AIR POLLUTION ALERTS

Rapid adaptation to the genetic data deluge

Neuroscience in the information ages: Brain atlasing goes digital

Simulating potential oil spill impacts on marine ecosystems
CONTENTS

Rapid adaptation to the genetic data deluge  p: 4

Air pollution alerts  p: 8

Neuroscience in the information ages: Brain atlasing goes digital  p: 14

Simulating potential oil spill impacts on marine ecosystems  p: 18

Restructuring of Norwegian national e-infrastructure  p: 23

Language Analysis Portal  p: 26

Research data must be shared: New Policy from the Research Council  p: 28

Cover picture: from the SeaWiFS Project, NASA/Goddard Space Flight Center, and ORBIMAGE
E-infrastructure is an important tool for science in several areas. Simulation of environmental effects of human activity is one example of these areas, a topic which is important also on the political agenda. The forecasting of air pollution using the supercomputers “Stallo” and “Vilje” is described in an article in this edition of META. Such forecasting can be of great importance, especially for people vulnerable to certain types of air pollution making them able to take precautions. Another article describing the simulation of potential oil spill impacts on marine ecosystems in the Barents Sea and Lofoten ecosystems, exemplifies how e-infrastructure is being used for creating part of the foundation necessary for taking decisions whether we should combine petroleum and fisheries activities in this environmentally vulnerable areas.

The data deluge refers to the situation where the sheer volume of new data being generated is overwhelming the capacity of institutions to manage it and researchers to make use of it. Human brain atlasing is an area where we can observe such a situation developing and consequently becoming a part of the “big data” challenge. The NorStore data services are able to satisfy some of the needs created by this development, but there is also an apparent need to develop more advanced services joining HPC and “big data” applications.

The use of community portals for task support and resource access may offer a substantial better user-friendliness and contribute to the dissemination of HPC to new user groups. In an earlier edition of META we published an article explaining the use of portals in the field of life sciences. Another article about portals in the current edition, used on the language technology area shows that portals are spreading to new areas.

UNINETT Sigma will be closed down as of 31 December 2014 and a new company, UNINETT Sigma2 will be established from 1 January 2015. In this edition of META you can read about the background for this and the restructuring of Norwegian national e-infrastructure. You will also find an interview with Gunnar Bøe, who will be Managing Director in the new company.

This is the last issue of META in which I have the editorial responsibility. I would like to thank everybody that have contributed to the magazine. As I understand, META will continue also in Sigma2, and I welcome Gunnar Bøe as the new Editorial Director.
The field of genetics has been revolutionised by the recent advent of high-throughput sequencing (HTS). The Department of Medical Genetics at Oslo University Hospital (OUS) has embraced this new technology both as a research and a diagnostic tool and is today the largest producer of sequencing data in Norway. However, HTS generates very large amounts of data and this has created challenges within a hospital setting. In this article, we share the experience of our rapid transition from a low- to high-performance computing environment. We focus particularly on the storage and computational challenges that have been overcome and the scientific as well as clinical benefits that have been reaped.

**HTS - A REVOLUTIONARY TECHNOLOGY**

Deoxyribonucleic acid (DNA) is the molecule that encodes the genetic information which forms the blueprint for all living organisms. It consists of a chain of chemical bases where each base is one of four nucleotides (adenine, guanine, cytosine and thymine). Sequencing is the process whereby one reads the sequence of bases in a DNA molecule with a certain probability of error for each base call. There are several different high-throughput sequencing technology providers, the key differences between technologies being the length of sequence reads (from a few hundred to several thousand bases) that can be produced and the error rate of each base call (from about 1/100 to 1/10,000). Typically, technologies that achieve long read lengths do so at the expense of a higher error rate. What all HTS sequencers (also known as “next generation”) have in common relative to the previous generation is that they have managed to achieve a massive
parallelization of the sequencing process. The most successful parallelization is achieved on the Illumina platform where approximately 4 billion DNA molecules can be sequenced in parallel on one HiSeq machine and 250 bases of each of these molecules can be read in a six day run. Such parallelization means that a single machine can sequence a human genome (approximately 3 billion bases) to over 300-fold coverage in less than 6 days, whereas it would have taken roughly 15 years on the previous generation of machines to produce just one-fold coverage. As a result, whereas in the past human geneticists were only able to gain information about a minute fraction of the genome, now the full raw sequence data for any individual is available within a few days and at a fraction of the cost (as low as one thousand USD on the largest machines).

