpatioTemporal Asset Catalogue (STAC) specification to allow for user interfaces and search services to be enhanced and facilitate interoperability with user tools and our partners. We are working to create a full-stack software implementation including an indexing framework, API server, web and programmatic clients, and vocabulary management. All components are open-source so that they can be adopted and co-developed with other organisations working in the same space.

We have built the "stac-generator", a tool that can be used to create a STAC catalog, which utilises a plugin architecture to allow for more configurability. A range of input, output, and extraction methods can be selected to enable data extraction across the diverse archive data and its use by other organisations. Elasticsearch was chosen to host the indexed metadata because it is performant, highly scalable and supports semi-structured data - in this case the faceted search values related to different data collections. As STAC's existing API was backed by an SQL database this called for the development of a new ES backed STAC API, which has now been merged back into the community developed API as an alternate database backend. We have also developed several extensions to the STAC framework to meet requirements that weren't met by the core and community functionality. These include an end-point for interrogating the facet values, as queryables, and a free-text search capability across all properties held in the index.

The developments of our search system has also included pilots for the Earth Observation Data Hub (EODH) and a future version of the Earth System Grid Federation (ESGF) search service, in which we have created an experimental index containing a subset of CMIP6, CORDEX, Sentinel 2 ARD, Sentinel 1, and UKCP data to investigate performance and functionality.

With the increasing demand on cloud-accessible analysis-ready data we are seeing in several of our upcoming projects. We have started to explore Kerchunk a lightweight non-conversion approach for referencing existing data, which works with open-source python packages like fsspec and xarray. And are looking to integrate this with our STAC work.

It is the aim of project to increase the interoperability of our search services, as well as foster collaboration with other organisation who share our goals. Additionally, it is hoped that this work will allow for greater and easier access to the data held at CEDA.

Topic

Data innovations: Data Analytics, Sensitive Data/FAIR Data

Primary author: EVANS, Rhys

Presenter: EVANS, Rhys

Session Classification: Maximizing Data Efficiency and Availablity: Innovations in Data Management Platforms

Contribution ID: **99** Type: **Demonstrations & Tutorials**

Improving Biodiversity Data and Services discoverability: the LifeWatch ERIC Metadata Catalogue

Tuesday, 1 October 2024 18:30 (30 minutes)

The LifeWatch ERIC Metadata Catalogue is a centralized platform for discovering and disseminating data and services, ensuring equitable access to information and promoting inclusivity in biodiversity and ecosystem research and conservation efforts. LifeWatch ERIC Metadata Catalogue was designed to tackle several challenges, critical for biodiversity and ecosystem research:

- **Data & Services Fragmentation**: Biodiversity data and services are often scattered across various sources, including research institutions, government agencies, non-profit organizations, and citizen science projects. This fragmentation makes it challenging to discover, access, and integrate relevant datasets or services, hindering comprehensive analyses and decision-making.
- **(Meta)Data & Services Heterogeneity**: Biodiversity data and services come in diverse formats, structures, and standards, making it difficult to harmonize and reconcile information from different sources. This heterogeneity poses significant barriers to data interoperability, integration and analysis, impeding collaborative research efforts.
- **Metadata Inconsistencies**: Inconsistent or incomplete metadata descriptions further exacerbate challenges related to data and services discovery and interpretation. Without standardized metadata practices, researchers may struggle to understand the context, quality, and limitations of available datasets and services, leading to potential misinterpretations or biases in analyses and conclusions.
- **Data Quality and Reliability**: Ensuring the quality, accuracy, and reliability of biodiversity data is paramount for robust scientific research and evidence-based decision-making. However, without comprehensive metadata documenting data and services provenance, methodologies, and quality assessments, it becomes challenging to assess the trustworthiness of available datasets, potentially compromising the integrity of research outcomes.
- **Limited Data Accessibility**: Biodiversity data and services accessibility remains a significant concern, particularly in regions with limited technological infrastructure or resources.

LifeWatch ERIC has adopted GeoNetwork as technology for its Metadata Catalogue, obtaining numerous advantages: open-source flexibility, geospatial capabilities, standards compliance, user-friendly interface, metadata management features, interoperability, scalability, performance, and community support. Different showcases have been developed in the latest years to demonstrate the data and services discoverability across different institutions and research infrastructures, like the one developed jointly with ANAEE and eLTer for the Research Sites in the ENVRI FAIR contest (https://envri.eu/home-envri-fair/), or the metadata harvesting from the GBIF network. The Metadata Catalogue has been continuously improving, progressively incorporating new features: a template based on a profile of EML 2.2.0 standard, developed for ecological datasets; a template with the LifeWatch ERIC profile of ISO 19139/119, developed for services; customizations to supply DOI assignment, if needed, using the Datacite services; metadata FAIRness evaluation, added thanks to the F-UJI tool (https://www.f-uji.net). Continuos developments are planned to keep on improving the quality of metadata, the next step will be the integration with the LifeWatch ERIC semantic repository Ecoportal (ecoportal.lifewatch.eu) and with AI technology to ensure metadata' s consistency.

Furthermore, LifeWatch ERIC is an EOSC Candidate Node and is working to federate the Metadata Catalogue in the context of the EOSC-Beyond project.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: FIORE, Nicola (LifeWatch ERIC)

Co-authors: Prof. BASSET, Alberto (LifeWatch ERIC); Dr ROSATI, Ilaria (CNR IRET); Dr VAIRA, Lucia (LifeWatch ERIC); MARROCCO, Vanessa (LifeWatch ERIC)

Presenter: FIORE, Nicola (LifeWatch ERIC)

Session Classification: Demonstrations & Posters

Contribution ID: **101** Type: **Short Talk**

Advancing Research and Innovation: The Role of the OpenAIRE Infrastructure in European Scholarly Communication

OpenAIRE is a pan-European non-profit organization that provides e-infrastructures to support scholarly communication. It connects human capital with advanced ICT services and is supported by a consortium of European research institutions. This organization plays a key role in promoting open science by making research more accessible, transparent, and collaborative. OpenAIRE aligns its services with the FAIR principles, which stand for Findable, Accessible, Interoperable, and Reusable, enhancing the discoverability and usability of research outputs.

This presentation will focus on the services OpenAIRE offers, emphasizing its commitment to promoting open science and improving scholarly communication across various fields. It will explain how OpenAIRE helps researchers meet open access requirements and supports the European Commission's Open Science agenda through targeted policies, tools, and guidelines. A significant part of the discussion will highlight OpenAIRE's collaboration with major European e-infrastructures such as EUDAT and EGI, showcasing how these partnerships contribute to the broader European Open Science Cloud (EOSC).

The structure of OpenAIRE, including its repository networks and the connections between research artifacts in the OpenAIRE Graph, will also be discussed. The aim is to provide insights into how OpenAIRE supports day-to-day research and plays a role in shaping strategic policies and infrastructure developments crucial for advancing research excellence and innovation in Europe. Finally, the presentation will look ahead to OpenAIRE's future, focusing on expanding its impact, improving integration of services, and enhancing engagement with researchers across all disciplines to create a more cohesive European research area. This session will invite attendees to explore collaboration opportunities within OpenAIRE's growing network.

Topic

EOSC Developments and Open Science: EOSC

Primary author: Dr MANGHI, Paolo (OpenAIRE AMKE)

Co-author: Dr MALAGUARNERA, Giulia (OpenAIRE)

Presenter: Dr MANGHI, Paolo (OpenAIRE AMKE)

Session Classification: Closing session

Contribution ID: **102** Type: **Short Talk**

Tagus river-to-ocean collaboratory for thematic digital twins and collaborative management

Tuesday, 1 October 2024 16:00 (15 minutes)

Digital Twins provide a virtual representation of a physical asset enabled through data and models. They can be used for multiple applications such as real-time forecast of system dynamics, system monitoring and controlling, and support to decision making. Recent tools take advantage of the huge online volume of data streams provided by satellites, IoT sensing and many real-time surveillance platforms, and the availability of powerful computational resources that make process-solving, high-resolution models and AI-based models possible, to build high accuracy replicas of the real world.

