Contribution ID: **117** Type: **Presentation**

Enabling Reproducible Computing on the EPOS ICS-D

Thursday, 11 October 2018 15:15 (15 minutes)

The EPOS-IP project is implementing solutions to enable user-driven reproducible computations exploiting the large wealth of data, data products and software discoverable through its Centralised Integrated Core Services (ICS)-C catalogue. The actual data is accessible through the web services that are managed by geographically distributed and interdisciplinary RIs organised in Thematic Communities called "Thematic Core Services". The variety of methodologies and interoperability requirements between data and software suggests the need for identifying and implement general use cases supported by flexible and scalable e-Science solutions. These must be integrated in the EPOS architecture with the preliminary objective of assisting the users in basic tasks, such as allocation of computational and storage resources and data-staging, incrementally accommodating more complex computational scenarios and reusable workflows.

We will present the approach envisaged for the integration of processing functionalities within the EPOS ICS portal. It will allow users to develop and execute new data-intensive methods and workflows within dedicated processing environments that are implemented as Jupyter notebooks and that are associated with contextual workspaces. Users of the EPOS ICS portal will select the data to be staged from one of their workspace, after having populated it with search results of interest obtained from the ICS catalogue.
Such service requires the data to be staged to remote computational facilities that adopts software containerisation and infrastructure orchestration technologies (Docker Swarm, Kubernetes) to dynamically allocate and prepare the needed resources. These will be heterogeneous and managed by national and European e-Infrastructures that will constitute the EPOS Distributed Integrated Core Services (ICS-D). We envisage that, beyond staging, many common operations could be encoded as configurable scientific workflows that will automatically preprocess the data before repurposing it to the researcher for further analysis, suggesting the need of a workflow as a service (WaaS) interface. Once data is staged and preprocessed, users can then define and evaluate their own methods via traditional scripting or still adopting advanced workflow technologies.

Thanks to containerisation, special attention is dedicated to portability and reproducibility of the processing environments, thereby allowing user to explicitly save, trace and access the different stages of their progress. Moreover, we will illustrate the approach for the adoption and integration of scientific workflow tools (CWL, dispel4py), that include validation and monitoring services. These are implemented on top of a provenance model and management system (S-ProvFlow), that allows the exploration of large lineage collections describing the obtained results. The system offers access to multi-layered, context-rich provenance information through interactive tools.

We will discuss the importance of the communication of such service with the EPOS ICS-C catalog and how it will contribute to produce and ultimately deliver research data that comply to the FAIR principles (Findable, Accessible, Interoperable and Reusable). The activities will be also presented in the scope of the cooperation with ongoing H2020 initiative such the newly funded project DARE (Delivering Agile Research Excellence on European e-Infrastructures).

Type of abstract

Presentation

Summary

The EPOS-IP project is enabling user-driven reproducible computations exploiting the large wealth of data and services discoverable through its Centralised Integrated Core Services. This must be integrated in the EPOS architecture with the objective of assisting the users in fundamental tasks concerning scalable allocation of resources, access to distributed data sources and management of reproducible complex methods. We will discuss an approach that tackles these challenges by combining Jupyter notebooks, software containerisation, workflows technologies and provenance.

Primary author: SPINUSO, Alessandro (KNMI)

Co-authors: Mr CARD, Chris (BGS - British Geological Survey); Dr BAILO, Daniele (INGV); Mr ROQUENCOURT, Jean-Baptiste (BRGM); Mr JONAS, Matser (KNMI); Mr SHELLEY, Wayne (BGS)

Presenter: SPINUSO, Alessandro (KNMI)

Session Classification: Thematic Services

Track Classification: Area 3. Computing and Virtual Research Environments