NICEST-2 Deliverable D4.4

Preparedness of the Earth System Modelling (ESM) research communities in the Nordics to efficiently exploit existing and new e-Infrastructures, and the role of e-Infrastructure Experts (eIEs) therein

Project manager: Anne Fouilloux
Work package leader: Jean Iaquinta
Internal reviewer: Alok Kumar Gupta

Authors: Jean Iaquinta, Anne Fouilloux and Alok Kumar Gupta.

The present document reports about a consultation that was carried out as the replacement for a physical workshop with national providers and users of High Performance Computing (HPC) resources which has never happened and was originally meant to discuss and characterise Nordics' Earth System Modelling involvement in the EuroHPC. The initial scope of the survey was modified to address several major concerns that the Nordics ESM communities consistently raised in the past few years, and that the Covid-19 pandemic may have contributed to further exacerbate.

Executive summary

Among the six topics that participants were asked to rank between critical and not relevant, the “lack of software development experts for supporting research software” came first, “unclear roles and responsibilities” [of researchers and national providers] second, and “who can help researchers select the most suited resources for their projects?” third (by a small margin). However, questions “I am happy to be contacted and further discuss my answers” and “I would like to participate in a hackathon” were completely shunned.

Clearly, the vast majority of respondents agrees that there are issues, but none of them seems to have the time and/or willingness to get personally involved in trying to resolving them. So, what can be done and who should initiate the process?

Our analysis concluded that the problems are closely linked to the *i)* existing gaps between e-infrastructure providers and user communities, *ii)* lack of HPC and data “professionals” in the climate community in the Nordics, and *iii)* lack of consensus on the roles of these data and software stewards.

A parallel can be made between e-infrastructures and “self-medication”, the whole point is to define the boundaries... Should someone (i.e., say a researcher) with a life-threatening “bug” directly go to the pharmacy and buy what he/she/they read on the internet and/or was/were led to believe is the best drug for his/her/their condition, without having ever consulted a specialist? And, should the pharmacist (i.e., resource provider) comply without ever questioning (and possibly make a juicy profit out of it) or send this “patient” to seek proper advice?

Several countries (including the US, UK, Sweden, Netherlands, Germany, Belgium, France, Australia, New Zealand, etc.) in similar situations are recruiting a new category of professionals also known as “Application experts”, “Research Software Engineers”, “Research Engineers” or “e-Science experts”, who combine a deep knowledge in particular scientific areas with expertise in method development and large-scale computational infrastructures (see what was discussed years ago in https://tinyurl.com/2p83f3s). It is also given genuine credit to work on artefacts (like software, workflows, datasets, etc.) as an integral part of the research output which is key for their career development.

Outstanding questions are typically whose responsibility such people (let’s call them e-Infrastructure Experts or eIEs for sake of convenience) should be placed under (so that they can remain as independent and impartial as possible), how to help them remain at the cutting edge of both their core scientific disciplines and IT technologies (which could be achieved by involving them as partners in research projects, and not as mere technical support or subcontractors), and finally whether this had to be replicated in every single country in the Nordics (including the least populated ones) or if it could be a cross-border collaboration.

Introduction

As part of NeIC/NICEST2 (the 2nd phase of the Nordic Collaboration on e-Infrastructures for Earth System Modeling) representatives from e-infrastructure national providers and users of their research computing and storage resources were invited to identify issues faced by their communities to efficiently exploit the existing and new e-Infrastructures, including the EuroHPC Joint Undertaking (i.e., Lumi, ranked third on the Top500 list of the world's fastest supercomputers). Based on internal discussions and on situations often experienced in everyday's work, a small number of prominent topics that were thought to deserve attention were shortlisted, and the corresponding questions were formulated to allow ranking on a scale from 1 = "Critical to discuss and address" to 5 = "Not relevant".

Additionally, participants had the opportunity to *i)* provide more information in case there were any other topics not mentioned and that they wanted to talk about, *ii)* indicate if they were happy to be contacted to further discuss their answers, *iii)* whether they would like to participate to a hackathon which was to take place either physically or online, and *iv)* add any other opinion or details they felt comfortable sharing with the survey organisers.

To begin with, the NICEST2 project's team members were asked to review the content of this survey. Based on their feedback, the introductory text for each question was reworded and drastically shortened. After a consensus was reached the team members were requested to fill it in for themselves and to distribute it to colleagues, in their institution, etc. This "5 minutes maximum questionnaire" was advertised more widely in NICEST2 April and June 2022's Info-boards (see for instance at https://tinyurl.com/2p8udan8).