CHALLENGES IN DATA PRODUCTION - THE NSC’S EXPERIENCE

The Norwegian Sequencing Centre (NSC) was established in 2009 as a consolidation of the sequencing platforms at the Department of Medical Genetics and the Centre for Ecological and Evolutionary Synthesis at UiO in order to provide HTS services to the Norwegian research community (see Figure 1). The machine park has rapidly grown both in diversity as new technologies have been commercialised (454, Illumina, PacBio, Ion Torrent) and in size (e.g. eight Illumina sequencers in 2014).

At the Department of Medical Genetics, where we exclusively operate Illumina machines, this growth has lead to data storage issues. Each run of a sequencing machine produces 1.5 TB of raw sequence data which later reduces to 0.5 TB after primary analysis. Ideally, this data would get copied directly over the network to a high-performance computing facility. However, the sequencer can only tolerate losing contact with the receiving storage device for a few hours as it is constantly producing data and has limited internal storage, thus direct transfer to the UiO HPC resources is undesirable. Because of this and the added constraint that the human sequence data produced by the machines is highly sensitive, we have not been able to make use of existing HPC facilities and have had to build up a significant local closed network computation and storage capacity. In the latest upgrade, we acquired a new dedicated server (64 core 512 GB memory) and expanded storage to 128TB. Illumina has however recently released an upgrade for their machines which almost doubles the output of the machines and halved the run times. This is not atypical of the rate of innovation within the field and requires us to begin planning an imminent doubling of server capacity and tripling of storage capacity.

As a national core facility, the NSC also has to deliver data to users throughout Norway. A considerable proportion of the sequence output is from human samples. This data is defined as sensitive and therefore has to be delivered by encrypted external hard drive. For non-sensitive data, we have been able to use NorStore’s excellent storage facilities which our users can access via a web server. The deliverable files from each project are uploaded to the NorStore storage which then serves both as a backup for our non-sensitive sequencing as well as a means for transferring data to users. In 2013, 7 TB of sequenced data were delivered via NorStore, while we predict a near doubling of this number in 2014. A recent issue has been limited network capacity on the...
variant detection (DNA-seq): sequencing an individual’s genome (or exome: all regions of the genome coding for protein) and determining in what way this genome differs from the reference.

* gene expression (RNA-seq): measuring the level at which different genes are expressed.

* epigenetic state (ChIP-seq): establishing the “packing state” of DNA which is a major determinant of gene expression levels.

Researchers at the Department of Medical Genetics have a wide range of research projects covering all these application areas with variant detection being the most common, since discovering pathological genetic variants is traditionally the core expertise of a medical genetics department. The previous generation technology would only enable the sequencing of a small fraction of the genome, so a researcher trying to identify the genetic cause of a congenital disease would typically need a robust hypothesis as to which genomic regions or genes bear the causal mutation before proceeding to sequencing. Thus, a large part of the research work involved formulating this hypothesis. Access to the whole genome through HTS has meant that the need to formulate a hypothesis ex ante is no longer a requirement. Instead sequencing is performed, variants are called and patterns in the data are then exploited to identify the causal mutations (Sorte et al.). This fundamental change has enabled breakthroughs for genetic syndromes which have been the subject of decades of research (Miscio et al.).

In our department, the technology was first used in 2009 as a research tool (Selmer et al.). However, it became immediately evident that the technology would be of great value as a diagnostic tool and by 2012 many diagnostic samples had been sequenced. When used as a diagnostic as opposed to a research tool, HTS is applied in a somewhat different way: although whole exomes are sequenced, only particular genes are actually analysed. Targeting the analysis in this way has two main advantages:

* it defines the set of variants that need to be assessed for their functional effect and it enables the analyst to draw a diagnostic conclusion.

* it limits the probability of an incidental finding: finding variants with a potential pathological effect but not related to the pathology for which the patient was referred for testing.

As of today, there are eight such diagnostic packages and more than 200 patient samples have been analysed (see table 1). For the affected patient groups, this is a major improve-
ment since the hospital did not previously have an adequate diagnostic tool for these conditions, as the number of candidate genes was too great to be sequenced by older sequencing technologies.

In the near future, the aim is to perform such targeted sequencing not just to find the cause of serious congenital pathologies, but also to tailor treatments to patients ("personalised medicine", see "Nasjonal utredning av person- tilpasset medisin i helsetjenesten"). For example, for certain drugs, it is well established that individuals carrying specific mutations either do not respond to the drug or may even suffer very serious side effects from it, but it is quite rare for clinicians to make use of genetic information to determine whether a drug is suitable for a specific patient. Instead, one starts with the compound that is most effective for the majority of the population and the patient is then monitored and changed to another drug, if he suffers side effects or is a non-responder.