The Tagus estuary is the largest estuarine region in the Iberian Peninsula and holds a multitude of services of huge economic, environmental and social value. The management of this large system is quite complex and there are often conflicting uses that require high resolution, complex tools to understand and predict its dynamics and support any interventions. Simultaneously, the Tagus basin raises concerns related to inundation and erosion (Fortunato et al., 2021) and water quality (Rodrigues et al., 2020). A variety of models have been applied here to address multiple concerns from physical to water quality and ecology. At the same time, the Tagus holds several observatories supported by data (e.g. CoastNet, http://geoportal.coastnet.pt/) and integrated model and data (UBEST, http://ubest.lnec.pt/). In spite of all these efforts, no integrated infrastructure, from river to ocean, accounting for the city of Lisbon and other important cities' drainage, was available to support management and research alike, allowing for users to interact with data and models to build customized knowledge.

The CONNECT project, funded through the CMEMs coastal downscaling programme, developed a multi-purpose collaboratory that combines digital twin technology, a smart coastal observatory tool (Rodrigues et al., 2021) and a monitoring infrastructure –CoastNet, to address both inundation and water quality concerns. The work takes advantages of the on-demand, relocatable coastal forecast framework OPENCoastS (Oliveira et al., 2021) to build a user-centered, multi-purposes DT platform that provides tailored services customized to meet the users' needs. A combination of process-based modeling in the estuary, using SCHISM suite, and AI modeling for the river inflow, using the AI4Rivers model builder, supports the automatic creation of both 2D and 3D predictions daily. Model performance is automatically shared with the users, both through online comparison with the in-situ and remote sensing data from CoastNet and CMEMS, and the calculation of indicators at several time scales.

Fortunato, A.B., Freire, P., Mengual, B., Bertin, X., Pinto, C., Martins, K., Guérin, T., Azevedo, A., 2021. Sediment dynamics and morphological evolution in the Tagus Estuary inlet. Marine Geology 440, 106590.

Oliveira, et al, 2021. Forecasting contrasting coastal and estuarine hydrodynamics with OPENCoastS, Environmental Modelling & Software, Volume 143,105132.

Rodrigues, M., Cravo, A., Freire, P., Rosa, A., Santos, D., 2020. Temporal assessment of the water quality along an urban estuary (Tagus estuary, Portugal). Marine Chemistry 223, 103824.

Rodrigues, M., Martins, R., Rogeiro, J., Fortunato, A.B., Oliveira, A., Cravo, A., Jacob, J., Rosa, A., Azevedo, A., Freire, P., 2021. A Web-Based Observatory for Biogeochemical Assessment in Coastal Regions. J ENVIRON INFORM.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: Dr RODRIGUES, Marta (LNEC)

Co-authors: OLIVEIRA, Anabela (National Laboratory for Civil Engineers); Dr DE JESUS, Gonçalo (LNEC); Dr B. FORTUNATO, André (LNEC); Mr MARTINS, Ricardo (LNEC); Ms MARDANI, Zahra (LNEC); Dr ALVES, Elsa (LNEC)

Presenter: OLIVEIRA, Anabela (National Laboratory for Civil Engineers)

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **103** Type: **Poster**

The interTwin IMS. Paving the way for the project exploitation

Tuesday, 1 October 2024 18:00 (1 hour)

interTwin co-designs and implements the prototype of an interdisciplinary Digital Twin Engine (DTE) - an open-source platform based on open standards, that offers the capability to integrate with application-specific Digital Twins (DTs). Its functional specifications and implementation are based on a co-designed interoperability framework and conceptual model of a DT for research - the DTE blueprint architecture. The ambition of interTwin is to create consensus on a common approach to the implementation of DTs that is applicable across the whole spectrum of scientific disciplines that will facilitate developments and interoperability across different DTs.

This poster depicts the interTwin Innovation Management System (IMS), the framework developed by the project that ensures that all project results are systematically captured, assessed for exploitation readiness and validated along with an improvement cycle to strengthen them. The IMS includes activities for 1) understanding the market, technological and political context of the project in order to provide the necessary market pull information to feed to the project solutions, 2) Capturing and identifying project results including the ownership, intellectual property and protection mechanisms and future access conditions to push technologies and services to the up to the market, 3) and preparing and monitoring exploitation, business and sustainability plans.

Main aim of the poster is to seek and foster discussion among conference participants about exploitation and collaboration opportunities, as the project is heading towards the final year, the first release of the SW components has been made available, early key exploitable results have been described. As in general to discuss how an innovation management framework can help to maximise the exploitation opportunities and therefore, the impact of Horizon Europe projects.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: SALAZAR, Xavier (EGI)

Presenter: SALAZAR, Xavier (EGI)

Session Classification: Demonstrations & Posters

Contribution ID: **104** Type: **Short Talk**

Collaborative research in the cloud with Notebook-as-a-VRE (NaaVRE)

Wednesday, 2 October 2024 17:15 (10 minutes)

Many scientific problems, such as environmental research or cancer diagnosis, require large data volumes, advanced statistical or AI models, and distributed computing resources.
To help domain scientists conduct their research more effectively they need to reuse resources like data, AI models, workflows, and services from different sources to address complex challenges. Sharing resources requires collaborative platforms that facilitate advanced data science research that offers: discovery access, interoperation and reuse of research assets, and integration of all resources into cohesive observational, experimental, and simulation investigations with replicable workflows. Virtual Research Environments (VREs) effectively supported such use cases offering software tools and functional modules for research management. However, while effective for specific scientific communities, existing VREs often lack adaptability and require substantial time investment for incorporating external resources or custom tools. In contrast, many researchers and data scientists prefer notebook environments like Jupyter for their flexibility and familiarity.

To bridge this gap we propose a VRE solution for Jupyter Notebook-as-a-VRE (NaaVRE).

The NaaVRE empowers users to construct functional blocks by containerizing cells within notebooks, organizing them into workflows, and overseeing the entire experiment cycle along with its generated data. These functional blocks, workflows, and data can then be shared within a common marketplace, fostering user communities and tailored Virtual Research Environments (VREs). Additionally, NaaVRE seamlessly integrates with external repositories, enabling users to explore, select, and reuse various assets such as data, software, and algorithms. Lastly, NaaVRE is designed to seamlessly operate within cloud infrastructures, offering users the flexibility and cost efficiency of utilizing computational resources as needed.
We showcase the versatility of NaaVRE by building several customized VREs that support specific scientific workflows across different communities. These include tasks such as extracting ecosystem structures from Light Detection and Ranging (LiDAR) data, monitoring bird migrations via radar observations, and analyzing phytoplankton species. Additionally, NaaVRE finds application in developing Digital Twins for ecosystems as part of the Dutch NWO LTER-LIFE project.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: PELOUZE, Gabriel (LifeWatch ERIC, Virtual Lab & Innovation Center (VLIC), Amsterdam, The Netherlands); KOULOUZIS, Spiros (LifeWatch ERIC, Virtual Lab & Innovation Center (VLIC), Amsterdam, The Netherlands); ZHAO, Zhiming (Multiscale Networked Systems, University of Amsterdam, Amsterdam, The Netherlands)

Presenter: PELOUZE, Gabriel (LifeWatch ERIC, Virtual Lab & Innovation Center (VLIC), Amsterdam, The Netherlands)

Session Classification: Simplifying Data-Driven Science with User-Friendly Platforms and Gateways

Contribution ID: **105** Type: **Short Talk**

Integrating data repositories with HPC resources for execution of VHT models

Tuesday, 1 October 2024 17:25 (15 minutes)

Introduction

The digital twin concept is gaining traction in research, demanding substantial computational power for simulations. Sano Centre for Computational Medicine, in collaboration with ACC Cyfronet AGH, is actively developing tools to optimize high performance computing (HPC) resources. Our focus is on providing scientists with a user-friendly toolkit for seamless model execution. This paper introduces the integration of the Model Execution Environment platform with data repositories, streamlining data management for researchers.

Description of the problem

Harnessing HPC resources necessitates specific expertise and extensive data management, posing challenges for researchers. Additionally, sharing processed data and research results among teams demands adherence to fair involvement rules, involving external services and consuming valuable time. Our aim is to alleviate these challenges by providing a comprehensive platform for efficient data management.

Related work

While Pegasus and others operate on various infrastructures, Model Execution Environment (MEE) focuses on an established execution framework. Our unique approach prioritizes seamless data staging across diverse repositories, such as Dataverse and Zenodo, enhancing flexibility, streamlining execution and fostering effortless collaboration.

