TEST-NeIC 2019 - Nordic Infrastructure for Open Science

Tuesday 21 May 2019 - Thursday 23 May 2019

Tivoli Hotel & Congress Center

Book of Abstracts

Contents

Plenary / 12

Welcome by NeIC chairman

Corresponding Author: sp@adm.dtu.dk

Plenary / 13

A regional approach towards Open Science

Corresponding Author: arne.flaoyen@nordforsk.org

Plenary / 14

Nordic Infrastructure for Open Science

Plenary / 33

A national approach towards Open Science

Plenary / 43

Bringing Open Science in Denmark into practice

Bringing Open Science in the US into Practice / 31

Bringing Open Science in the US into Practice

Author: Christine Kirkpatrick[1]

[1] *San Diego Supercomputer Center, Univ of California San Diego*

Corresponding Author: christine@sdsc.edu

With dozens of government agencies and foundations funding research at over 200 universities and hundreds more institutes and businesses, the United States comprises a challenge to comprehensive open science offerings. Approaches by various government agencies to require and incentivize open access to data will be mentioned, as well as platforms and services that enable data sharing and discovery. Models working in the EU, are being replicated in the US and providing a basis for increased awareness and value in open data. The role and challenges of including industry or private research will also be discussed.

Open Science: Is the Research Data Alliance a help or a hindrance? / 24

Open Science: is the Research Data Alliance a help or a hindrance?

Author: Hilary Hanahoe[1]

[1] *Research Data Alliance*

Corresponding Author: hilary.hanahoe@rda-foundation.org

(Title given by organisers: Insights on Open Science & EOSC from an RDA perspective)
Within the complex, international landscape of open science and open data, the research data landscape is highly fragmented, by disciplines or by domains. When it comes to cross-disciplinary activities, the notions of "building blocks" of common data infrastructures and building specific "data bridges" are accepted metaphors to approach data complexity and enable data sharing.

The Research Data Alliance (RDA) develops solutions, specifications and best practices enabling data to be shared across barriers through focused Working Groups and Interest Groups, formed of data professionals from all around the world.

This presentation will address where the RDA community stands within the realm of open science and, specifically, the European Open Science Cloud. Does it support or just add another element of complexity?

Improved Observation Usage in Numerical Weather Prediction (iOBS) / 21

Improved Observation Usage in Numerical Weather Prediction (iOBS)

Author: Jørn Kristiansen[1]

[1] *MET Norway*

Corresponding Author: jornk@met.no

Observations from the "Internet of things"(IoT), such as intelligent cars, phones, buildings and personal weather stations (PWS), including commodity weather sensors, provide detailed information on local to hyper-local meteorological phenomena. This NordForsk infrastructure project (iOBS) will accommodate an increasing amount and diversity of observation data, and provide a system of harmonised data pooling and merging. The targeted breakthrough and measurable benefit of iOBS is the effective assimilation of diverse observations in regional high-resolution NWP models for the delivery of reliable and accurate weather forecasts and warnings for the benefit of operations, business and society. The basis will be the current operational NWP model, AROME-MetCoOp and/or the very recent addition of a nowcasting suite. At the same time, there is currently a significant and unnecessary diversity at the different National Meteorological Institutes in formats, file structures and (local) software used for observation handling and pre-processing. This fragmented data handling introduces redundancies, errors and missing observations, and the consequence is that valuable information is lost. iOBS wil therefore introduce the Scalable Acquisition and Pre-Processing system (SAPP) for a joint observations handling.

The project will enable use of high-resolution and high-frequency observations. This requires to improve, develop and implement timely quality control (QC) algorithms for a massive amount of private observations of surface pressure. To our knowledge, if successful this will be the first time private pressure observations are assimilated in an operational NWP system.

The observation data flow will be built in parallel on two future generation e-infrastructures: MET Norway's PPI and Glenna-2. PPI provides flexibility, scalability in computing data storage capacity and full end-to-end data integrity to meet modern requirements on data consistency. PPI offers the benefits of both building on existing operational solutions, run as an operational environment and act as a reference to the cloud service. Glenna-2 will make effective use of hybrid environments combining specialized HPC resources and for example container technology with the more flexible cloud delivery model. Having two e-infrastructures solutions offers redundancy and flexibility, addressing the needs and requirements of Nordic (and beyond) research and operations.