THE PATH AHEAD

As the clinical usage of HTS continues to grow, and with the increased yield in output per sequencer, there is a growing need for an IT-infrastructure for clinical HTS data. In light of this, and to meet the broader challenges of the new era of personalised medicine, the innovation project "Norwegian Clinical Genetic Analysis Platform" was started in 2011 (genAP). The project is a collaboration between our department, The University Center of IT (USIT), and Department of Informatics (IFi) at UiO, and aims to build an IT infrastructure for personalised medicine i.e. both the software and the infrastructure that will make it possible to go from raw sequence reads to clinically relevant genetic information. A key part of this infrastructure is Tjeneste for Sensitive Data (TSD), which has been developed by USIT in collaboration with NorStore and genAP.

TSD is a secure infrastructure that features massive storage (it currently has a 400 TB allocation from Astrastore’s 4PB disk) and an HPC computer cluster, all behind a secure firewall that requires two-factor authentication to log in. OUS IKT and our department have decided to build on the work of genAP, and are currently conducting a project to adapt procedures and workflows to use TSD for analysis of sequencing data.

Looking ahead, it is already evident that the needs of the Department of Genetics will eventually outgrow the capacity of TSD. For example, Illumina is marketing the HiSeq X Ten setup which will produce approximately 1.5 PB per year. This means that TSD can only serve as a temporary solution until OUS either builds or leases the necessary HPC capacity.

CONCLUSION

It is difficult to predict precisely what the future holds for this fast-moving field. One certainty is that it will continue to stretch the limits of current IT infrastructure but, hopefully, also continue to contribute to the implementation of solutions (such as TSD). Moreover, challenges are not limited to infrastructure: an active research area within the genAP project is the development of algorithms and interfaces that can assist laboratory personnel in understanding the functional effect of a variant and its clinical relevance. Here, building and maintaining tight collaboration between professionals with very different expertise is key. There can be no doubt that the advent of high-throughput sequencing has created a strong and growing need for many different types of informatics competence within the field of genetics.

References


Norwegian Sequencing Centre (NSC) - www.sequencing.uio.no


Norwegian Clinical Genetic Analysis Platform (genAP) - www.mn.uio.no/if/i/forseekning/prosjekter/genap

Tjeneste Sensitive Data (TSD) - www.uio.no/tjenester/it/ forskning/sensitiv/mer-om/intro.html
Air quality is not a local problem. Like clouds move through the sky, pollution is transported from one location to another by wind patterns in the atmosphere.

**INTRODUCTION**

Small amounts of potentially harmful gases and small particles directly affect our health, even if we might not be aware of it. Research shows that adverse health effects do not only result from acute events but also from chronic exposure, reducing human life expectancy by more than eight months on average and by more than two years in the most polluted cities and regions worldwide. On a larger scale, atmospheric composition represents the full state of the global atmosphere covering phenomena such as desert dust plumes, long-range transport of atmospheric pollutants and ash plumes from volcanic eruptions, but also variations and long-term changes in the background concentrations of trace components.

Observations from satellites and from ground based measurements can provide a snapshot of the air quality, but have no real predictive capability. Forecasting of air pollution is necessary for several reasons. Some people are more vulnerable to the effects of air pollution, and air pollution alerts can tell them when they are advised to take precautions. People with asthma or other lung and heart conditions are most vulnerable to the effects of air pollution.
Children are especially vulnerable, since their lungs are still growing and developing, and they spend more time outdoors, and breathe faster than adults. Older adults, and adults who work or exercise outdoors for extended periods of time, are also vulnerable. Secondly, forecasting of high air pollution events can alert politicians to take mitigation measures at the local scale. Such measures can e.g. be temporary speed limits for cars, 'low emission zones' (restricting the use of cars with high emissions), or reductions of industrial emission (e.g. through reduced production or use of cleaner fuels).

A novel forecasting system for atmospheric composition has been developed in preparation for the Copernicus Atmosphere Monitoring Service (CAMS) in a series of research projects funded by the European Union since 2005. MET has participated in all of these projects, and modified the EMEP/MSC-W state-of-the-art chemical transport model (CTM) to be able to forecast air quality over Europe. The EMEP/MSC-W model is one of 7 European CTMs that together constitute the prototype ensemble forecast system for CAMS. This system is now operated daily to forecast regional air quality over Europe. The forecasts are an important input to air quality warning systems for protecting human health.

THE FORECASTING SYSTEM
The Copernicus prototype forecasting system consists of a global model and an ensemble of regional air quality models for the European domain, which use the predictions of the global model as boundary conditions to simulate long-range transport of air pollutants. Satellite observations are fed into the global model to improve the initial conditions of the forecasts, and also to infer the latest distribution of emissions, both from anthropogenic and natural sources (e.g. forest fire and desert dust). The regional models use surface observations, as well as satellite data, to correct their initial conditions.

