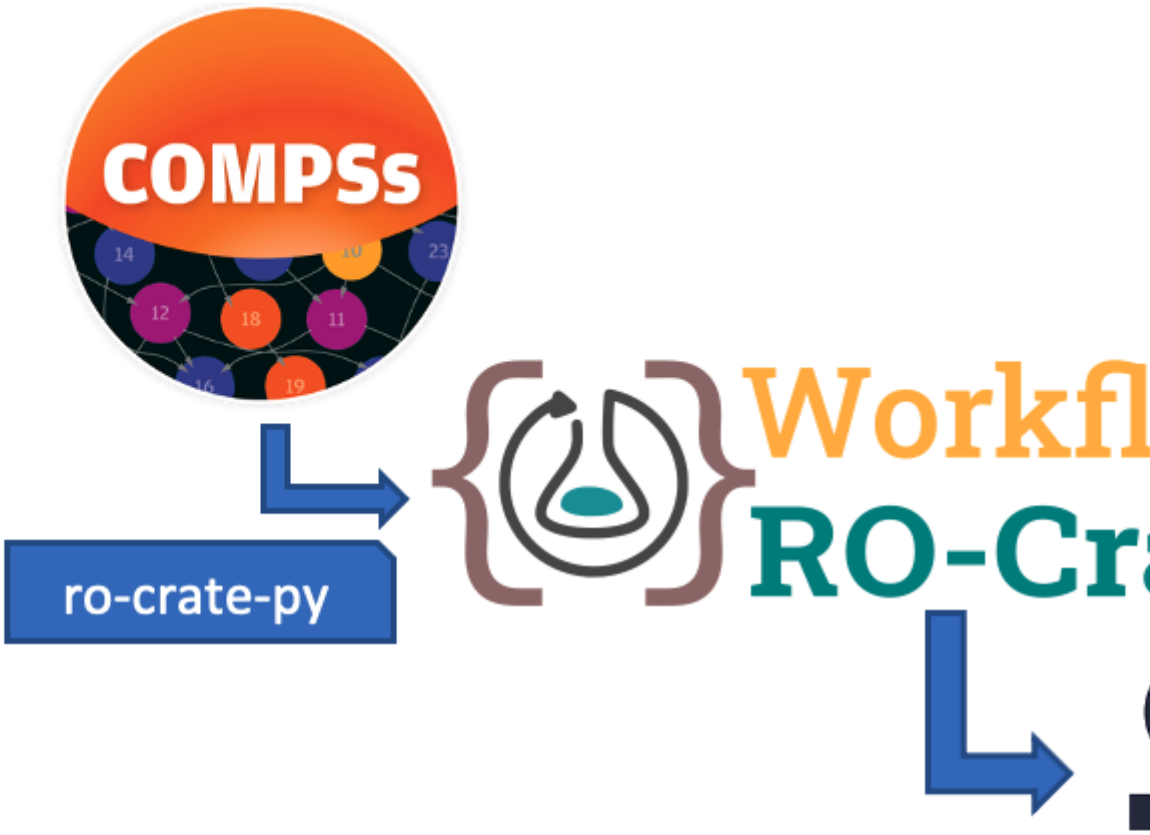
Provenance, Metadata and Reproducibility



Record metadata of experiments as provenance information, and leverage it for Governance, Reproducibility, Traceability and Knowledge Extraction.

Summary

Supercomputers and distributed systems are essential tools to help in the progress of many different scientific domains. The vast computing capacity they offer enable to run very complex simulations, as close as possible to reality, and large calculations, which often take the form of scientific workflows. As the computing capacity of these systems increases, this extra computing power is leveraged by users increasing the complexity and the time of the simulations they run, as well as the number of experiments executed. This ends up in a **growing number of generated experiments and data results**, and such a big number can be a problem when users want to keep track of **"what has been executed and where"**.