Solution of the problem

Our platform integrates with Dataverse and Zenodo APIs, enhancing efficiency and collaboration by eliminating intermediaries. Customizable repository rules ensure fair data sharing, safeguarding confidentiality.

Conclusions

Our research has resulted in a sophisticated toolkit for medical research efficiency. Future plans include broader integration, simplified data retrieval via Digital Object Identifiers, enhancing accessibility of our toolkit.

Acknowledgements. This publication is (partly) supported by the European Union's Horizon 2020 research and innovation programme under grant agreement ISW No 101016503, supported by the European Union's Horizon 2020 research and innovation programme under grant agreement Sano No 857533 and was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016227.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: Mr MEIZNER, Jan (Sano Centre for Computational Medicine); Mr ZAJĄC, Karol (Sano Centre for Computational Medicine); MALAWSKI, Maciej (Sano Centre for Computational Medicine); NOWAKOWSKI, Piotr (Sano Centre for Computational Medicine); Mr ZHYHULIN, Taras (Sano Centre for Computational Medicine)

Presenter: Mr ZHYHULIN, Taras (Sano Centre for Computational Medicine)

Session Classification: Bridging the Gap: Integrating the HPC Ecosystem

Contribution ID: **106** Type: **Short Talk**

Reviving Virtual Access to fund data usage in the EOSC Federation?

Wednesday, 2 October 2024 11:20 (20 minutes)

In order to fulfil its "catalysing and leveraging role"in the development of European Research Infrastructures (RIs) and e-Infrastructures, the European Commission (EC) introduced mechanisms in the previous Framework Programme to provide researchers who participated in EC-funded projects with access to European RIs. Access to "depletable"resources, including physical and remote access to facilities, was regulated by Trans-national Access (TNA), while access to "non-depletable"resources (like e.g. data) was done via Virtual Access (VA). TNA was restricted to partners in the consortium, while VA could be extended to users outside as well.
TNA and VA allowed projects to use money from grants to reimburse providers for the costs incurred in the provision of the service, including support-related costs, and covered also any travel costs of researchers accessing the services. This approach helped pool resources across Europe to "properly address the cost and complexity of new world-class RIs"and ensured "wider and more efficient access to and use of"European RIs. By transferring the money directly to the provider, the EC enabled researchers to use facilities around Europe free at the point of use.
While continued for Destination INFRASERV calls of the current Framework Programme, VA and TNA are not included as eligible costs for projects awarded in Destination INFRAEOSC calls, causing several digital services to become discontinued, and making access by researchers to others difficult due to the lack of funding sources that allow them to pay for their use.
However, demand for access to datasets, data processing applications and other data-related services is expected to continue increasing. Processing, analysis and storing of data carry considerable costs when incurred by researchers outside of their own communities, linked to the infrastructure, maintenance, and operating staff. In the context of the Open Science paradigm, the EC and the EU Member States (EU MS) have agreed that enabling "secondary use"of data is needed aims to provide access to any potential user to all data obtained with public funding. This key ingredient in the "EOSC Federation"put forward by the EC, the EOSC Association and the EU MS is expected to result in a further increase in the costs incurred by RIs, since the additional access to and processing of data by researchers not included in the original user base have in general not been foreseen when planning RIs. Some RIs will therefore face problems to fulfil the requirements placed on them.
We argue that a mechanism that replaces VA and TNA is needed in the future FP10 for a successful implementation of EOSC as a federation of "EOSC nodes". The EC and EU MS must agree on a way by which data providers can be reimbursed for the extra costs generated by the "secondary use"of data such that access to data remains essentially free at the point of use for researchers.
In our talk we will evaluate the current situation according to the plans to build the EOSC Federation, and will suggest possible ways forward to be discussed with the audience.

Topic

Data innovations: Business models

Primary author: REY MAZON, Miguel (Graz University of Technology)

Co-authors: ROBERTSON, Dale; JONG, DE, Franciska (CLARIN ERIC); PICARD, John (CLARIN-ERIC)

Presenter: REY MAZON, Miguel (Graz University of Technology)

Session Classification: Empowering Open Science: EGI Community's Impact on EOSC

Contribution ID: **107** Type: **Long Talk**

Interoperable Workflow Efficiency: Exploring the Integration of OpenEO, CWL , and EOEPCA for Seamless Data Processing and Modeling

Tuesday, 1 October 2024 15:55 (15 minutes)

Contemporary HPC and cloud-based data processing is based on complex workflows requiring close access to large amounts of data. OpenEO process graphs allow users to access data collections and create complex processing chains. Currently OpenEO can be accessed via one of the clients in JavaScript, R, or Python. Direct access to data is provided via Spatio Temporal Asset Catalogues (STAC). As part of our ongoing research under the InterTwin project, the focus is on extending the capabilities of OpenEO to support the management and execution of OGC Application Packages.

An Application Package allows users to create automated, scalable, reusable and portable workflows. It does so by creating, for example, a Docker Image containing all of the application code and dependencies. The workflow is described with Common Workflow Language (CWL). The CWL document references all of the inputs, outputs, steps, and environmental configurations to automate the execution of an application.

The execution is handled by an Application Deployment and Execution Service (ADES) coming from the EOEPCA project. It is a Kubernetes based processing service capable of executing Application Packages via OGC WPS1.0, 2.0 OWS service, and the OGC API-processes. In many ways, the core goals and objectives between EOEPCA and the InterTwin project align well. The focus is on allowing workflows to be seamlessly executed without the need of substantial code rewrites or adaptations to a specific platform.

OpenEO is on its way to become an OGC community standard. It currently supports a large set of well-defined cloud optimized processes that allow users to preprocess and process data directly in the cloud. The goal is to integrate the Application Deployment and Execution Service (ADES) from the Earth Observation Exploitation Platform Common Architecture (EOEPCA) project to create a fusion between OpenEO process graphs and Application Packages.

Application Package support in OpenEO is a means of providing users the ability to bring their applications directly to the platform. Instead of having to reimplement the code in a process graph, it is possible to wrap any existing application. The fusion of process graphs and CWL based application workflows extends OpenEO for users that would like to perform testing of their models, ensemble models etc. while utilizing the same process graph and direct access to data.

Many complex workflows require some kind of data preprocessing. This preprocessing can be done using OpenEO process graphs and then be directly sent for execution to an Application Package to run the actual process. OpenEO complements STAC by providing a standardized interface and processing framework for accessing and analyzing Earth observation data. Having all of the data and tools readily available on a single platform creates an accessible, interoperable, and a reproducible environment for users to create efficient workflows.

The ability to create standardized, reusable workflows using CWL and execute them on distributed computing resources via ADES can significantly reduce the time and effort required for data processing tasks. Researchers can focus on algorithm development and data analysis rather than worrying about infrastructure management or software compatibility issues.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary authors: JACOB, Alexander (Eurac Research - ACCADEMIA EUROPEA DI BOLZANO); ZVOLENSKY, Juraj (Eurac); CLAUS, Michele (EURAC); CAMPALANI, Piero (Eurac Research); FERRARIO, iacopo (EURAC Reseach)

Presenters: JACOB, Alexander (Eurac Research - ACCADEMIA EUROPEA DI BOLZANO); ZVOLENSKY, Juraj (Eurac)

Session Classification: Cloud Compute federation and national initiatives

Contribution ID: **108** Type: **Long Talk**

PLG Portal - platform for managing distributed computing resources in a federated infrastructure

Wednesday, 2 October 2024 15:20 (20 minutes)

PLGrid is a nationwide computing infrastructure designed to support scientific research and experimental development across a wide range of scientific and economic fields. PLGrid provides access to supercomputers, quantum computers, specialized accelerators for artificial intelligence, cloud computing, disk storage, optimized computing software and assistance from experts from the entire Poland. The Polish PLGrid infrastructure is managed by the PLGrid Consortium, established in January 2007, which includes the following computing centers: Cyfronet, ICM, PSNC, CI TASK, WCSS, NCBJ.

In order to make it easier for users to use the available distributed resources, it was necessary to create a centralized platform that includes many applications, tools, and solutions, with the PLGrid Portal as its main component. The platform has been developed since the beginning of the PLGrid Consortium, going through successive new versions. Thanks to the experience gained over 10 years, it has become a mature and flexible solution. This allows us to easily adapt to changing requirements, which makes us able to effectively respond to new challenges in terms of both user and operational convenience in federated infrastructure.