The initial idea behind this consultation was to proceed with a face to face meeting, or hackathon, with representatives of all the parties involved in order to come up with a clear definition of the roles and responsibilities, "guidelines" to facilitate communications and understanding between national providers and the scientific communities, as well as a list of concrete actions to address these points in an efficient and cost-effective way.

Overall 50 people answered, most of them remained anonymous and only 4 individuals expressed interest in a hackathon. The next section of this document provides the details of how each question was answered, and what percentage of respondents considered them worth addressing. In view of the three topics that stand out after summing-up all answers obtained in the range of 1 to 3 (out of the 5 possible ratings) we go back to the rationale behind each of these questions, to replace them into context.

In a nutshell: there are outstanding issues that all parties consulted seem to be perfectly aware of, however they are not willing to be actively involved in trying to resolve or even discuss them face to face. As a result it was decided to further investigate the situation and possible reasons behind it, to try to come up with ideas about how to get out of this deadlock and initiate discussions with all the stakeholders.

Questions and raw answers (by order of importance)

Making a parallel with the "levels of alert" often used in emergency situations, the following five levels of

"importance" were defined and represented by colours:

- ***Purple*** *is the uppermost level of "risk", something that can be interpreted as highly disruptive in everyday's work and demanding immediate attention;*
- ***Red*** *also represents a very significant "threat", requiring the implementation of measures to prevent problem escalation;*
- ***Orange*** *is something to seriously investigate and clarify, although it is not perceived as a "threat" yet;*
- ***Yellow*** *can be interpreted as "vague" concerns, something to keep an eye upon because it can quickly tip-up and become an issue;*
- ***White*** *is when there is no real "danger", which is seen here as "less than 50% of those asked" found it relevant [Comment: None of the questions fell into this category].*

1. Lack of software development experts for supporting research software

There are not enough software developers to support researchers, and experts (like for GPUs) lack the domain background and/or are not receiving sufficient training to be able to work efficiently with scientists.

	1	2	3	4	5	
Critical to discuss and address	64%	20%	10%	6%	0%	*Not relevant*
		94%				

2. Unclear roles and responsibilities

There is a significant gap between the expectations of the researchers and what is actually offered by national providers (and vice versa) with no real ownership of responsibilities: who is to resolve technical issues to facilitate the completion of the scientific work and make the best possible use of the resources?

	1	2	3	4	5	
Critical to discuss and address	38%	32%	18%	10%	2%	*Not relevant*
		88%				

3. Who can help researchers select the most suited resources for their projects?

There is a wealth of compute/storage resources made available by local ITs, by national providers or at European level, but also by commercial companies: how can researchers make the "right choice" and avoid lock-in solutions?

	1	2	3	4	5	
Critical to discuss and address	27%	29%	21%	17%	6%	*Not relevant*

4. Slow pace of technologies transfer and adoption

New technologies are often made available without much guidance/documentation and meaningful context for domain specific usage. Requests sometimes also come from end-users but national providers and support staff do not have the time/resources to respond.

Critical to discuss and address	1	2	3	4	5	*Not relevant*
	32%	28%	16%	18%	6%	

5. Communication mismatch between users and support staff

There are often discrepancies between what the researchers believe is a sufficient level of detail to describe their problem and what is required by support staff to investigate and fix it. Conversely support staff replies can be perceived as gibberish, or irrelevant.

Critical to discuss and address	1	2	3	4	5	*Not relevant*
	19%	17%	35%	23%	6%	

6. Very “uneven” level of support provided

Depending on who is asking, which project and which discipline.

Critical to discuss and address	1	2	3	4	5	*Not relevant*
	2%	26%	28%	10%	34%	

Rationale behind the questions asked

1. Lack of software development experts for supporting research software

Using and developing software is part of the daily work of most scientists, however this requires knowledge outside their domain, such as efficient usage of software engineering tools, developing architecture design based on their requirements, publishing Open Source Software (https://doi.org/10.5281/zenodo.5530444). A recent survey in the US (https://doi.org/10.5281/zenodo.814220) highlighted that even though 95% of the postdocs use research software, less than half had any training in software development.
Simply providing training/teaching for scientists, as is done by the SoftwareCarpentries (https://software-carpentry.org) or CodeRefinery (https://coderefinery.org) is better than nothing, but clearly not sufficient and probably not the right approach because this is not their core activity, they are not paid to do that, and even if they significantly invested time and efforts in this area it would not be any good for their career or evaluation. This training is fundamentally aimed at making scientists somewhat aware of what is available/possible (for better use of common vocabulary, new concepts, best practises, etc.) and at allowing them to discuss with more “professional” software developers or experts.