On top of that, **results need to be traceable**: users need to be able to understand how results have been generated, which systems and software have been involved in their generation, the quality and correctness of results, which individuals have created and/or run the experiments that led to these results, and more. Traceability is commonly offered by many systems by providing verification or validation tools, or even visualisation environments that enable to discover the history of steps behind a specific result generation. One example of such environments is the WorkflowHub registry (https://workflowhub.eu/), where workflows and their executions can be shared and permanent references created to **make workflows FAIR** (Findable, Accessible, Interoperable and Reusable). These kind of tools feed from metadata obtained during the execution of the experiments.

Another of today's key problems is the difficulty to reproduce research results, that some authors in the literature entitle as **Reproducibility Crisis**. A large number of experiments presented in scientific papers are not able to be reproduced by other peers in the corresponding research community domain, which diminishes their credibility: other researchers have to "trust" the results provided in a paper, rather than being able to run them by themselves and verify that the results presented are true. While a common approach is to provide a container image that facilitates the installation and deployment of software artifacts, containers are heavy to be shared, and do not cover all reproducibility cases (e.g. when a specific hardware needs to be used, when scalability results are presented for a specific machine, etc...). **Reproducibility** is gaining importance in many scientific Conferences and Journals, such as the Reproducibility Initiative at the The International Conference for High Performance Computing, Networking, Storage, and Analysis (https://sc23.supercomputing.org/program/papers/reproducibility-initiative/), where artifacts related to submitted papers are requested in order to verify their scientific claims and results.

In addition, the execution of experiments can be configured in many different ways. Experiments can use different resolutions, numerical methods can be run with different precisions, and users need to consider the trade-off between how much resources are needed and the resolution of their results (i.e. obtain results faster using less details, or use slower runs to get richer details). Also, the usage of different resource configurations during the execution of experiments can lead to scenarios where a better performance and improved use of the computing infrastructure can be achieved. Questions such as **"what is the best resource configuration for my experiment to obtain enough resolution on the results, while using a reasonable amount of resources"** are not easily answered. When the recording of metadata of experiments is enabled, and a large set of experiments and their results are available, possibilities such as the study of these different execution cases arise. First, the metadata of the experiments can be stored in a database, and later queried to know details about previous experiments (e.g. how many runs have been done, how many resources have been used for each case, what has been the execution time, ...). But also, **Knowledge Extraction** of the metadata becomes a possibility, since metadata can be analysed to find behavioural patterns across the different sets of experiments.

Metadata recording in the form of **Provenance** can help dealing with all the problems mentioned above. **Provenance** is defined as the chronology of the origin, development, ownership, location, and changes to a system or system component and its associated data. This means we need to register metadata of the experiments in order to gather all these details about them, while these details can later be used for **Governance, Traceability, Reproducibility and Knowledge Extraction**, among others.

The representation of metadata is possible by following different **standards and specifications** available in the literature. One of the main problems of metadata is to make it **interoperable** among different systems, to ensure it can be used for all the previous mentioned purposes. We intend to study and test metadata standards and specifications to understand in detail the best way to use them, as well as contribute to these initiatives with our extensive background in scientific experiments execution in both supercomputers and distributed systems.

Objectives

- Ensure Governance of experiments for any scientific domain: keep record of all their related locations (applications, third party software, results, ...), and allow their publication and sharing to their communities
    - FAIR workflows (Findable, Accessible, Interoperable, Reusable)
- Provide mechanisms for accomplishing easy and automatic Reproducibility and Replicability of experiments, so other peers can verify scientific results obtained (e.g. published in a scientific paper or elsewhere)
    - Reproducible science
- Understand the history of experiments and optimise future runs through Knowledge Extraction, by allowing the querying of metadata related to these experiments and metadata analytics through Machine Learning techniques and others
    - Data science and analytics
- Enable the building of tools that allow the Traceability of results for applications (i.e. through verification, validation and visualisation tools)
    - Explainable AI (XAI), correctness
- Study and contribute to emerging standards and specifications for metadata and its semantic interoperability across scientific domains and systems

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---