As the main application in PLGrid Infrastructure, the Portal PLGrid consists of many elements from creating an account to requesting distributed resources through PLGrid grants. All the necessary functionalities for the User like creating and managing an account, affiliations, subordinates, teams, services (access to resources), applications, and ssh keys. However, it is the process of requesting resources that is specific. The user fills out a grant application, which is negotiated with resource administrators - and after the grant is completed, it must be settled. From our perspective, the most important thing was to realize such Portal that would be flexible. We have a variety of types: accounts, teams, services, and grants. In an easy way, new types of resources can be defined, that will have other limitations. For example, a specific account type cannot create a specific grant type or a given type of grant has other fields that are required to be provided by the user. On the operational side, we have many roles, where the main like Resource/Service Administrator have dedicated web views, the ability to replicate data to LDAP, and API access which allows you to synchronize all data among various HPC centers and different types of resources (computing, storage, object storage, cloud, etc).

Topic

Needs and solutions in scientific computing: Federated operation

Primary authors: ZEMLA, Andrzej (CYFRONET); Mr KASZTELNIK, Marek (ACK Cyfronet AGH)

Presenter: ZEMLA, Andrzej (CYFRONET)

Session Classification: National Perspectives: EGI Member Countries' Latest Developments and Future Initiatives

Contribution ID: **109** Type: **Short Talk**

Blue-Cloud 2026 –a Federated Data Discovery and Access Service

Wednesday, 2 October 2024 11:30 (15 minutes)

The Blue-Cloud 2026 HE project aims at a further evolution of the pilot Blue-Cloud open science infrastructure into a Federated European Ecosystem to deliver FAIR & Open data and analytical services, instrumental for deepening research of oceans, EU seas, coastal & inland waters. It also strives to become a major data and analytical component for the Digital Twins of the Oceans (DTO' s) as well as a blue print for a thematic EOSC node.

One of the key services is the Blue-Cloud Data Discovery & Access service (DD&AS), which federates key European data management infrastructures, to facilitate users in finding and retrieving multi-disciplinary datasets from multiple repositories through a common interface. In Europe, there are several research infrastructures and data management services operating in the marine and ocean domains. These cover a multitude of marine research disciplines, and providing access to data sets, directly originating from observations, and to derived data products. A number are ocean observing networks, while others are data aggregation services. Furthermore, there are major EU driven initiatives, such as EMODnet and Copernicus Marine. Together, these infrastructures constitute a diverse world, with different user interfaces. The Blue-Cloud DD&AS has been initiated to overcome this fragmentation and to provide a common interface for users by means of federation.

The pilot Blue-Cloud Data Discovery & Access service (DD&AS) already federates EMODnet Chemistry, SeaDataNet, EuroArgo-Argo, ICOS-Marine, SOCAT, EcoTaxa, ELIXIR-ENA, and EurOBIS, and provides common discovery and access to more than 10 million marine datasets for physics, chemistry, geology, bathymetry, biology, biodiversity, and genomics. As part of Blue-Cloud 2026 project, is is being expanded by federating more leading European Aquatic Data Infrastructures, such as EMSO, SIOS, EMODnet Physics, and EBI –Mgnify. In addition, upgrading is underway for optimising the FAIRness of the underpinning web services, incorporating semantic brokering, and adding data sub-setting query services.

The common interface includes facilities for discovery in two steps from collection to granular data level, and including mapping and viewing of the locations of data sets. The interface features a shopping mechanism, facilitating users to compose and submit mixed shopping baskets with requests for data sets from multiple BDIs. The DD&AS is fully based and managed using web services and APIs, following protocols such as OGC CSW, OAI-PMH, ERDDAP, Swagger API, and others, as provided and maintained by the BDIs. These are used to deploy machine-to-machine interactions for harvesting metadata, submitting queries, and retrieving resulting metadata, data sets and data products.

Presentation:
During the presentation more details will be given about the federation principles, the semantic brokerage, and the embedding of the DD&AS in the Blue-Cloud e-infrastructture, serving external users as well as users of Blue-Cloud Virtual Labs and EOV WorkBenches.

Topic

EOSC Developments and Open Science: EOSC

Primary authors: SCHAAP, Dick (Mariene Informatie Service MARIS BV); BOLDRINI, Enrico (CNR-IIA)

Presenter: KRIJGER, Tjerk (MARIS)

Session Classification: Maximizing Data Efficiency and Availablity: Innovations in Data Management Platforms

Contribution ID: **110** Type: **Short Talk**

Pioneering Integrated Data Access and Analysis Across Research Fields: A Dutch Initiative

Tuesday, 1 October 2024 16:45 (20 minutes)

This talk introduces a significant Dutch initiative designed to transform the landscape of data access and analysis for all research fields, using the ODISSEI metadata portal as a specific example for the social sciences and humanities (SSH) community. Our integrated workflow begins with the Data Access Broker (DAB), developed by SURF, which standardizes data access requests and data transfers across diverse data providers, addressing the complexities of handling sensitive data with varying access scopes and policies.

Following data acquisition, the workflow advances to the Secure Analysis Environment (SANE), a Trusted Research Environment (TRE) facilitated by SURF. SANE allows researchers to work securely with sensitive data, while the data provider stays in full control of the data and tools within the TRE. Through Federated Identity and Access Management (FIAM) we simplify the collaboration between data providers and researchers. This cloud-based solution simplifies the often intricate process of sensitive data analysis, providing researchers with essential tools and access needed to drive forward their investigations.

In this presentation, we will not only explore the architectural and operational aspects of DAB and SANE but also outline our strategy for expanding these services beyond SSH to encompass all research domains. As a leading Dutch initiative for integrated data systems, our aim is to reveal the potential and current capabilities of this innovative framework, showcasing how the ODISSEI metadata portal serves as a model for other research communities.

Attendees will gain insight into the full workflow, from data discovery through to detailed analysis, and understand how this initiative is shaping a new frontier in research capabilities across the Netherlands.

Topic

Trust and Security: Trusted computing:

Primary authors: HESAM, Ahmad (SURF); Mr VAN DER MEER, Lucas (Erasmus University Rotterdam)

Presenter: HESAM, Ahmad (SURF)

Session Classification: Managing & Processing Sensitive Data

Contribution ID: **111** Type: **Demonstrations & Tutorials**

A Hybrid Reference Architecture for Cloud-based Quantum Computing Microservices with an Aerial-Ground Cooperative Robot Mapping Use Case

Wednesday, 2 October 2024 14:00 (30 minutes)

The increasing accessibility of quantum computing technology has opened up new avenues for exploring their potential applications in various scientific fields such as artificial intelligence, manufacturing, and finance. Many research scientists heavily depend on cloud computing infrastructures for their investigations. However, accessing actual quantum hardware resources, often located remotely, involves deploying and configuring different software components.

In this demonstration, we present our reference architecture [1,2], which combines cloud computing and quantum resources for easier initiation of experiments across diverse quantum compute resources. Our solution simplifies distributed quantum computing simulations in traditional cloud environments and provides access to remote quantum compute resources. The reference architecture is portable and adaptable to different cloud platforms, offering efficient utilization and application opportunities for research communities. It incorporates essential quantum software development kits (SDKs) with machine learning support and access to various quantum devices. Furthermore, we provide practical examples serving as references for constructing solutions to predefined problems. Our reference architecture prioritizes a user-friendly interface.

Additionally, our reference architecture enables continuous deployment of quantum applications, allowing seamless orchestration with traditional cloud-based applications. These quantum applications are deployed as microservices, accessible through standard REST APIs following open standards. This combination simplifies the design and deployment of quantum services, showcasing the effective utilization of standard methodologies from traditional service-oriented computing in this hybrid context.

The reference architecture is deployed amongst others within the National Laboratory for Autonomous Systems in Hungary 3 (abbreviated as ARNL) and within the National Research on Hydrogen-Powered, Cooperative Autonomous Remote Sensing Devices and Related Data Processing Framework project (TKP2021-NVA-01). We would like to demonstrate our reference architecture through a selected use case involving global route planning for autonomous vehicles such as unmanned ground vehicles (UGVs).