Furthermore, there are not enough software developers to support scientists, and experts (like for GPUs) lack the domain background and/or are not receiving any training to be able to work efficiently with scientists in the Nordics.

In several countries (including the US, UK, Netherlands, Germany, Belgium, France, Australia, New Zealand, etc.) a “category” of personnel qualified as “Research Software Engineers” (RSEs) has emerged over the last few years and it is becoming increasingly recognized (https://researchsoftware.org). These people who combine professional software expertise with an understanding of research are precisely meant to serve as a link between users, scientists and researchers of all levels and e-infrastructure providers.

More than mere “application” experts these RSEs have a larger overview of the whole “computing ecosystem” (than providers’ user support teams) as well as a relevant scientific background which makes it easier for them to understand both parties, investigate issues and facilitate dialog. These RSEs also provide impartial advice (without any ambiguity as to potential interest, or what they would gain from it, since they have different employers) and/or relevant solutions. We believe that something like these RSEs is dearly missing in the Nordic countries, and that NeIC ought to advocate such a movement.

2. Unclear roles and responsibilities

There are significant discrepancies between the expectations of the researchers and what is actually offered by national providers (and vice versa). The primary job of researchers is to carry out scientific research. National providers supply compute/storage resources as well as the related e-infrastructure and domain agnostic support to researchers. However, it is not clear how national providers can support research communities and what actual means and manpower they ought to leverage (or for how long?) to efficiently fulfil this task.

In addition, there is no “ownership” of the responsibilities: who is to resolve any technical issues to facilitate the completion of the scientific work and make the best possible use of the resources, how far should support go beyond basic introduction to the services and sorting-out access or machine hardware/software issues, what about application support, can the very same staff cover all the needs of all the users and effectively contribute without in-depth knowledge of the specific applications?

Basically what is considered to be an integral part of the research work, in particular when it comes to research software and software stacks as well as to usage of the storage and computing resources (including management of workflows and data/metadata), and what is the user support offered by national providers supposed to help with? How far should these be intertwined? When do user demands/needs exceed “normal” support to become an “Advanced User Support” (AUS) request? Can this AUS be a substitute for postdocs, application experts or domain engineers? Should representatives from the national providers play a more active role in the research work and be seen as project partners?

This is not really clear for anybody, and misunderstandings often lead to situations where users are frustrated because they have technical problems that need resolving, but no time, skills/knowledge or experience to do it themselves, and nobody to ask. From the provider’s perspective addressing individual user’s issues is perceived as going way beyond what they are commissioned for: they provide “generic” technical support and solutions, try to remain ahead of the technology and keep it in working order (sorting out lots of things “in the background”, without it being noticed at all by end users), but they cannot allocate an engineer behind every user.

Most of the time researchers write proposals where they “promise” to deliver work which heavily relies on the availability of relevant hardware/software but without ever involving the actual providers of this hardware/software or taking their advice into consideration. On their side providers have to anticipate future usage, and invest huge amounts of money in new hardware/software years before it will effectively become operational. Doing so can be a “hit or miss”, as for example with GPUs where it was expected that researchers would be happy to “convert” their historical software (thereby misjudging the amount of work that this requires) whereas most GPU users exploit them for new AI applications (without bothering much about what there is underneath).

3. Who can help researchers select the most suited resources for their projects?

There are lots of hardware resources for compute/storage, either made available by local IT services (servers, personal computers), by universities (storage, light HPC, cloud, dedicated computers for Machine Learning or particular groups/applications), by national e-infrastructure providers (larger HPCs, national archives, toolkits and services), or at a European level (EOSC, EuroHPC/Lumi, PRACE, EGI) but also by commercial companies (Amazon, Google, Oracle, Alibaba, etc.).

When it comes to computing there is a wealth of tools available, and software stacks themselves have become complex, with dependencies for some applications numbering in the hundreds. Packaging, distributing, and administering software stacks of that scale is a complex undertaking, with esoteric compilers, hybrid hardware, and a lot of uncommon combinations (https://doi.org/10.48550/arXiv.2211.05118). Users always want to use the latest versions of software packages and libraries, when providers are concerned about optimising and maintaining this ever growing stack, using as much as possible automated procedures (to replicate installations over thousands of nodes, sometimes), keeping it safe and secure, satisfying as many users as possible for the most common applications, and not necessarily all of them.

How can researchers make the “right choice” and avoid lock-in solutions knowing that each provider may push their own "products" with little or no understanding of end-users' current and future needs, whereas this can have significant impacts on effective usage (access, speed, availability, maintenance, support), efforts required to adopt and/or adapt to a particular technology, inter-interoperability (or the lack thereof), sustainability, reproducibility and costs?