Each of the regional models has a different formulation of the chemical and physical processes. The spread of forecasts of the individual regional models is used as an indication of the uncertainty of the predictions, and the ensemble median has a better overall forecast skill than any of the individual models.

CHEMICAL DATA ASSIMILATION
Atmospheric trace gases and aerosols have been observed from satellites for more than two decades, but only recently have these observations been exploited for air pollution applications. Throughout the last years, a ‘data assimilation module’ has been added to the EMEP/MSC-W model.

At present, the daily air quality forecast run by the EMEP/MSC-W model contains assimilation of NO₂ from the OMI instrument (on the NASA Aura satellite), but there is also a research version including AOD (using the MODIS instrument on NASA Terra and Aqua). Within the next half year, assimilation of ozone will also be included (using the OMI or GOME-2 on METOP).
CAN LOCAL AIR POLLUTION DIMINISH BY REDUCING EMISSIONS FROM LOCAL SOURCES?

Many people believe that air pollution problems in their city are caused solely by local emissions (e.g. car traffic or industrial activity within or close to the city). The extent to which this is true strongly depends on the location of the city, its surroundings and the weather conditions. In central Europe, sub-urban areas often have high emissions and the distance between the cities and neighbouring countries are short. Within the scientific community it is well established that air pollution can travel over long distances, and thus emissions from sources far away can affect local air pollution. It has been shown that the trans-boundary component (originating from another country) of the local air pollution for PM$_{10}$ can be as large as 50% [ref 1].

Once a week, the EMEP/MSC-W chemical forecast is run for a few selected cities in Europe – switching on and off emissions either within or outside the city area. From these runs, the effect of emission reductions within or outside the city area is forecasted (see Figure 5.). For aerosols, the reductions of emissions lead to reductions in concentrations – although not to the same extent as the reductions (due to aerosols being imported from outside the city). For ozone, reductions of the precursor emissions often have the opposite effect – ozone increases. The latter is a result of so-called titration effects: when NO$_2$ concentrations are high, they tend to decrease ozone. In order to decrease ozone, the reductions have to be large enough so that NO$_2$ reaches the regime where it produces ozone.

This kind of source-receptor analyses will be further developed to include more...
European cities and to distinguish between emission categories (e.g. anthropogenic versus natural contributions, traffic versus industry, etc.). This product has been presented at several international user workshops and has received considerable interest, especially among potential policy users.

Computational Resources
Our research depends heavily on the resources provided by Notur/NorStore. For this research project, mainly the supercomputers Stallo and Vilje have been used. Most of the programming and testing has been performed on Stallo and then ported to Vilje, but both computers are still used for this project.

In 2015, the Copernicus Atmosphere Monitoring Service will be moving into an operational phase. Around the turn of the year, a number of calls for this service will be published. Up to now, the Norwegian Meteorological Institute has played a central role in the development of the project, and if the applications succeed, several developments are foreseen. One important aspect in our future work is the importance of the different emission sectors for long range transport contributions to densely populated regions. For instance, it has been proposed that wood burning and agriculture emissions played an important role in the high pollution episode in central Europe in March 2014.

Analysis of the importance of the different emission sectors from urban, domestic and transboundary emissions to regional air pollution is envisaged with the EMEP/MSC-W model. Nevertheless our most important contribution to CAMS is likely to be the air pollution forecasts, their continuous further development and maintenance, as well as periodic evaluation of the modelling system against measurements.

Figure 2. Building blocks of the Copernicus forecasting and data assimilation system for chemical weather. Observations are fed into the global and regional forecasting systems. An ensemble of seven regional air quality models (one of them being the EMEP/MSC-W model) use the global forecast as boundary conditions for air quality forecasts for the European domain.
Figure 5. Effect on ozone and aerosol concentrations (in %) from a 15% reduction of emission inside the city area, as modelled for the period 20-23 October 2014. For ozone there is a positive effect, e.g. concentrations of ozone within the Paris area would have increased if emissions in Paris had been reduced in that period. For aerosols (PM$_{10}$) there is a negative effect, e.g. aerosol concentrations within the city area would have been reduced by emission reductions.

Footnotes:
1 Copernicus is an EU-wide flagship programme that aims to support policymakers, business, and citizens with improved environmental information.
2 Data assimilation is the process by which observations are incorporated into a computer model of a real system. Observations of the current state of a system are combined with the results from a numerical model (the forecast) to produce an analysis, which is considered as ‘the best’ estimate of the current state of the system. This is called the analysis step. Essentially, the analysis step tries to balance the uncertainty in the data and in the forecast. The model is then advanced in time and its result becomes the forecast.
3 AOD [Aerosol optical depth] is a measure of the total aerosol burden in the atmosphere.