The aerial-ground cooperative mapping system aims to construct a comprehensive 3D model of an unknown environment by leveraging the perspectives of different agents. Drones enable rapid exploration, but it may result in incomplete 3D reconstructions due to limited aerial coverage. Aerial robot-collected data is used in structure from motion (SfM) pipelines to create initial 3D reconstructions. To address the aforementioned issue, we propose automatically locating unmapped regions and guiding the ground robot to complete the 3D model. Leveraging the initial aerial 3D model, we determine the shortest traversable paths between unmapped regions and utilize quantum computing to generate an optimal global route.

The research was partially supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Autonomous Systems National Laboratory Program. Project no. TKP2021-NVA-01 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

1 Quantum | HUN-REN Cloud - https://science-cloud.hu/en/reference-architectures/quantum
2 A. Cs. Marosi, A. Farkas, T. Máray and R. Lovas, "Toward a Quantum-Science Gateway: A Hybrid Reference Architecture Facilitating Quantum Computing Capabilities for Cloud Utilization," in

IEEE Access, vol. 11, pp. 143913-143924, 2023, doi: 10.1109/ACCESS.2023.3342749.
3 National Laboratory for Autonomous Systems - https://autonom.nemzetilabor.hu/

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: Dr MAROSI, Attila Csaba (HUN-REN SZTAKI); Mr BUGÁR-MÉSZÁROS, Barnabás (HUN-REN SZTAKI); Mr FARKAS, Attila (HUN-REN SZTAKI); Mr EMŐDI, Márk (HUN-REN SZTAKI); Mr SZABÓ, Péter (HUN-REN SZTAKI); Dr LOVAS, Robert (HUN-REN SZTAKI)

Presenter: Dr MAROSI, Attila Csaba (HUN-REN SZTAKI)

Session Classification: Demonstrations & Posters

Contribution ID: **112** Type: **Short Talk**

Leverging Federated Data Infrastructure for a European Open Web Index

Wednesday, 2 October 2024 12:00 (15 minutes)

In an era where web search serves as a cornerstone driving the global digital economy, the necessity for an impartial and transparent web index has reached unprecedented levels, not only in Europe but also worldwide. Presently, the landscape is dominated by a select few gatekeepers who provide their web search services with minimal scrutiny from the general populace. Moreover, web data has emerged as a pivotal element in the development of AI systems, particularly Large Language Models. The efficacy of these models is contingent upon both the quantity and calibre of the data available. Consequently, restricted access to web data and search capabilities severely curtails the innovation potential, particularly for smaller innovators and researchers who lack the resources to manage Petabyte Platforms.

In this talk, we present the OpenWebSearch.eu project which is currently developing the core of a European Open Web Index (OWI) as a basis for a new Internet Search in Europe. We mainly focus on the setup of a Federated Data Infrastructure leveraging geographically distributed data and compute resources at top-tier supercomputing centres across Europe. We then detail the use of the LEXIS platform to orchestrate and automate the execution of complex preprocessing and indexing of crawled data at each of the centres. We finally present the effort to adhere to the FAIR data principles and to make the data available to the general public.

Topic

Data innovations: Data Analytics, Sensitive Data/FAIR Data

Primary authors: Prof. GRANITZER, Michael (University of Passau); Mr HAYEK, Mohamad (Leibniz Supercomputing Centre)

Co-authors: Dr WAGNER, Andreas (CERN); Dr GOLASOWSKI, Martin (VSB - Technical University of Ostrava); Dr SHARIKADZE, Megi (Leibniz Supercomputing Centre); Mr DINZINGER, Michael (University of Passau); Ms FATHIMA, Noor Afshan (CERN); Mr ZERHOUDI, Saber (University of Passau); Mr MOIRAS, Stavros (CERN); Dr HACHINGER, Stephan (Leibniz Supercomputing Centre)

Presenter: Mr HAYEK, Mohamad (Leibniz Supercomputing Centre)

Session Classification: Maximizing Data Efficiency and Availablity: Innovations in Data Management Platforms

Contribution ID: **113** Type: **Long Talk**

Perun Authentication and Authorization Infrastructure

Thursday, 3 October 2024 09:20 (20 minutes)

To conduct research and foster innovation, collaboration and resource sharing have become the primary focus for research communities and national e-infrastructures. It can reduce duplication of work, leading to reduced costs, while opening possibilities for achieving common goals by combining data, skills, and efforts. However, offering such functionalities brings complex challenges to building an environment that is secure and easy to administer. That's where the authentication and authorization infrastructure (AAI) steps forward.

At CESNET, we are continuously addressing those needs by implementing Perun AAI, which is actively used as a fundamental component for large communities such as the LifeScience cluster on the European Open Science Cloud (EOSC) level and the Czech national e-infrastructure. In terms of collaboration, we create an environment that enables easy integration of services users can access by logging through their own organization accounts without forcing them to remember extra usernames and passwords. Our primary focus is also to protect users'resources by enabling multi-factor authentication, anti-phishing protection, and advanced authorization mechanisms i.e., GA4GH passports that provide a convenient and standardized way of communicating users'data access authorizations based on either their role (e.g. being a researcher), affiliation, or access status. Last but not least, we put our efforts into coming up with automatic processes, such as user life cycle within the organization, together with the automatic provisioning and de-provisioning of accounts without manual interference.

This talk will introduce the benefits Perun AAI can bring to research communities and national e-infrastructures to help them foster collaboration, improve the security of all operations, and lower administration costs.

Topic

Trust and Security: Access control

Primary author: BALCIRAK, Peter (CESNET)

Presenter: BALCIRAK, Peter (CESNET)

Session Classification: Trust & Security

Contribution ID: **114** Type: **Poster**

Scientific dataset management system for the research institute based on Onedata

Tuesday, 1 October 2024 18:00 (1 hour)

As the volume and complexity of scientific data continue to grow, the efficient management of data across its entire lifecycle has become paramount. In this context, we have decided to create a system for CEITEC Research Institute, which would allow emerging data sets to be registered and managed, using the existing Onedata system as the data layer.

At its core, Onedata oversees the entire data lifecycle, commencing with the acquisition of data from various connected instruments (cryo-EM, NMR, light microscopy) at the moment of data generation. The automated processes employed by the system enable the organisation of acquired data into coherent datasets, enriched with metadata harvested directly from the instruments themselves and the execution of workflows designed to generate data-aware metadata annotations where feasible, in accordance with defined metadata schemas established in specific fields. This facilitates the creation of FAIR datasets which are ready for publication in thematic data repositories, as and when required.

The ability to integrate heterogeneous storage capacity with heterogeneous high-performance computing (HPC) platforms, such as Jupyter notebooks and Kubernetes container clouds, is a significant advantage. By facilitating the connection between storage capacity and direct access to compute resources, Onedata enables access to compute resources for data analysis, thereby accelerating scientific discovery.

Finally, the ability to share live data via Onedata enables data sharing within and beyond the research group. Once the analysis has been completed, the system is prepared to allow scientists to easily complete and publish the final dataset to the thematic data repositories.

The objective of this poster is to illustrate the development of tools that will facilitate and streamline data sharing among scientific communities at the national and international levels. These tools are intended to support the principles of FAIR and Open Science.

Topic

Data innovations: Data Management/Integration/Exchange

Primary authors: Mr SVOBODA, Tomáš (Masaryk University); Mr ROSINEC, Adrian (CESNET); Dr RACEK, Tomas (Masaryk University); Mr HANDL, Josef (Masaryk University); KRENEK, Ales (CESNET); Mrs SVOBODOVA, Radka (Masaryk University)

Presenters: Mr SVOBODA, Tomáš (Masaryk University); Mr ROSINEC, Adrian (CESNET)

Session Classification: Demonstrations & Posters

Contribution ID: **115** Type: **Poster**

EGI Cloud Container Compute 101

Tuesday, 1 October 2024 18:00 (1 hour)

This poster offers a straightforward, step-by-step approach to leveraging Kubernetes for scientific tasks. Kubernetes provides robust features for deploying and managing containerized applications across distributed environments. The guide begins with containerizing scientific computations and proceeds to prepare for deployment by configuring essentials such as pods, deployments, services, ingresses, storage mounts, and secrets. It also covers setting up kubectl and authentication to the Kubernetes API using Rancher. The EGI offers container compute service based on the Rancher, for the seamless usage of Kubernetes in scientific computing projects. By following this guide, researchers can get familiar with basic concepts of container computing.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary authors: Mr ROSINEC, Adrian (CESNET); HEJTMÁNEK, Lukáš (Institute of Computer Science, Masaryk University)

Co-author: LUNA VALERO, Sebastian

Presenters: Mr ROSINEC, Adrian (CESNET); MORAVCOVA, Klara

Session Classification: Demonstrations & Posters

Contribution ID: **116** Type: **Demonstrations & Tutorials**

OiPub: Research Communities Made Simple

Wednesday, 2 October 2024 12:30 (30 minutes)

Keeping a research community updated with all the most relevant and impactful research and information is a never-ending task. With over 4 million articles published in 2021, growing rapidly at over 5% per year1, it's hard for anyone to keep up with a given topic.