The benefits of this 2-year project include:
- Improved NWP forecast quality from increased number of observations
used in data assimilation
- Improved QC algorithms for pre-processing
private observations
- Reduced cost for software maintenance and development
- Improved conditions for Nordic research collaboration on both
novel technologies and handling of different observation types
- Knowledge transfer across scientific disciplines and technological
solutions
- Redundancy and flexibility by using both a cloud
based research infrastructure (Glenna-2) and a proven operational
infrastructure (PPI)
- Raise awareness of benefits of public-private
partnerships, e.g. our QC will inform data manufacturers about
their data quality

The project partners are CSC, FMI, MET Norway and SMHI.

Conference dinner inside the Copenhagen Tivoli (registration available through NeIC2019 registration page) / 38

Conference dinner

The dinner takes place at Gemyse restaurant inside the famous and magical theme park, Tivoli Gardens. To join the dinner you must select attendance when registering to the conference. You can look forward to a delicious three-course gourmet dinner in beautiful surroundings. The required entrace ticket to Tivoli Gardens is then included.

Welcome!

Contact info:
Google maps link
Bernstorffsgade 5
1577 Copenhagen

+45 88 70 00 00
gemyse@nimb.dk

<table> <tr><td><a target="$_{b}$lank"href = "https : //indico.neic.no/event/18/images/24 − Hovedindgangen.jpg"" >< imgsrc = "https : //indico.neic.no/event/18/images/30 − Hovedindgangen.jpg""alt = "Tivoli" >< /a >< /td >< td >< atarget = "$_{b}$lank"href = "hhttps : //indico.neic.no/event/18/images/25 − Gemyse$_{R}$estaurant.jpg"" >< imgsrc = "https : //indico.neic.no/event/18/images/31−Gemyse$_{R}$estaurant.jpg""alt = "Gemyse" >< /a >< /td >< /table >

Plenary / 19

Shaping up the Nordics for EOSC

Author: Lene Krøl Andersen[1]

[1] *NeIC*

Corresponding Author: lene.krol.andersen@deic.dk

Shaping up the Nordics for EOSC is based on the work behind the EOSC-Nordic project proposal coordinated by NeIC, which was submitted to the EC during autumn 2018. *Shaping up the Nordics for EOSC* aims to facilitate the coordination of EOSC relevant initiatives within the Nordic and Baltic countries and exploit synergies to achieve greater harmonisation at policy and service provisioning across these countries, in compliance with EOSC agreed standards and practices. The project brings together a strong consortium of 24 complementary partners including e-Infrastructure providers, research performing organisations and expert networks, with national mandates and experience with regards to the provision of research data services, and a unique capacity to realise the outcomes of the EOSC design as outlined by the EOSC Implementation Roadmap.

Plenary / 20

Lifeportal

Authors: Sabry Razick[1]; Nikolay Vazov[1]

[1] *University of Oslo*

Corresponding Authors: sabry.razick@usit.uio.no, n.a.vazov@usit.uio.no

Lifeportal(lifeportal.uio.no/) is a web-based interface developed for researchers who do not have advanced computer science expertise but need to perform resource-consuming computational analyses. Lifeportal promotes open science by enabling the users to share and reuse the results of these analyses, workflows and data among their collaborators or entire workgroups within one single platform.
Lifeportal is built on the galaxy platform (galaxyproject.org) and it is customized to fit the needs of the researchers and students. The unique features of the Lifeportal are :

- HPC backend
- Open ID Connect
- Project and user management module We will show perform an analysis involving different software using the Abel HPC cluster and display results using only a web-browser. Then we will show how these data and analysis pipeline could be shared freely with other Lifeportal users, making the science reproducible and methods reusable.