From a user support perspective: most of the times users ask for this or that without explaining why or what for, and without context it is difficult to tell if the request is relevant, useful at all, or a total waste of time and effort. Sometimes it is possible to exchange and better define the actual issue, instead of jumping on a technical "solution" which may not solve anything.

A caricatural example is that of a user requesting access to a machine with huge memory to load his dataset in one go, when this approach was simply not scalable and he/she should have been better-off processing it in chunks. Another example is a research group using HPCs for both model development and execution (despite long queuing time, setbacks when due to maintenance, porting issues, etc.) whereas it is more comfortable, efficient and reproducible with containers and a laptop (or Virtual Machine) to develop/test whilst saving (expensive) HPC resources for production. A very common case is that of users genuinely believing that they are taking advantage of GPUs for their application when in fact it runs on CPUs only.

Overall, more dialog is needed to understand individual situations, assess available strategies/resources and offer the most appropriate response. There is also a large part of “interpretation” which is required, to expose the deeper roots, since problems are not in

general expressed "correctly", which obviously prevents them from being resolved appropriately.

Some elements for further discussions

The topics reaching the highest scores concern both the e-infrastructure national providers and users of research computing/storage resources, so "someone" ought to do something about them. Developing "proper solutions" to address these largely exceeds the scope of this project and should involve all stakeholders (including representatives from funding agencies, publishers, research institutes & universities, national providers, unions, etc.) because the problems are much deeper than it may seem at first sight, furthermore some of them are rather tightly connected. However we can already point out a few sensitive topics/issues and try to identify the underlaying causes to clarify the situation.

- **Gap between researchers and resource providers**

This is not new, although with all the progress that has taken place in the last decade or so, and the subsequent complexification of the landscape and widening of the offer, the gap between the scientists carrying out scientific research and those in charge of the e-Infrastructures has never been wider. Users who were still able to fully apprehend and somehow manage their computing hardware and software environment when it was still restricted to their local resources (i.e., personal computers, small clusters, etc.) are now literally facing a "jungle" with novelties continuously coming out. Not only do they have to keep-up with progress in their own scientific disciplines and deal with the related tasks or "chores" (including writing applications for funding, publishing the results of their work, sometimes teaching, knowledge transfer towards policy makers, etc.), they also have to cope with new e-Infrastructures technology developments and learn new skills.

Separation of concern is paramount: to researchers the scientific work and to e-infrastructure providers the responsibility for supplying compute/storage resources, but then who is to ensure the technical IT support and bridge the gap in practice?

- **Reward and recognition in the academic research system**

When it comes to scientists' career advancement and research production it has become clear that the system is biassed and entangled in a drift, with more focus on quantity (i.e., the number of publications) rather than their intrinsic quality in terms of relevance and/or originality of the work, innovations, reproducibility, reusability, value-for-money, etc. (as discussed in https://tinyurl.com/mf4wzc8b). Without entering a debate on the excess and misuse of citations, we can emphasise the ever growing importance of technical support in the research work, and in particular IT support. However, rare are the researchers who associate as co-authors of their publications the IT engineers without whom the work could not have been performed. At best they are mentioned, but acknowledgments is not something that can enrich a CV, like a publication, or be easily traced back.

How to move away from a system where the number of publications (whatever form they take) and citations are the only means to evaluate contributions to progress?

- **Spotlight on technical contributions**

With Open Science and the foundational principles behind FAIR -Findable, Accessible, Interoperable and Reusable- (https://www.datafairport.org) and CARE -Collective Benefit, Authority to Control, Responsibility, Ethics- (https://www.gida-global.org/care), as well as the increased use of PIDs (Permanent IDentifiers) for digital objects and ROs -Research Objects- (https://reliance.rohub.org) it will be much easier to trace the origin of artefacts and attribute the correct authorship, but also to cite them, reuse them (this is behind the concept of Reproducible Research), make derivative works building on them, etc. All these advances will eventually contribute to making the traditional process of “citations” obsolete and more difficult to abuse, fraud and infringe ethics. With traceability comes new metrics (still to be defined) which should contribute to provide a much broader view of the overall landscape and reward the relevant people. The more “something” (including datasets, software, scripts, etc.) is referred to, reproduced (giving the same results for software for instance), modified and expanded (for different purposes, with other data or hypotheses), the more all those who contributed will get recognition, and not only the head of the department.

Making more visible and rewarding technical staff contributions to the research work should lead to more attractive career paths for a new category of personnel who are not necessarily “pure” scientists but more interested in the technical aspects of science, and in particular IT.