Other academic sharing and networking platforms rely on users to add and share papers within specific groups, placing a heavy burden on them to maintain relevance and currency. This results in incomplete and messy data, including grey literature and errors, alongside peer-reviewed research. Community curation demands significant time and effort, often leading to outdated or inactive communities. Additionally, fragmented information sharing limits interdisciplinary discussion and collaboration. Finally, impact scores are publication-based rather than community-specific, hindering the recognition of relevant impactful research.

The result is that online research communities are currently much more limited in scope than they could be and often not good information sources, limiting online collaboration and making finding and following research areas much harder than necessary.

This is why we have developed OiPub - to discover and share cutting-edge research effortlessly.

OiPub automatically tracks all the latest research from recognised respected sources (e.g. CrossRef, OpenAire, ORCID, Scholexplorer) and categorises it into Topic-based information and discussion hubs. Topics can also be combined by users into custom Communities following specific interests. These are automatically kept up to date with feeds of all the latest research data relevant to them, with filtering and sorting tools allowing users to easily find the exact information they need.

This makes keeping up to date with research in any and every niche easier than ever before, allowing you to spend less time finding research and more time doing research.

In this talk we will provide a live demo of OiPub, focusing on our design concept along with insights into the data perspectives and outcomes of our work in:

1. **Topic tagging** research
2. Adapting paper-based impact scores to **Topic-based impact scores**
3. Combining Topic-based impact scores into user-customisable **Community-of-Topics based scores**
4. The **Hot score** system we are using which considers impact and recency

Omni Iota Science Limited received valuable support from EGI-ACE in developing OiPub, starting from the EOSC Digital Innovation Hub (https://eosc-dih.eu/oipub/) through the EOSC Future project, and continuing with EGI computational services and support from the EGI DIH. This along with support from other EOSC-DIH service partners including OpenAire Nexus (https://www.openaire.eu/integration-of-openaire-graph-on-oipub-sme-platform), and funding grants through the Malta Council for Science and Technology and the Malta Information Technology Agency for the development of its platform OiPub.

You can find and use OiPub at https://oipub.com/.

Topic

Data innovations: Data Analytics, Sensitive Data/FAIR Data

Primary author: BIANCHI, Robert (Omni Iota Science Ltd.)

Presenter: BIANCHI, Robert (Omni Iota Science Ltd.)

Session Classification: Demonstrations & Posters

Contribution ID: **117** Type: **Short Talk**

A Digital Twin Application: Climate Extremes Detection and Characterization using Deep Learning

Tuesday, 1 October 2024 17:25 (20 minutes)

Climate Extreme Events and their impacts are getting a lot of attention lately, because their occurrence, severity and spatial coverage are increasing and will likely increase further toward mid and end of century. Many countries are experimenting significant impact of those climate extremes. It becomes more and more important to better assess the change of characteristics of climate extremes, according to users and society needs.

However, it is not straightforward to correctly assess and quantify uncertainties. It is also a challenge to find and characterize climate extremes in all available and relevant climate simulations. This is mainly due to the very large number of simulations, along with significant data volumes. It is unfortunate to limit the number of climate simulations used in a climate change assessment study, only because of those technical and time constraints, as we should use all available information.

A novel approach and methodology is being developed to detect and characterize the changes in climate extreme events using Artificial Intelligence (AI). This is a generic method based on Convolutional Variational Autoencoders (CVAE). This deep learning technique, that uses neural networks, can process large climate datasets much faster than traditional analytical methods, and also use efficient hardware architecture like GPUs. It has the potential to better assess and quantify uncertainties associated with the various projected IPCC (Intergovernmental Panel on Climate Change) scenarios. This has been integrated in a Digital Twin Engine (DTE) architecture provided by Core Components and a Data Lake within the interTwin projects.

In this presentation, first results of the method applied on Global Coupled Climate Model datasets will be shown for several greenhouse gas scenarios, over Western Europe. A comparison to analytical methods will also be presented to assess the robustness of the method.

In summary, this DT application will enable end users to perform on-demand what-if scenarios in order to better evaluate the impact of climate change on several real-world applications in specific regions to better adapt and prepare the society.

This project (interTwin) has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement N°824084.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary authors: DURIF, Anne (CERFACS (CECI)); Mr PAGE, Christian (CERFACS (CECI))

Presenter: Mr PAGE, Christian (CERFACS (CECI))

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **118** Type: **Short Talk**

EUDAT, a Pan-European e-Infrastructure to support data-driven science.

As one of the largest pan European e-Infrastructures EUDAT delivers data management support through an integrated suite of services and resources that focus on scientific data management and embrace the FAIR principles. The EUDAT Collaborative Data Infrastructure (or EUDAT CDI), a membership organisation, is a network of European data and computing centres, as well as research organisations, which commit to sustain these services and to offer them to European research organisations and research infrastructures. The presentation will highlight the EUDAT infrastructure and its services and the close collaboration with a wide range of different scientific disciplines and communities involved in the development process of the services. Also, the structure of EUDAT with its members hosting the data management solutions and the continuous services development process as well as the services management framework guaranteeing smooth operations and support are explained. Finally the position of EUDAT among other large e-Infrastructures such as EGI, OpenAire, GÉANT and PRACE, its role in EOSC and the EOSC node development will be discussed.

Topic

Topic not listed

Primary authors: TESTI, Debora (CINECA); VAN WEZEL, Jos. (SURFsara); SANDEN, Mark (SURFsara BV); CARRILLO, Rob (TRUST-IT Services); APWEILER, Sander (JUELICH); LE FRANC, Yann (e-Science Data Factory)

Presenter: LE FRANC, Yann (e-Science Data Factory)

Session Classification: Closing session

Contribution ID: **119** Type: **Short Talk**

Cybersecurity, AI, Open Access, and Human Data

Wednesday, 2 October 2024 17:20 (10 minutes)

Within the framework of the European data strategy (European Commission. European Data Strategy (2020), (eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_it), the establishment of European data spaces per specific domains (e.g., the European Health Data Space - EHDS) have been proposed with the concomitant strengthening of regulations for governing cybersecurity.
The European Health Data Space aims to create a common space where individuals can easily control their electronic health data by defining directly applicable common rules and principles. It will also enable researchers and policymakers to use such electronic health data reliably and in compliance with privacy regulations.
The sectionalization of data movement spaces has led to increased legislation on cybersecurity, in addition to the GDPR, the EU Institutions have improved the regulation on artificial intelligence, the proposal for data governance regulation, and the proposed regulation on data, Directive (EU) 2016/1148 concerning the security of network and information systems (NIS Directive). The new legal framework is operating within a scenario where the principles of Open Access, Open Science, and FAIR principles are increasingly affirmed. The principles to be introduced within the European legal framework will impact the application of Open Science, Open Access, and FAIR principles. The impact of the ongoing regulatory adoption will also be significant in the context of research projects that concern human data. In particular the Genome Data Infrastructure (GDI) project which will enable access to genomic, phenotypic, and clinical data across Europe. GDI aims to establish a federated, sustainable, and secure infrastructure to access the data. It builds on the outputs of the Beyond 1 Million Genomes (B1MG) project to realize the ambition of the 1+Million Genomes (1+MG) initiative. Additionally, the ELIXIRxNextGenerationIT project for empowering the Italian Node of ELIXIR, the European Research Infrastructure for Life Science Data, has the primary goal of enhancing six platforms: Data, Compute, Tools, Interoperability, Omics, and Training, and integrating the activities of the national infrastructure dedicated to Systems Biology. This contribution aims to outline the legal framework currently being defined and verify the impact of regulations on research activities in the field of Biological Data.