Plenary / 30

Repurposing Climate Data

Author: Anne Fouilloux[1]

[1] *University of Oslo, Norway*

Corresponding Author: annefou@geo.uio.no

Advances in the development of climate models and associated data viewers and processing tools is achieving unprecedented maturity in the environmental scientific community. This was accompanied by the standardization of model output formats (conventions for Climate and Forecast metadata), the availability of open databases (i.e., the Earth System Grid Federation), and often of the

climate model codes themselves.
Applying such **FAIR** principles virtually makes it possible to re-run climate model runs or undertake other experiments. On the one side such new opportunities should attract interest from other communities such as social and human sciences. On the other side, the climate models, viewers and processing tools are generally far too complex for non-specialists and computationally demanding thus hindering cross-disciplines transfer.

In this presentation we will show how climate models can be run out-of-the-box, without much effort, using an online web platform. We will also show how climate model outputs can be visualized or how deep-learning techniques can be applied using the same web portal.

Plenary / **25**

Integrated Nordic-Baltic Genebank Information Management System

Author: Jan Svensson[1]

Co-authors: Anna Palmé [1]; Karolina Aloisi [1]

[1] *Nordic Genetic Resource Centre*

Corresponding Authors: jan.svensson@nordgen.org, anna.palme@nordgen.org, karolina.aloisi@nordgen.org

The Nordic and Baltic genebanks are responsible for conservation of plant genetic resources for food and agriculture. The e-infrastructure used by genebanks is termed Genebank Information Management System (GIMS). Implementation and development of a new Nordic Baltic integrated GIMS with functionalities that allows for incorporation of more data (phenotype/genotype) will be of great benefit for breeders and researchers using plant genetic resources. Efficient use of genetic resources is dependent on an informative database which allows for simple to complex queries, from Boolean searches to more complex queries using combined Boolean searches together with filtering for phenotypic (for example, morphology, disease resistance, yield, quality parameters) and geographic information. In the future there will also be a need to integrate genotypic (genomics) data on the collections. The aims are to fully integrate all information on clonal material from primary collections to clonal archives, develop batch tools for registration of material (including pictures, passport- and phenotype-data), deploy tools to support seed/clone health information (phytosanitary documentation), set up direct links to FAO and ITPGRFA for reporting on PGR, direct export to European (EURISCO) and Global (Genesys PGR and GBIF) databases , provide advanced viewing and filtering methods for phenotypic data, develop capabilities to integrate geographic information, increased ability for Boolean searches across more database tables, prepare for future genotypic (genomic) data on collections, and a "one-stop-shop"for researcher to find and order material from all Nordic-Baltic genebanks.

Plenary / **11**

Glenna

Workshops II / **22**

Introduction to the Rahti container cloud

Author: Risto Laurikainen[1]

[1] *CSC - IT Center for Science*

Corresponding Author: risto.laurikainen@csc.fi

CSC has a new cloud platform called Rahti. It is based on OpenShift - Red Hat's distribution of Kubernetes. It is a generic cloud platform that is suitable for a wide range of use cases from hosting web sites to scientific applications. What differentiates it from previous cloud plaforms such as cPouta is the ease with which applications can be managed, scaled up and made fault tolerant.

In this session, we will introduce the Rahti platform, tell you how to get access and show how it can be used with demos of setting up applications such as Apache Spark and Rocket.Chat.

Workshops I / 26

JupyterHub for research facilities

Authors: Thor Wikfeldt[1]; Richard Darst[2]; Radovan Bast[3]; Sabry Razick[4]

[1] *KTH/NeIC*

[2] *Aalto University*

[3] *UiT/NeIC*

[4] *UiO/NeIC*

Jupyter notebooks combine the accessibility of an interactive web-frontend, the reproducibility of a laboratory notebook, and the collaborative potential of a cloud-based deployment. The accessibility and interactivity lowers the barrier for researchers to prototype, write, and share data analysis pipelines, and the literate programming approach of Jupyter makes it particularly simple to reproduce, reuse, and adjust notebooks by colleagues and peers.

