Sharing costs of cross-border computing resources for beautiful climate data

Anne Fouilloux, Simula Research Laboratory, Oslo, Norway
Jean Iaquinta, University of Oslo, Scientific Computing Services, Norway
Oskar Landgren, Norwegian Meteorological Institute, Norway
Prashanth Dwarakanath, Linkoping University, Sweden
Abdulrahman Azab, University of Oslo, Norway

> **"Climate change carries no passport and knows no national borders. Countries must work toward the common interest, beyond narrow national interests."**
>
> *– Ban Ki-moon, Secretary General of the United Nations (4 November 2015)*

Summary

Researchers communicate about their achievements at conferences or in journals, and in the spirit of FAIR[A] and Open Science, they are now asked to also share their data, models, software and workflows, along with a plethora of useful additional information to exploit them (so-called metadata). In the near future, they could be rewarded (or at least recognised) for their effort. But whom do they really share it for? What do they share? For what purpose? At what cost? And for how long?

Collecting and/or producing large amounts of climate data and developing models/software comes at a price. Data is a resource, an asset for everyone, and all should be empowered to understand global change-related challenges, assess impacts, and contribute to actions. Beyond FAIR and data sharing, it is paramount to define a cost-sharing model for cross-border computing enabling anybody, regardless of their geographical location and means, to explore, process, visualise and understand climate data, from anywhere, without having to download entire datasets then install, set-up and execute complex tools to do it. This is an opportunity for the Nordic countries to become pioneers in the sharing of cross-border computing costs for a good cause, hence serving as a paragon of selflessness for others and especially for the EOSC[B].

Policy recommendation

A cost-sharing model for cross-border computing to support climate change adaptation actions.

Problem

The amount of weather and climate data continues to grow, and with the Destination Earth digital twins, this is expected to exceed 1 petabyte per day. In this context, the traditional approaches that consist in locally downloading entire datasets for processing become totally unrealistic. As an alternative, a shift towards *moving the compute closer to the data*, is gaining prominence. This involves performing data analysis or processing tasks directly on the storage or computing infrastructure where the data resides, rather than transferring it to a local machine or server. Such a strategy leverages distributed computing frameworks and cloud-

[A] Findability, Accessibility, Interoperability and Reusability
[B] European Open Science Cloud

based technologies to process data *in-situ*, thereby reducing data movements. The technical feasibility of this approach was demonstrated during the NeIC-NICEST2[C] project, and a European framework even exists that covers these needs: this is called the European Open Science Cloud or in short, the EOSC. However, moving the computer (software and tools, including an environment requirement, possibly a container) to the data raises other questions, in particular: How secure is it to run codes from *a priori* untrustworthy users? Who should pay for the processing cost? Is it the data depositor or the end-user? How many resources should be made available to ensure an acceptable level of service?

Background and context

Funded by NeIC[D], the NICEST2[E] project (2020-2023) aimed at boosting the position of the Nordic Region in the climate communities. The project focused on enhancing scientists' ability to leverage current and future computing/storage resources (e.g., EOSC, EuroHPC[F]) to perform climate simulations, analyse the results to improve model performance in the Nordic region and adopt Open Science and FAIR practices. Climate models are largely legacy codes that are difficult to master (port, install, use, develop) but encapsulating them into containers can significantly streamline the process. Similar steps can be performed from the data creation (running models on HPC, or collecting observations with edge computing) to its exploitation by different end-users (scientists, local authorities, citizens). In which case, it then becomes straightforward to move all of the compute (software) steps to the data.

Steadily more researchers are learning how to *work openly*, making their data FAIR, to share it along with their models, software and workflows, plus a wealth of metadata. However, climate data are usually large, stored in complex binary formats, and need a lot more context to be fully exploitable. The potential of what one can really do with such data does not lie in the data itself, nor in software, workflows or the metadata, rather, it is the combination of all of these *research artefacts and the links between them* that conditions what can or cannot be done with them. This matter was resolved with the introduction of the concept of FAIR ROs[G] aggregating research artefacts including data, documentation, papers, software/tools/ workflows and *machine actionable* metadata. However, *an infrastructure is required to give it life*, and it should be available for all, whether they only need to have a peek, or want to further build on FAIR ROs – this is how otherwise *ordinary* climate data becomes *beautiful climate data*.

