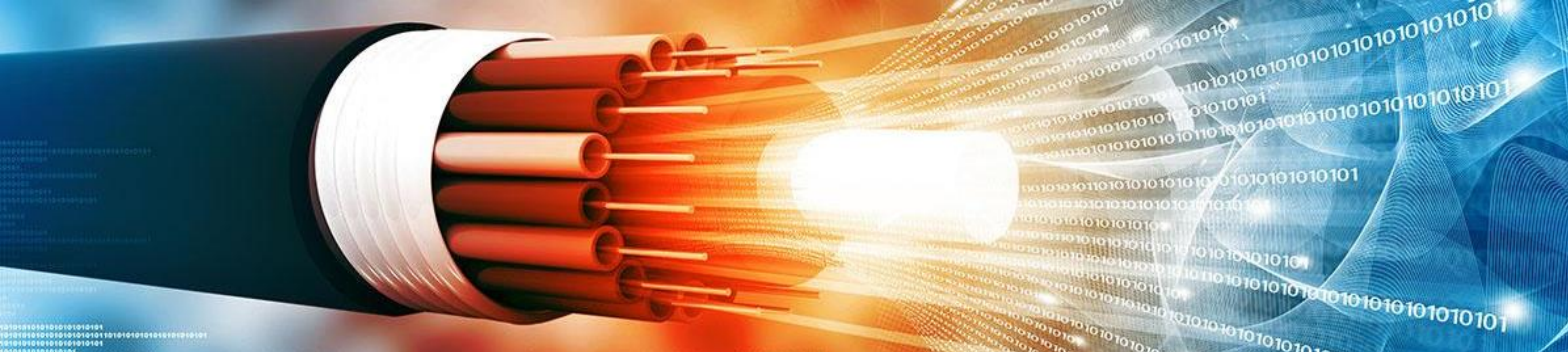


EOSC NORDIC

UNIVERSITY OF ICELAND

FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Reproducible Cross-border (High Performance) Computing for Scientific Portals

Kessy Abarenkov, Anne Fouilloux,
Helmut Neukirchen, Abdulrahman Azab

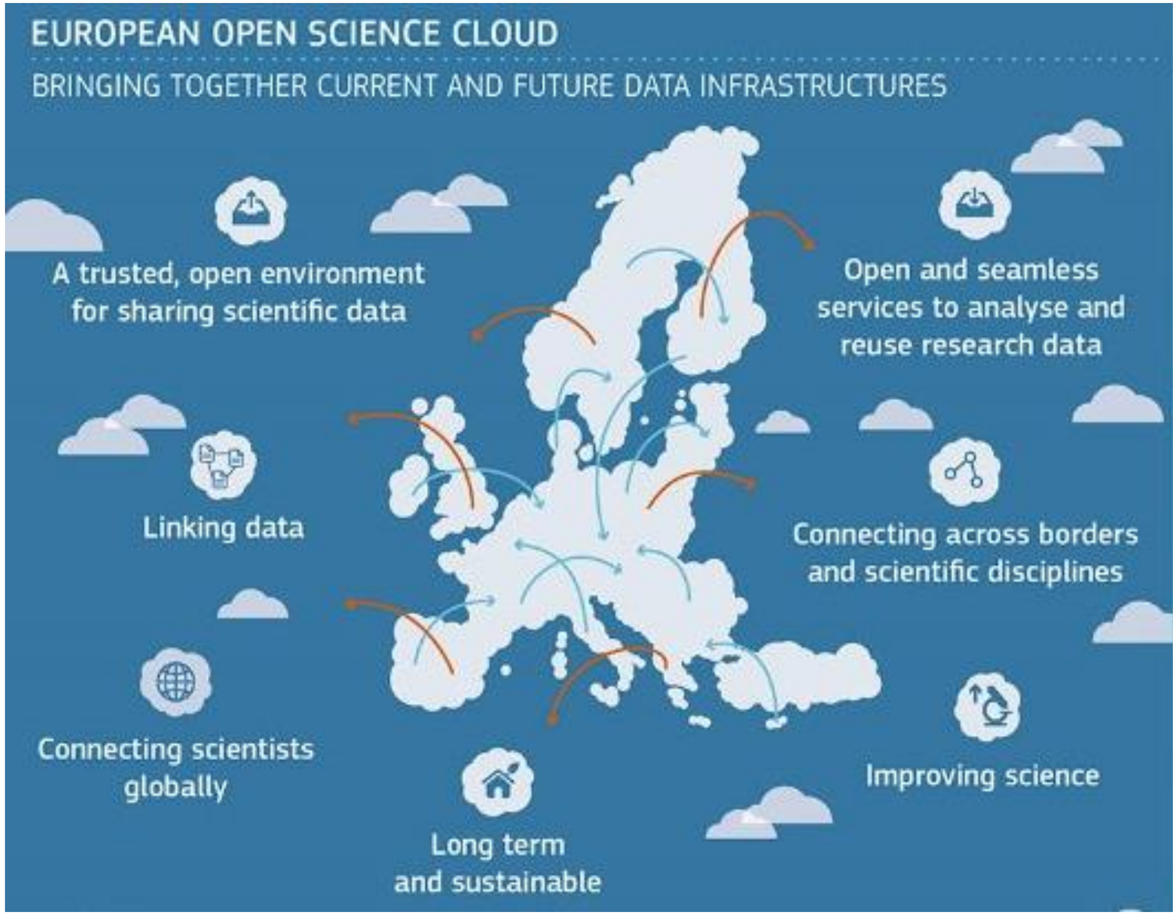
https://uni.hi.is/helmut/

The EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652.