Jupyter has another use: providing access to remote resources via JupyterHub. Many typical JupyterHub deployments have used cloud-based resources for one-off purposes, but there is also good support for JupyterHub as an interface to HPC clusters and other pre-existing research facilities. JupyterHub can provide a stepping stone for light computing on existing clusters - as well as a more user friendly interface for preparation and visualization for existing power users.

In this workshop, we will demonstrate the use of JupyterHub and provide guidance so that attendees can set up their own JupyterHub deployments. There will be a show-and-tell of Jupyter itself and existing JupyterHub deployments. We will go over the basic requirements and practical implementation for a JupyterHub setup. The workshop includes discussion about the difference between traditional batch and interactive workloads, and how the parameters of HPC systems can be tuned to interactive uses. At the conclusion of the workshop, participants will be well prepared to begin deployment of JupyterHub to their own facilities and a Nordic JupyterHub community will begin.

Pre-workshop

Prerequisites: since we do not cover Jupyter itself, we will share links to talks/lessons on basic Jupyter notebooks in an updated abstract so participants can learn and experiment in advance.

Workshop outline

1. **Intro and tour of the Jupyter ecosystem** (talk, 10 min, speaker TBA): What is Jupyter and why is it taking over the world? Most of us have heard about Jupyter, but what is JupyterHub, JupyterLab, repo2docker, mybinder, etc?
2. **Tour of some cloud JupyterHub deployments** (talk + discussion, 10 min, speaker TBA)

3. **Tour of Aalto HPC JupyterHub** (focus on integration to existing infrastructure) (talk + discussion, 10 min, Richard Darst): A brief description of one existing system to frame our final perspective.
4. **JH internals from a sysadmin's point of view** (talk + discussion, 20 min, speaker TBA): JH makes a lot of sense once you know the details. This talk goes into those details.
5. **JH with batchspawner** (talk + discussion, 10 min, Richard Darst): "batchspawner" is a JH spawner for batch systems. It allows harnessing existing batch systems to run the notebooks. It is the core connector of JupyterHub and existing batch systems.
6. **JH with Kubernetes** (talk + discussion, 10 min, speaker TBA): Kubernetes is a container orchestration system, the alternative cloud-like viewpoint for large-scale deployments.
7. **Tuning HPC for interactive uses** (talk + panel, 20 min): Making JH work is relatively easy. But the typical workload of HPC clusters is quite different than that required by the interactive use of Jupyter. How should batch systems be configured so that it can co-exist with Jupyter? This is the hard part.

References

- "I don't like notebooks" presentation by Joel Grus
- SWAN: Jupyter instance at CERN: https://conferences.oreilly.com/jupyter/jup-ny/public/schedule/detail/68359

Open Science / 41

From re-useless data to Artificial Intelligence: the new frontier of the knowledge becomes a reality with Open Science.

Corresponding Author: maria.iozzi@uninett.no

Artificial intelligence (AI) / machine learning (ML) methodologies are more and more the basis of ground-breaking research. Through proper training of the algorithms, AI/ML can make direct predictions of phenomena that earlier hardly could be investigated by solving first principles problems. That is one of the reasons AI/ML workflows are becoming the driving force in many frontiers of research and innovation, from climate modelling to health care innovation, from marketing to politics and social administration. Obviously, the larger the dataset used to train the AI/ML algorithm, the more reliable the predictive potential of AI/ML methodologies becomes. Data fusion, i.e. the ability to integrate multiple, possibly large, data sources to produce more consistent, accurate, and useful information than that provided by any individual data source alone, can really enable solid predictions, when combined with suitable computing. In the era of digital data explosion and FAIR data, data driven analysis based on AI/ML methodologies can therefore be seen as one of the most promising and innovative outcomes enabled by Open Science.
This session will focus on potentials and challenges related to AI/ML, from both a scientific and a technical point of view. By learning from already existent experiences, we will try to address the fundamental question: How can scientists, innovators, technology providers and funding bodies efficiently collaborate in the Nordics in order to transform this high momentum into a solid collaborative reality?