Towards a European federated and Open Science Cloud

The EOSC aims to provide a *federated and open* environment for researchers, innovators, companies and more generally European citizens. It enables *seamless access* and reliable reuse of research data and Digital Objects (DOs) following FAIR principles. EOSC's goals are to develop a *Web of FAIR Data and services* and to promote Open Science practices. On the surface EOSC is indeed an attempt to address the issues mentioned hereinabove given that it federates several resource providers and in principle makes it possible to select the *computer closest to the data requested* by end-users. And if computing is available close to the data, and if the data in question comes as an object containing methods (tools) that can be applied to the data, then there is no need to transfer anything anymore but only return the information actually requested by the end-users, typically in the form of a table, graph, map, data summary, *etc*.

[C] 2nd phase of Nordic e-Infrastructure Collaboration for Earth System Modelling Tools (https://neic.no/nicest2)

[D] Nordic e-Infrastructure Collaboration

[E] 2nd phase of Nordic e-Infrastructure Collaboration for Earth System Modelling Tools (https://neic.no/nicest2)

[F] High Performance Computer

[G] Research Objects

Budget for the Data Processing

Should the responsibility of cost be on the users requesting data that needed to be processed or on the one who initially produced and deposited it? How should the cost be shared when the data resides in a country but is accessed and processed by a user from a different country? When it comes to cross-border computing a cost-sharing model and clear accounting rules are necessary. The EOSC has not fully addressed this issue yet, focusing instead on delivering services and onboarding new communities. However, users will only come to the EOSC if they can obtain more (or at least similar) computing and storage facilities than they already get from their own institution and/or at the national level. To produce climate data, the modelling community is dependent on being able to access the largest possible HPCs – moving to the EOSC would mean that our national providers agree to join the EOSC. But why would they do that? And what could be the benefit *versus* the cost?

Policies are therefore needed at the national level to support providers (who may lose currently tied-in customers tempted by alternative offers) in joining the EOSC, then a governance at international level to sort out the technical mechanisms (including security and authentication) and split the costs of cross-border computing and storage resources (inspired by that done by GÉANT[H]*, the collaboration of European National Research and Education Networks behind eduroam and eduGAIN).*

Benefits of cross-border cost sharing

An exhaustive list of possible savings and new opportunities would be quite long, and out of the scope of this brief, but we can definitely mention obvious *benefits of having beautiful climate data for the community*. Not transferring large amounts of data and instead having it analysed according to what those who produced it intended (i.e., no risk of variable misinterpretation, or human errors introduced in this process) will significantly:
discharge network bandwidth for other purposes; minimise the time wasted by end-users queuing and downloading;

- reduce the number of unnecessary copies of the data (and associated storage costs);
- preserve the full provenance back to the original producer of the data;
- facilitate recording data usage (for instance assess if the initial investment was worth it);
- grant anybody access to data, regardless of their location, technical means or skills;
- avoid issues by not having to write code to disentangle the data and make sense out of it;
- foster contributions from a wider range of actors who would not otherwise get any support;
- set minds free to explore innovative solutions once released from technical burden (Figure 1).

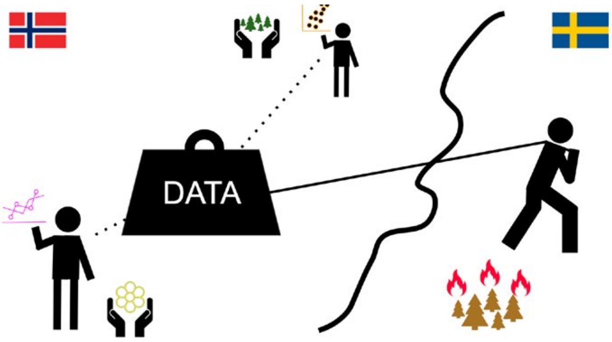

Figure 1: Symbolic issues when no cross-border cost sharing is implemented with on the left side end-users easily getting useful outputs from climate data processing carried out elsewhere and able to concentrate on mitigation actions, and on the right side another end-user wasting time and resources to procure and process by himself/herself all the data made available by a user from another country instead of thinking about actual solutions to minimise the impacts of climate change.

[H] Gigabit European Academic Network (https://geant.org)

References

NeIC website: https://neic.no/
NICEST2 website: https://neic.no/nicest2/
EOSC website: https://eosc-portal.eu/
Research Objects website: https://www.researchobject.org/

Rights and permissions

Sponsor