Topic

Data innovations: Data Analytics, Sensitive Data/FAIR Data

Primary authors: Dr CESTARO, Alessandro (CNR Ibiom); Dr VARVARA, Angelo Sante (CNR); Prof. FOSSO, Bruno (Università degli Studi di Bari A. Moro); Dr LOGIUDICE, Claudio (CNR); Dr DE LEO, Francesca (CNR Ibiom); Prof. PESOLE, Graziano (Università degli Studi di Bari A. Moro - CNR); Dr TANGARO, Marco (CNR Ibiom); Dr CHIARA, Matteo (Università degli Studi di Milano); Dr FOGGETTI, Nadina (CNR Ibiom); Dr COX, Sharon Natasha (Uniba)

Presenter: Dr FOGGETTI, Nadina (CNR Ibiom)

Session Classification: Powering Collaboration: Technical Computing and Data Continuum Requirements

Contribution ID: **120** Type: **Short Talk**

Leveraging leadership class computing toward addressing climate and environmental challenges

Tuesday, 1 October 2024 16:15 (15 minutes)

Frontier and Summit, two of the largest supercomputers in the world, are hosted at the Oak Ridge Leadership Computing Facility (OLCF), and managed on behalf of the US Department of Energy (USDOE). They are also counted among "leadership class"systems in the world offering capability computing that accommodate modeling and simulations as well as data analytics and artificial intelligence applications at scale, not readily available at most capacity computing centers. The portfolio of recent computing projects at OLCF include kilometer scale earth system modeling, using the DOE Energy Exascale Earth System Model (E3SM) and the ECMWF Integrated Forecasting System (IFS), and the development of AI foundation models for climate and environmental applications. The presentation will summarize recent advances and highlights from computational earth and environmental sciences projects at OLCF, including: [a] global 3.5 km simulations using the DOE Simple Cloud Resolving E3SM Atmosphere Model (SCREAM); [b] the Oak Ridge Base Foundation Model for Earth System Predictability (ORBIT), a 113 billion parameter vision transformer model trained on CMIP6 simulations; and [c] two geoAI foundation models trained on large volumes of earth observation data from satellites.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary author: Dr ANANTHARAJ, Valentine (Oak Ridge National Laboratory)

Presenter: Dr ANANTHARAJ, Valentine (Oak Ridge National Laboratory)

Session Classification: Replicating and predicting complex systems with scientific Digital Twins

Contribution ID: **121** Type: **Poster**

Integrated Modeling of Energy Consumption Behavior in French Households: Combining Approaches to Align Needs and Consumption

Tuesday, 1 October 2024 18:00 (1 hour)

The energy consumption behavior of French households presents a complex puzzle, influenced by an interplay of socio-economic, environmental, and technological factors. This article introduces an innovative approach aimed at untangling this puzzle by integrating multiple methods to model energy consumption behavior. Our goal is to comprehend the gaps between needs and energy consumption, thereby providing insights for more effective energy policies and sustainable practices.

By combining behavioral models, dynamic simulation techniques, longitudinal data analyses,machine learning methods, and feedback loop models, our integrated model strives to capture the complexity of underlying factors influencing energy consumption decisions. By identifying the key determinants of gaps between needs and consumption, this approach will enable better guidance of energy policies aimed at promoting more efficient and sustainable energy utilization among French households.

The energy consumption behavior of households is a multifaceted phenomenon, shaped by various socio-economic characteristics, environmental considerations, technological advancements, and policy interventions. However, existing models often fall short in capturing the full spectrum of influences and interactions between variables. To address this gap, our integrated modeling approach synthesizes methodologies to provide a comprehensive understanding of energy consumption patterns.

Our methodology involves the use of behavioral models to elucidate the underlying motivations and decision-making processes driving energy consumption choices. Additionally, dynamic simulation techniques allow for modeling temporal dynamics and feedback loops within energy consumption systems. By incorporating longitudinal data analyses, our approach captures the evolution of energy consumption behavior over time and identifies trends and patterns that may not be apparent in cross-sectional analyses. Lastly, machine learning methods are leveraged to uncover complex relationships and nonlinear interactions between variables, enabling more accurate predictions and insights into future energy consumption trajectories.

Moreover, feedback loop models are integrated to assess how variations in environmental conditions influence household consumption behaviors and vice versa. By comparing consumption behavior forecasts with actual consumption data, we can identify the factors contributing to the gap between needs and consumption.

By combining these methodologies, our integrated modeling approach provides a framework for understanding the complexities of energy consumption behavior in French households and offers valuable insights for informing targeted energy policies and interventions.

Topic

Environmental informatics: Climate Change/Environment

Primary author: BOURGEOIS, Elisabeth (Université Savoie Mont-Blanc (USMB))

Presenter: BOURGEOIS, Elisabeth (Université Savoie Mont-Blanc (USMB))

Session Classification: Demonstrations & Posters

Contribution ID: **122** Type: **Long Talk**

SPECTRUM placeholder

Wednesday, 2 October 2024 15:00 (20 minutes)

SPECTRUM aims to deliver a Strategic Research, Innovation and Deployment Agenda (SRIDA) and a Technical Blueprint for a European compute and data continuum.

(abstract to be completed)

Topic

Needs and solutions in scientific computing:: Compute Continuum

Primary authors: ANDREOZZI, Sergio (EGI.eu); SALAZAR, Xavier (EGI)

Co-author: FRANCK, Gwen

Presenter: ANDREOZZI, Sergio (EGI.eu)

Session Classification: Roadmapping for high energy physics and radio astronomy

Contribution ID: **123** Type: **Short Talk**

SRCE ecosystem for Croatian science community

Wednesday, 2 October 2024 16:00 (10 minutes)

The University of Zagreb, University Computing Centre (SRCE) has been providing all layers of e-infrastructure for Croatian science and higher education for more than 50 years. The latest expansion achieved through the project Croatian Scientific and Educational Cloud (HR-ZOO) brought five new data centers in four major cities, significant bandwidth improvement in the national educational and research network and resources for the two new advanced ICT services –Virtual Data Centers and Advanced Computing.
The Advanced Computing service provides users with two advanced computing resources - Supekand Vrančić. Supek is a supercomputer based on HPE Cray with sustained computing power of 1.25 PFLOPS. Vrančić is a cloud computing platform based on widely used open-source platforms
OpenStack and Ceph that provides 11520 CPU cores, 16 GPUs and 57 TB of RAM.
Furthermore, SRCE provides several data services options –PUH for storage and sharing of data during education and research and digital academic archives and repositories framework DABAR for establishment and maintenance of reliable and interoperable institutional and thematic repositories.
Both services are integrated with the Advanced Computing in a sense that users can easily access store data on PUH or DABAR.
Finally, SRCE is actively participating in the development and maintenance of the Croatian Research Information System –CroRIS –the central place for reliable information about institutions, researchers, projects, equipment, publications, etc. Access to Advanced Computing resources for researchers is currently fully integrated with CroRIS, which enables advanced resource usage reporting –based on institutions, projects or funding streams, but also automatically correlating publications with resources.
The SRCE ecosystem provides users from science and higher education with a variety of features described above, together with workshops that enable speedy on-boarding, as well as expert support that prepares scientific applications on advanced computing resources as well as other tools to simplify usage of provided services. Information systems such as CroRIS glue all this together by enabling the development of rich and transparent usage reports.

Topic

Needs and solutions in scientific computing: National and scientific perspectives

Primary authors: CELJAK, Drazenko; IMAMAGIC, Emir (SRCE); UDOVČIĆ, Petra

Presenter: IMAMAGIC, Emir (SRCE)

Session Classification: National Perspectives: EGI Member Countries' Latest Developments and Future Initiatives

Contribution ID: **124** Type: **Demonstrations & Tutorials**

EOSC CC for PA , URBREATH, BeOpen

Wednesday, 2 October 2024 14:00 (30 minutes)

Topic

Presenters: FILOGRANA, Antonio; MUREDDU, Francesco (The Lisbon Council); SALAZAR, Xavier (EGI)

Session Classification: Demonstrations & Posters

Contribution ID: **125** Type: **not specified**

Landscape analysis of research infrastructure practices and needs for green computing - Survey findings

Tuesday, 1 October 2024 11:20 (20 minutes)

The GreenDIGIT project run a survey among research infrastructures to understand their status practices, plans and needs towards lowering the environmental impact of their digital services. This presentation will present the data and findings from this survey.