- **Who should be in charge of what?**

On the one side, placing e-infrastructure support technical staff under the “supervision” of non-technical management (i.e., whether it is scientific or administrative lead) is not a good idea because they will not understand each other, will not “see” things from the same angle, and will not put weight on the same aspects of the work, which will be a source of disagreements and frustration. On the other side, having them as integral part of the e-infrastructure providers’ personnel may make them more inclined to promote ”their” technical solutions instead of the “best” one to address particular questions.

Really, they ought to belong to an entity as independent as possible from both users and providers of e-Infrastructures in order to deliver the most impartial advice (níncluding to help researchers select the most suited resources for their projects) and support.

- **How to remain at the cutting edge of both research disciplines and technology?**

One of the main disadvantages of being highly specialised is that it does not last forever: one needs to keep up with the new stuff, whether it is in a scientific discipline or a more technical domain. This requires continuously following recent advances, training to learn new skills and practising (in the sense of using the newly acquired knowledge so as to further gain experience). For IT support to research in particular scientific disciplines it also means to take part in actual research projects to remain up-to-date with the developments in this field, but also in the preparatory phases (proposal and description of the work) to influence the technical choices all along the project (instilling best practices, for example) instead of “suffering” from them and making-do with whatever is on hands to salvage what can be saved.

E-infrastructure support staff ought to be integral partners in projects and not only be seen as the fifth wheel of the coach, asked upon when there is a problem to fix in an emergency or after completion of the work to clean-up the mess, archive and magically make everything FAIR.

Conclusions, possible way forward and role for NeIC

This consultation was conducted with e-Infrastructure users and national providers to better understand the current situation and as an attempt to initiate discussions on several issues which remained largely overlooked. These issues were not clearly expressed into plain English words or were hidden behind more obvious problems and hence never properly identified, addressed or resolved.

Preparation for the survey required authors and NICEST2 staff members to come up with a number of questions on topics meant to lead to a clear definition of the roles and responsibilities of the various parties involved. The survey provides some kind of “guidelines” to initiate dialog between national providers and the scientific community, and we tried to interpret the answers to come up with suggestions about how best to tackle these points in an efficient and cost-effective way.

Details about the questions asked and answers obtained are given in the main body of this document. What is striking now is how most of the concerns revolve around two points, namely communication mismatches between user communities and providers of e-Infrastructures added to the absence of recognition and professional career paths for data and software stewards in the Nordics.

One simple and foolproof way to sort this out, and seamlessly solve all the issues discussed in this report, would be to create a category of personal playing the role of “e-Infrastructure Experts” (eIEs) that would be placed under the administrative umbrella of a body independent from both users and national providers.

This responsibility would typically rest with an entity like NeIC (or similar) who would be ideally placed to:

- recruit experts (from relevant scientific communities);
- supervise a cross-border roster (to minimise duplications and optimise workload);
- allocate the “e-Infrastructure Experts” (on projects where they are needed most);
- train them if/when needed (in collaboration with e-Infrastructure providers);
- manage the eIEs contingent promotions (depending on their actual contributions to Open Science) and careers (as a whole, regardless of their origin/discipline).

In the Nordics these “eIEs” would need no translator to speak with researchers or computer scientists and would facilitate communications between them, they would be involved in the most recent research projects and hence always remain up-to-speed, they would work with best-in-class e-Infrastructures and also have a full overview of the technologies available to provide the most appropriate advice to users, they would take part in training and community onboarding and thereby be seen as “beacons”, having a very good overview of the actual

community needs they could also be involved in planning and e-Infrastructure developments, etc.

Models where scientists could afford to own and maintain their own computing and storage infrastructure and dedicated software engineer teams, or where national providers could meet all the needs of all the research communities in a perfectly standardised way and with generic solutions are over. Now more than ever has come the time to save resources (both in terms of electricity and manpower), minimise re-work (building on Open Science and FAIR principles) and optimise e-Infrastructure usage (by taking advantage of the most suited technologies available to get the best possible results without errors, losses, useless transfers, repetitions, etc.).

A no-brainer to reach this endeavour, to not miss the boat, make the machinery work smoothly and by the same token to address the major concerns that research communities have consistently raised over the past few years, is to create a corps of "e-Infrastructure Experts" in the Nordics, in a similar way to what is being done in other leading countries.

Such jobs have to offer attractive career paths, to appeal to the best candidates (who should not be considered as researchers having failed in an academic environment), and they have to be consulted at all the levels of the e-Infrastructure construction and operation to realise its full potential.

Incidentally the basis for evaluation of the actual production in research, and definitions of what should be accounted for as significant contributions to the research work and achievements, would have to be revisited, to better represent modern practices and accompany the implementation of FAIR and Open Science principles.