Workshops I / 51

Test-contribution for uploading materials

Corresponding Author: hikingm@yahoo.no

Open-science-two-block / 45

test-session1

Workshops I / 29

Reimagining research computing

Authors: Radovan Bast[1]; Richard Darst[2]; Sabry Razick[3]; Thor Wikfeldt[4]

[1] *UiT/NeIC*

[2] *Aalto University*

[3] *UiO/NeIC*

[4] *KTH/NeIC*

In modern times, computation power is becoming more and more important. However, at the same time, the rest of the world is becoming consumerized: while the general expectation is that information technology is easier to use, the design of high-performance computing (HPC) systems has not kept up with modern developments in computer usability. There are many historical artifacts of how HPC systems are set up: HPC systems are often optimized for data transfer over scp, while users often prefer solutions where remote drives are mounted. We expect computations to fit into nice "rectangular"boxes of number of cores × time × memory, while with modern data science workflows, the time and memory can be unknown at the start of a job, and, in particular, interactive usage leads to highly intermittent CPU and memory requirements. Why is knowing Linux shell scripting a requirement for every job when we want our facilities to be usable by anyone? How can we empower users to have more control over their software stack?

In this workshop, we will explore the largest usability barriers in HPC systems, existing solutions, and create a joint vision of a modern HPC system. The first talks will be presentations on vision and usability from invited speakers from both HPC and human-computer interaction (HCI). After that, there will be brainstorming sessions (guided, in small groups, unconference, or panel discussions) where we identify the biggest pain points. Then, there will be group discussions in a speed-blogging format to create a shared vision document which will be the result of this workshop. After this workshop, there should be additional Nordic infrastructure cooperation to improve the accessibility, and possibly standardization, of large computational resources beyond those who traditionally use them.

"Homework": This is an interactive workshop, so please come prepared. Talk to people at your institution and/or other meeting at NeIC. Poll the people around you: what are the biggest issues with using your institution's computational facilities? Issues can be both general and specific, e.g. "all files have to manually be transferred, but due to the use of ssh proxy hosts there it is difficult from outside the campus network" or "it is easier to pay Amazon than pay us".

Workshop outline

- **HPC and accessibility** (30 min, speaker TBA): Inspiring talk about why accessibility of resources is important.
- **Strides towards accessibility at Aalto University** (10 min, Richard Darst): Demo of some of our recent ideas and problems we have seen.
- **Strides towards accessibility** at [TBA] (10 min, speaker TBA)
- **Unconference planning**, divide into layers and discuss (10 min)
- **Unconference** (30 min)
- lunch break
- **Unconference** (30 min)

- **Presentations by groups** (30 min)
- **Panel follow-up** (30 min) (panel members TBA)

Notes

- Resources first or usability first. Traditional computing center strategy is resources first, then think about how to use them. Google/Amazon strategy is make them usable, then scale up (credit: jh)
- Google/AWS cluster as a service as a baseline.
- Intro talk points: we know that there are solutions, but they are all different. Not as useful to have many different solutions. Life isn't just HPC, but it is a standard baseline.

Workshops II / 23

Introduction to Deep Learning and image classification using Convolution Neural Networks

Author: Gurvinder Singh Dahiya[1]

[1] *Uninett AS*

Corresponding Author: gurvinder.singh@uninett.no

We, at Uninett and Sigma2, have been working on enabling researcher to utilize machine learning by providing "one click"install to most used deep learning frameworks e.g. Tensorflow, Pytorch with GPUs resources to train the models. This workshop will provide participants a possibility to utilize the given platform.

The workshop will provide introduction to Deep learning with hand-on session to perform image classification using Convolution Neural Network (CNN) in PyTorch framework. We will also go through a Kaggle competition and engage in one competition for image classification task as well. The hands-on tutorial will use Jupyter Notebook running on the above-mentioned infrastructure.

Participants will gain practical experience in Deep Learning from applying state-of-art methods in image classification and ability to transfer this knowledge to their corresponding fields. Some experience with Python and Jupyter notebook can be beneficial to get most out of workshop but no prerequisite is required from Deep learning part.