References
UNINETT Sigma invites Norwegian researchers to the 1st NorStore Research Data Management training seminar in Oslo on Wednesday 7 January 2015.

Today many research projects produce sets of digital data (data sets) from which the research findings are derived. During the project’s lifetime the data sets need to be reliably managed. Once the project has finished and the findings have been published researchers are increasingly finding value in keeping the data sets and reusing them in future projects either as a reference or to augment the new data sets. The NorStore project was established to provide a central set of services for data management that cover the entire data life-cycle.

The aim of the Research Data Management training seminar is to provide researchers with information on the available set of NorStore services for managing research data. The seminar will cover some of the challenges in data set management and will describe the NorStore services that exist, how to apply for and use those services and the services that are planned in the near future.

PRACTICAL INFORMATION:
Venue: IFI Room C (3437), Ole Johan Dahls hus, Blindern, Oslo
Address: Gaustadalléen 23B, Oslo
Date: Wednesday 7 January 2015
Time: 09h00 - 17h00

Participation is free. Please visit www.norstore.no for more information.
The human brain contains roughly 100 billion nerve cells, 3.2 million kilometres of wires, a million billion connections, all packed in a volume of 1.5 liters weighing a bit more than 1.5 kilograms, and consuming about 10 Watts of power. In order to understand the functions and dysfunctions of this marvellously complex organ, there is a need to integrate many levels of description, ranging from molecule and gene, cell and synapse, to perception, cognition and behaviour. With increasing number of advanced methods being employed and huge volumes of data being collected, neuroscience is being propelled into the information age, demanding organization and synthesis.
With a new generation of three-dimensional digital brain atlases, new solutions for integrating and disseminating brain data are being developed. In many ways, digital brain atlases are similar to current geographical atlases, such as Google Maps and Google Earth, which provide interactive access to huge amounts of high resolution image data, together with additional information (annotations, practical information, photographs) and more detailed visualizations (e.g., “street view”) for specific areas. Digital brain atlases play an important role in several large international projects, including the European Union ICT Future Emerging Technologies Flagship project, the Human Brain Project.

Understanding the brain: the reasons

From a basic science perspective, understanding the brain is a goal in itself, potentially of interest for all members of society. From a health perspective, deep knowledge about the brain and its organization and function is highly relevant, since it will impact on strategies for prevention and treatment of mental health conditions, drug abuse, degenerative diseases, and memory loss and disturbances following normal aging. With increasing life expectancy, cognitive dysfunction and dementia are emerging as significant health problems, affecting a large proportion of the ageing population. Even small reductions of brain related health problems will be of benefit for millions of individuals and families and will have significant impacts on societies. From a technological perspective, better understanding of the principles of information processing in the brain will influence new computing architectures and even the creation of brain-like computers and robotics applications. Taken together, brain sciences, brought into a context where the large volumes of data collected can be efficiently integrated, will have a major impact on individuals and societies.
ATLASES: THE EARTH AND THE BRAIN COMPARED

One approach used to organize data from the brain is to relate them to spatial locations using atlases. Traditional atlases of the structures of the brain are books or articles containing maps, typically two-dimensional diagrams of sections through a brain, with a limited amount of accompanying relevant information (Figure 1). These atlases are quite similar to conventional maps of the surface of the Earth and typically show a range of landmarks and boundaries of regions, together with names of each region. In the brain, such regions are referred to as structures in a hierarchy. The highest level in the hierarchy corresponds to the continents on the Earth, and the lowest level to small communities in a region within a country. But maps of the Earth only deal with the surface of the planet, whereas maps of the brain need to cover the full interior of the organ. In conventional brain atlases, this is solved by showing series of slices through the brain, one by one. Within the plane of each slice, microscopic images will provide information at high resolution. Across planes, resolution will be determined by several factors, including the thickness of the slices. If high cross-plane resolution is required, very thin slices need to be collected, resulting in very large numbers of slices. This data format is, needless to say, quite cumbersome to work with, and it lacks flexibility, since each series of diagrams are locked to the single plane of orientation of the slices.

The new generation multilevel digital atlases build on recent progress in brain imaging. While limited in resolution compared to microscopy, magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI) produce volumetric data with higher flexibility, allowing the image volumes to be re-sliced and viewed in arbitrary angles without loss of image quality. Atlases of structures of the brain, following the hierarchical organization as outlined above, can to advantage be produced with MRI/DTI volumes as templates, since the image volumes demonstrate a range of landmarks and structural features in the brain. Further detail is provided by aligning high resolution microscopic images from slices through the brain to the image volumes. Thus, a large range of scales of images are included (Figure 2), together with various categories of detailed information (Figure 3). The template itself is used as a basis for delineating the basic structures of the brain (Figure 4).