EOSC, the European Open Science Cloud

EOSC NORDIC

- Environment for researchers to store, analyse, and re-use data.
- Providers offer services,
- Researchers use services,
- Distributed over Europe.
  - Problem, e.g.: How to give users from one institution access to a compute resource from another institution?
  - Once you have access to that remote compute resource:
    - The software that you need is likely not available there ⇒ Reproducibility.


EUROPEAN OPEN SCIENCE CLOUD
BRINGING TOGETHER CURRENT AND FUTURE DATA INFRASTRUCTURES
A trusted, open environment for sharing scientific data
Open and seamless services to analyse and reuse research data
Linking data
Connecting across borders and scientific disciplines
Connecting scientists globally
Improving science
Long term and sustainable

Source: https://ec.europa.eu/commission/news/european-open-science-cloud-becomes-reality-2018-nov-23_en

Reproducibility: Standard ingredients – season with salt and pepper

- Provide matching software versions,
  ⇒Package software as container.
- Input data needs to be available,
  ⇒FAIR data (findable, accessible, interoperable, reusable) – the DNA of EOSC.
- Automated workflows,
  ⇒Workflow engines.
- User-friendly,
  ⇒Web-based scientific portals (typically come with a workflow engine),
- To reproduce compute-intensive tasks:
  compute resources needed, maybe even specific HPC resources.
  - HPC and scientific portals are different administrative worlds:
    - Formal allocation committees vs. informal easy adding of users to scientific portals.
  - How to give scientific portal user access to HPC (or cloud infrastructure) resource?

Options to give portal user access to HPC cluster

A. Each portal user has a matching account on that cluster.
   - Tedious: need to apply for each portal user for an account on that cluster.

B. One single "robot" user account on that cluster to submit jobs of *all* portal users.
   - Quota issues: one community might consume the whole HPC resource quota.
   - Security: HPC administrators do not like an account shared between all portal users.

C. (Our proposal): Separate "robot" user account for *each* scientific community of the portal.
   - Quota shared within group that works anyway together on the same scientific problem.
   - Security concerns still apply, but the group of users gets narrowed.

- Both for options B and C, security can be improved:
  - Users never get the credentials to access HPC cluster directly – only via the portal.
  - Portal will log for each HPC cluster access who was the responsible portal user.

Case study: PlutoF portal

- PlutoF: a scientific web portal for bioinformatics (DNA sequencing):
  - Supported already workflows and to submit jobs to one specific HPC cluster:
    - Copy data to and from the HPC cluster via ssh/scp, submit HPC jobs via Slurm.
- We added for reproducibility:
  - Package software to be executed using containers, automated setup.
  - Selecting more than one HPC cluster.
  - Logging of which portal users submitted what HPC job via community robot user.
  - Support for clouds (instead of HPC only): create VMs, eased by our auto container setup.
- Tested on Swedish SNIC cloud & HPC and Estonian ETAIS HPC (national infras.):
  - PlutoF with SNIC cloud and ETAIS HPC resources: in productive use (260 users).
  - SNIC HPC did not allow our robot user, so used only a 1:1 portal/HPC users mapping.

Case study: Galaxy Climate Europe portal

- Galaxy: generic scientific web portal use by Galaxy Climate Europe community.
  - Supported already workflows, packaging, remote jobs (to cloud & HPC).
    - Nothing new to be implemented, just setting up and configuring needed.
- We added for reproducibility:
  - Packaged software using EOSC-Life methodology framework to enhance reproducibility.
  - Automated setup:
    - Add new software using GitHub pull request, create container using GitHub actions, on new compute resource: fetch resulting container from GitHub.
  - Expose remote storage resources (S3) to run jobs independently from storage location.
- Tested on Finnish CSC cloud (cPouta) & Czech CESNET cloud.
  - No HPC, because the targeted HPC clusters did not accept our robot user proposal.

Conclusions

- Improved reproducibility in two web-based scientific portals:
  - Rigorous use of container and automated setup.
  - Per-community Robot account to access cross-border cloud or HPC:
    - Enable portal users to reproduce science using remote compute resources.
    - Added logging of which portal user is actually using the resource:
      - Still: only 1 out of 3 national HPC infrastructures allowed our robot accounts.
    - Policies preventing use of robot users are political administrative problem.
      - Needs to be solved in EOSC to add compute service (in addition to data service).
      - At least the 1:1 user mapping should become less tedious:
        - » Géant (pan-European data network) introduced MyAccessID on top of eduGAIN as account for HPC access, e.g. to HPC system LUMI in Finland (#3 TOP500 list).

Thank you for your attention!

EOSC NORDIC

https://www.eosc-nordic.eu

https://twitter.com/EOSC_Nordic