Open Science / 28

Open Science with Sensitive Data

Author: Antti Pursula[1]

[1] *CSC*

Corresponding Author: antti.pursula@csc.fi

Many scientific fields are using, or would like to use, personal or sensitive data in the research. Such fields include for example genomics, health, social sciences and language research. The sensitive data that has been cleared for secondary use, should be properly managed and made findable under the same principles than non-sensitive research data. This naturally needs to be done under strict ethical

and legal compliance and via secure IT services. However, providing secure e-infrastructure for large cross-border research projects dealing with sensitive data is still in great demand and remains in some extents an unsolved challenge. Moreover, the emphasis on open science and FAIR data by the science communities and policy-makers increase the demand for professional research data management in connection to sensitive data.

This workshop will discuss the current status, opportunities and challenges of secure e-infrastructure services from various angles, and tries to form conclusions as well as inspire action for supporting open science with sensitive data. The topics for the short presentations and panel discussions are selected to highlight different sides of the topic. Topics include for example the following: openness in a sensitive data landscape, experiences from Tryggve project both from service provider and user perspective, secure processing of distributed data, impact of NeIC sensitive data activity, as well as Research Data Management and sensitive data.

The workshop programme includes short talks followed by panel discussion on a few selected topics. We plan to use online tools for one channel for addressing questions to speakers as well as enable online surveys as an option for speakers to interact with the audience.

Open-science-two-block / 46

test-session2

Open-science-two-block / 47

test-session3

Open-science-two-block / 48

test-session4

Open-science-two-block / 49

test-session5

Open-science-two-block / 50

test-session6

Social activities / 39

Social activities

Open Science / 37

FAIR

Corresponding Author: asc@kb.dk

The provision of research data in accordance with the FAIR principles could be seen as a cornerstone of Open Science and will be a major deliverable of the European Open Science Cloud (EOSC). In this session we will explore various aspects of the implementation of FAIR in the Nordic countries, covering perspectives from the researcher to the policy-maker. Topics can include the roles of the national policies compared to university and funder policies, cooperation between universities and service providers, roles of national and international networks and cooperation, outreach to user communities, scalability and long-term sustainability for FAIR data and other implementation aspects of FAIR.

Workshops II / 32

Building and Managing Linux Containers for Centralized and Distributed Systems

Author: Abdulrahman Azab Mohamed[1]

Co-author: Gurvinder Singh [2]

[1] *University of Oslo*

[2] *UNINETT*

Corresponding Authors: azab@usit.uio.no, gurvinder.singh@uninett.no

Linux containers, with the build-once-run-anywhere approach, are becoming popular among scientific communities for software packaging and sharing. Docker is the most popular and user friendly platform for running and managing Linux containers. Singularity is a platform for deploying lightweight containers for HPC systems. Kubernetes is a portable orchestration system for managing containerised workloads. This hands-on tutorial workshop will cover the following:

- Overview of the Linux containers technology
- Docker: Installation, building and managing Docker containers
- Singularity: Installation, building and running singularity containers, and creating singularity containers from Docker containers
- Containers for HPC: using Docker and Singularity containers in HPC job scripts using HTCondor
- Container Orchestration: Introduction to Swarm/Kubernetes and Hands-on

Workshops I / 40

Security in the Nordics

Corresponding Author: urpo.kaila@csc.fi

This workshop will identify common needs to share and develop joint security measures among Nordic e-infrastructures. The workshop focus on identifying requirements and solutions for security compliance to protect the infrastructures and sharing of data. The workshop will cover fields of potential joint interests, such as vulnerability management, security assessments, development of security skills, and ways to share critical information on security. The participants are also requested to contribute with suggestions for joint security initiatives.

The target audience for the workshop is security professionals, service managers and persons responsible for external relations and liaisons at Nordic e-infrastructures.

Research perspectives on AI / 35

Research perspectives on AI, transparency, privacy, law.

Closing remarks / 36

Closing remarks

Corresponding Author: gudmund.host@nordforsk.org