Session Classification: Green Computing: towards greener digital services

Contribution ID: **126** Type: **not specified**

Greening EGI - practices and plans

Tuesday, 1 October 2024 11:40 (20 minutes)

Session Classification: Green Computing: towards greener digital services

Contribution ID: **127** Type: **not specified**

Discussion

Tuesday, 1 October 2024 12:00 (30 minutes)

Session Classification: Green Computing: towards greener digital services

Contribution ID: **128** Type: **not specified**

Core services for EGI - An overview

Wednesday, 2 October 2024 16:45 (10 minutes)

Session Classification: Advancing Together: New Features and Roadmaps of the EGI Federation Core

Contribution ID: **129** Type: **not specified**

Discussion

Wednesday, 2 October 2024 17:40 (20 minutes)

Session Classification: Advancing Together: New Features and Roadmaps of the EGI Federation Core

Contribution ID: **130** Type: **Short Talk**

METROFOOD-RI

Wednesday, 2 October 2024 17:00 (10 minutes)

Topic

Presenters: ZOANI, Claudia (ENEA - Italian National Agency for new technologies, energy and sustainable economic development); Dr PRESSER, Karl (Premotec)

Session Classification: Powering Collaboration: Technical Computing and Data Continuum Requirements

Contribution ID: **131** Type: **not specified**

Discussion

Wednesday, 2 October 2024 17:40 (20 minutes)

Topic

Session Classification: Powering Collaboration: Technical Computing and Data Continuum Requirements

Contribution ID: **132** Type: **not specified**

Additional services

Wednesday, 2 October 2024 17:25 (15 minutes)

Session Classification: Advancing Together: New Features and Roadmaps of the EGI Federation Core

Contribution ID: **133** Type: **not specified**

EuroScienceGateway

Wednesday, 2 October 2024 17:35 (20 minutes)

Session Classification: Simplifying Data-Driven Science with User-Friendly Platforms and Gateways

Contribution ID: **134** Type: **not specified**

Q&A

Wednesday, 2 October 2024 17:55 (5 minutes)

Session Classification: Simplifying Data-Driven Science with User-Friendly Platforms and Gateways

Contribution ID: **135** Type: **not specified**

Reproducibility with EGI Services

Thursday, 3 October 2024 10:10 (15 minutes)

Presenters: FERNANDEZ, Enol (EGI.eu); LUNA VALERO, Sebastian

Session Classification: Reproducible Open Science: making research reliable, transparent and credible

Contribution ID: **136** Type: **not specified**

Q&A

Thursday, 3 October 2024 10:25 (10 minutes)

Session Classification: Reproducible Open Science: making research reliable, transparent and credible

Contribution ID: **137** Type: **not specified**

AI Landscape and Roadmap in EGI

Thursday, 3 October 2024 09:00 (10 minutes)

Topic

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning

Contribution ID: **138** Type: **not specified**

Advancing Research Frontiers: NBIS's Impact on Sweden's Computational Infrastructure and European Bioinformatics Collaborations

Wednesday, 2 October 2024 09:05 (20 minutes)

NBIS (National Bioinformatics Infrastructure Sweden) is one of the largest research infrastructures in Sweden. With approximately 120 multidisciplinary experts positioned across Sweden's major universities, NBIS constitutes the SciLifeLab Bioinformatics platform and represents Sweden within ELIXIR, the European infrastructure for biological information.

NBIS's team is composed of specialists in various bioinformatics fields, such as protein bioinformatics, mass spectrometry (MS), next-generation sequencing (NGS), large-scale data management, metagenomics, systems biology, biostatistics, and RNA sequencing. Committed to advancing research, NBIS delivers tailored support for numerous projects, providing sophisticated infrastructure and analytical tools for bioinformatics.

Establishing dynamic partnerships with SciLifeLab platforms, the SciLifeLab Data Centre, the National Academic Infrastructure for Supercomputing in Sweden (NAISS), and an extensive network of stakeholders, NBIS encourages collaboration and cohesion. These partnerships ensure exceptional support for researchers, enhancing bioinformatics research throughout Sweden and internationally.

As the Swedish node for ELIXIR, NBIS plays a crucial role in collaborative European initiatives, contributing to prominent projects like the Federated European Genome-phenome Archive (EGA), Genomic Data Infrastructure (GDI), BigPicture, EUCAIM, and the European Joint Programme on Rare Diseases (EJP-RD). By connecting local expertise with worldwide efforts, NBIS not only bolsters Sweden's computational capacities but also contributes to a concerted effort to manage the complexities of biological data across Europe.

Presenter: SHU, Nanjiang (KTH Royal Institute of Technology)

Session Classification: A view from the bridge: a look at innovative practices in our community

Contribution ID: **139** Type: **not specified**

TBC

Wednesday, 2 October 2024 09:30 (20 minutes)

Session Classification: A view from the bridge: a look at innovative practices in our community

Contribution ID: **140** Type: **not specified**

Unleashing Potential: High-Performance Computing, Big Data and Quantum Computing for Innovation

Wednesday, 2 October 2024 10:00 (20 minutes)

Through the National Recovery and Resilience Program (NRRP), Italy has funded the constitution of an unprecedented national infrastructure targeting digital resources and services for science and industry. Specifically, the National Center on HPC, Big Data and Quantum Computing ("ICSC") is an initiative funded with €320M to evolve existing public state-of-the-art network, data, and compute services in the Country, establish new facilities and solutions, and drive an ambitious program for fundamental as well as industrial research. In this contribution, the current state of work of ICSC will be given, exploring the instruments and collaborations that ICSC has been enacting to maximize the impact of this initiative at the national and international levels.

Topic

Presenter: SALOMONI, Davide (INFN)

Session Classification: A view from the bridge: a look at innovative practices in our community

Contribution ID: **141** Type: **not specified**

Demo pitches

Session Classification: A view from the bridge: a look at innovative practices in our community

Contribution ID: **142** Type: **not specified**

Cloud Compute: introduction

Tuesday, 1 October 2024 15:15 (10 minutes)

Presenter: FERNANDEZ, Enol (EGI.eu)

Session Classification: Cloud Compute federation and national initiatives

Contribution ID: **143** Type: **not specified**

HPC: introduction

Tuesday, 1 October 2024 16:45 (20 minutes)

Presenter: FERNANDEZ, Enol (EGI.eu)

Session Classification: Bridging the Gap: Integrating the HPC Ecosystem

Contribution ID: **144** Type: **Short Talk**

GlobalCoast Cloud: enabling equitable coastal resilience for the Future

Wednesday, 2 October 2024 17:30 (10 minutes)

The primary objective of the CoastPredict Programme is to provide decision-makers and coastal communities with integrated observing and predicting systems to manage risk in the short-term and plan for mitigation and adaptation in the longer-term context of future climate and ocean change.To accomplish the CoastPredict goals, the GlobalCoast initiative has been launched to create globally replicable solutions, standards, and applications that enhance coastal resilience. The advancement of CoastPredict innovation will be facilitated through the creation of an open and freely accessible digital platform known as the GlobalCoast Cloud - GCC. It is a cloud platform to transform the way we can improve and expand monitoring and forecasting of the global coastal ocean. By harnessing extensive data and establishing an infrastructure for cloud-based data and computing, this platform will expedite the dissemination of science-driven tools and information, making sure they are available and practical for the benefit of the public, decision- makers, coastal communities, and the research community. The GlobalCoast Cloud will hosts products and services for 125 Pilots Sites across the world ocean coastal areas.

Presenters: HOSLOP, E.; Dr COPPINI, Giovanni (CMCC); TINTORE, J.; Prof. PINARDI, Nadia (UNIBO); KOURAFALOU, V

Session Classification: Powering Collaboration: Technical Computing and Data Continuum Requirements

Contribution ID: **145** Type: **not specified**

E-infrastructure Assembly - Introduction and discussion

Thursday, 3 October 2024 14:00 (45 minutes)

Presenter: FERRARI, Tiziana (EGI.eu)

Session Classification: Closing session

Contribution ID: **146** Type: **not specified**

Conference awards and closing

Thursday, 3 October 2024 14:45 (15 minutes)

Session Classification: Closing session