Atlases of the Earth make use of geographic coordinate systems that enables every location on the surface of the planet to be specified by a set of numbers. Brain atlases require the use of a 3-D coordinate system to keep track of location. A set of conventions are used to define geographic coordinate systems. In a similar way, landmarks in the brain or the surrounding skull are used to define the origin and orientation of the coordinate system applied to the brain. New standards are being applied to deal with this important aspect. Thus, brain atlases have some features in common with geographic atlases of the Earth. But due to their 3-D nature, brain atlases have different requirements and are more complicated.

TOWARDS DATA-ENRICHED BRAIN ATLASES

The current digital geographical atlases, such as Google Maps, provide valuable examples of how information can be combined, aiming at data integration and sharing. Google Maps are dynamic atlases holding diverse information, ranging from directions to distribution of restaurants, traffic, and real estate prices. The linked information is provided by organizations and individuals. We use these maps to examine information about locations and analyzing it in a manner only recently made possible with barely a second thought. Moreover, Google have provided simple access to resources and an infrastructure that allows developers to easily create their own tools that take advantage of these maps and related information. Furthermore, earth maps, also delivered through the web, are informed by satellite imagery and computer technology and provide detailed overlays of different types of information. Building on the same principal approaches, digital atlases of the brain that are currently being constructed can hold diverse information, ranging from distribution of molecules to presence of various markers, cell types and networks (Figures 2 and 3). With a common atlas framework and a set of tools to interact with the framework, research groups can connect their data to atlas space, share the data through new data systems being constructed, and search and find other relevant data through the same systems.

NAVIGATING THE BRAIN

In geographical navigation a global positioning system (GPS) is used to determine location in a map. For brain data, it is equally important to assign spatial coordinates to data originating from specific locations. It will therefore be necessary to develop digital systems for navigating across brain data and atlases (analogous to the various GPS navigation devices used for geographical maps) to study the organization of the brain, and to search for and locate various types of data that are associated with a given structure or system in the brain. Using a big science approach, this is currently done by
Retrieval based on spatial coordinates is therefore a more robust approach. For this reason, it is essential to align all data to standardized atlas templates and assign spatial coordinates to the data.

**STORING AND ACCESSING DATA FROM THE BRAIN**

A large proportion of neuroscience data consist of 2-D high-resolution images acquired with microscopy techniques and volumetric image data obtained by radiological methods, including magnetic resonance imaging and positron emission imaging. Basic requirements for managing such image data include access to large storage space, software tools for viewing and analysis, and smart approaches for querying the data. Metadata coupled to the images are typically semantic, describing names of brain structures covered by the images, and experimental conditions and other essential information. But retrieval of data based on semantic metadata will often be insufficient. Among many challenges, standardized naming conventions for all parts of the brain are difficult to arrive at, not at least since the criteria for defining brain structures are in many instances part of the research itself and will change over time.

**FUTURE PERSPECTIVES**

Digital brain atlasing requires capabilities for storing, managing, viewing, analyzing, and sharing and disseminating very large data sets. Digital brain atlasing is therefore part of the “big data” challenge. With the NorStore national infrastructure as a key resource for managing and storing data, Norway is taking part in international coordinated developments in the digital brain atlasing field. Data systems are at this stage delivered through the Rodent Brain WorkBench, hosted by the University of Oslo. New resources for atlasing in experimental research are recently published and released through portals for sharing of data and tools. Gradually more advanced systems will be developed, through the Human Brain Project international consortium, the International Neuroinformatics Coordinating Facility, and other mechanisms, enabling new modalities for collaborative brain research.

**Further reading**

HBP, The Human Brain Project, humanbrainproject.eu
INCF, International Neuroinformatics Coordinating Facility, incf.org
Allen Brain Atlas brain-map.org
Interoperable atlases of the human brain, Neuroimage 99, 525-32 (2014) doi: 10.1016/j.neuroimage.2014.06.010
Rodent Brain WorkBench, portal for sharing of rodent brain image data, rbwb.org
The Barents Sea region hosts a valuable marine ecosystem. It potentially also holds large subsea hydrocarbon deposits of great interest to the petroleum industry. As an aid to impact assessments, a new modelling system links ocean-atmosphere, oil drift and marine ecological models into a single computation framework. The system simulates possible impacts from combinations of petroleum and fisheries activities in the Barents Sea and Lofoten ecoregion. The first prototype version of this system combines today’s knowledge and modelling systems with the powerful computational capabilities of Notur’s Stallo supercomputer to bring forth more advanced ecological understanding into environmental management.
INTRODUCTION

The Barents Sea ecoregion supports a productive and diverse marine ecosystem of major significance for the wider Arctic. The region is home to a rich benthic fauna including cold-water coral reefs and sponge communities coupled closely with a productive pelagic marine ecosystem of diverse fisheries and marine mammals (Wassmann et al, 2006). The area also supports major harvesting activities, particularly for cod, haddock and capelin, and is the mainstay of major fishing investment by coastal communities from Norway and Russia. There are also significant sea-based industrial activities underway in maritime transport, offshore energy, tourism, coastal development, and aquaculture. These activities are being realized partly because of new discoveries of natural resources but also due to improved access routes and newly available infrastructure in the region.

To protect these valuable natural resources, both the Norwegian government and the European Union have adopted regulatory policies that address the management of ocean resources using an ecosystem-based approach. Ecosystem-based management (EBM) is the accepted policy instrument used to jointly address multiple policy objectives, such as biodiversity conservation, fishery production, and impact mitigation. To effectively realize these policy objectives, stakeholders recognize there is a need to improve our existing ecology based assessment methodologies with the aim of achieving more realistic predictions of potential impacts.

An advanced system to support EBM is being developed for Northern Norway’s Lofoten and Barents Sea region, a region where oil and gas activities are expanding. The system, known as SYMBIOSES, is developed by a large international consortium. The main partners include Akvaplan-niva and UiT The Arctic University of Norway (Tromsø), SINTEF (Trondheim), Institute of Marine Research (Bergen), the University of Oslo and the Norwegian Institute of Water Research (Oslo), Radboud University (Nijmegen), University of Ghent and IMARES (Wageningen UR). Utilizing the advanced computational capabilities of Norway’s supercomputer at UiT, the SYMBIOSES development team is creating an innovative system that performs integrated ecosystem-based impact assessments that addresses the potential impacts of fisheries and petroleum industry operations on fish and plankton communities in Lofoten and the Barents Sea. This new advanced impact assessment system will improve the scientific foundation for sustainable management of these important marine areas.

THE SYMBIOSES SYSTEM

The objective of the SYMBIOSES system is to be an open platform for integrated simulations of marine ecosystems. SYMBIOSES allows from the outset independent models to exchange information bidirectionally, enabling feedbacks where relevant.
The current version of the SYMBIOSES program is targeted at simulating the effects of oil on zooplankton and cod larvae in the Lofoten and Barents Sea region. Simulations are performed step-wise by computing oceanographic parameters for the region of interest. These parameters are subsequently used to advect phytoplankton and zooplankton fields (calanus finmarchicus), as well as cod eggs and larvae. The parameters are also used to advect oil concentration fields, as simulated by an oil fate model. Where dissolved oil and biology overlap, ecotoxicological models are used to alter life parameters (e.g., growth, survival, reproduction) of the various species, according to both the composition and concentration of the hydrocarbon exposure.

**DESIGN PRINCIPLES**
The SYMBIOSES program system is created by coupling existing state of the art environmental models. Leveraging existing models, tested and validated previously, allows us to substantially reduce the development efforts. The key design principle was that all coupled models should be easily replaceable, without affecting other parts of the system. This allows upgrading parts of the system with relative ease as new models become available. Furthermore, we required the method for linking models to be light-weight, and to interfere as little as possible with the internals of the individual environmental models.

Coupling independently developed (legacy) models can be a challenge. Some models are particle based, while others compute properties on grids. Models also often have different temporal resolution and behaviour. In order to achieve the necessary generality, the SYMBIOSES system focuses only on physical, observable properties, and all data transfer between models is done point-by-point. Point-wise access is done by requesting properties at a specified latitude, longitude and depth. It is then up to the framework to satisfy the request, and ensure its validity. Accessing data in a point-wise manner is computationally less efficient than directly accessing grid data. However, grids are not physical observables, and in the long run no two programs will ever agree on a global grid definition.

The individual environmental modules should be truly interchangeable. The system provides a mechanism to register and query model components using generic feature specifications, hiding all details of the underlying modules. For example, most models rely on the availability of accurate hydrodynamic parameters, and the oceanographic component can be requested using the unique feature name “hydrodynamic/ocean”. Similarly, models needing information on calanus finmarchicus can request the “aquatic/zooplankton/calanus” feature.

**ARCHITECTURE**
The SYMBIOSES system consists of a central framework library which acts like a junction box for model components. All communication between models is handled by the framework. The framework provides generic interfaces to physical model properties in points in space. The interfaces are partitioned into classes, each describing some general physical property. For example, the hydrodynamic model class provides information about velocities, temperatures, salinities and other physical oceanographic quantities. The framework provides interfaces for hydrodynamics, aquatics and atmospheric parameters as well as oil properties and chemical concentrations. The framework also acts as a plugin manager, dynamically handling model registration and feature requests.

Environmental models can be coupled to the framework either as data providers, data consumers, or both. Coupling a model provider to the system is done by implementing the appropriate abstract model interfaces and registering the corresponding features. Retrieving data from the system is then done by first requesting the relevant feature handler.
from the framework, and then using the corresponding data access procedures.

To minimize changes to coupled models, each model runs using its own, normal input and output channels as if the model was running stand-alone. The system provides methods to translate general user-defined SYMBIOSES input files to model specific inputs, transparent to the user.

OPERATIONAL SETUP
In the current configuration, the SYMBIOSES system consists of a driver program and four models, coupled through the framework library, as shown in Figure 1. The list of independent models can be found in Table 1. All models are fully 4-dimensional, data series are produced in 3 spatial dimensions \((x,y,z)\) plus time \((t)\), with the exception of the adult fish population model \(\text{GADGET}\). \(\text{GADGET}\) is a regional model that describes population changes for the entire habitat. Some models provide more than one feature to the system, i.e. \(\text{SINMOD}\) provides both oceanographic and zooplankton data. Each model comes with its own library of input data files.

The driver program is written entirely in Python. It initializes and directs computations. The initialization takes care of the dynamic loading of models, and the setup of compartmentalized data files needed by the individual models. The driver then enters the global time loop, invoking models in the right order at the appropriate times.

A range of outputs, from time series of water current fields and plankton distributions to toxicological endpoints and effects for cod larvae, are generated during a SYMBIOSES simulation. The list of parameters to save can easily be changed. The SYMBIOSES driver produces output in NetCDF-4 format, in compliance to the Climate and Forecast (CF) Metadata Conventions (www.cfconventions.org). Simulation data is saved every 12 hours.

RESOLUTION
The production configuration for SYMBIOSES runs on a spatial domain of 460 km by 1340 km in the region of the Lofoten/Barents Sea (Figure 2). The computational grid has a resolution of 4 km by 4 km, and 30 depth layers. The domain is nested into a large scale domain for the Nordic Seas (20 km), providing boundary and initial conditions for the smaller domain. The size of the domain has to be kept modest, as the computational demands of the fully coupled model is vastly higher than in the uncoupled

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
<th>Responsible/Owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSCAR©</td>
<td>Oil transport &amp; behaviour</td>
<td>SINTEF Materials &amp; Chemistry</td>
</tr>
<tr>
<td>SINMOD</td>
<td>Ocean dynamics</td>
<td>SINTEF Fisheries &amp; Aquaculture</td>
</tr>
<tr>
<td>LARMOD</td>
<td>Zooplankton ecology &amp; effects of oil</td>
<td>Institute of Marine Research</td>
</tr>
<tr>
<td>GADGET</td>
<td>Fish eggs &amp; larvae &amp; effects of oil</td>
<td>Institute of Marine Research</td>
</tr>
</tbody>
</table>

Table 1. Names and descriptions of the four pre-existing, independent models linked within the SYMBIOSES system.
case. For example, in a normal standalone SINMOD calculation, 30-40 fields are advected at each time step. In the coupled model, with 25 oil components and full toxicology, the number of variables advected is of the order of 1000. Most components of the system have been successfully parallelized, but the scaling outside a single compute node is not yet optimal. Running on 20 cores on a single compute node on Stallo, the computational efficiency is approximately 14 simulated days/day or ~1 month computational time to simulate one year. Using the aforementioned domain, the core memory requirements are around 15 gigabytes [GB]. A three month simulation produces nearly 50 GB of data per run. In the future we will work to reduce both the computational demands and the scaling, so that larger domains can be simulated faster with higher accuracy.

CONCLUSIONS AND OUTLOOK

The modelling system, known as SYMBIOSES, has been developed to perform realistic simulations of the impacts of fisheries and petroleum industry operations on fish and plankton communities in Lofoten and the Barents Sea. With SYMBIOSES, we are able to explore the impact of hypothetical oil spill scenarios in a key commercial fisheries region of the Arctic. This advanced eco-regional assessment model will lend key support to science informed decision making and stakeholder communication.

References

From January 1 2015 Norway will introduce a new model for investment in and operation of national high performance computing and storage solutions for research data.

Replacing the existing UNINETT Sigma, the new company UNINETT Sigma2 will administer the joint funding initiative between the Research Council of Norway and the four major universities. Since February 2014, a short term assigned interim-board has been working with the transition and establishment of UNINETT Sigma2. The members of the interim-board are Professor Torbjørn Dignernes from NTNU, Dean Helge Dahle from UiO and Prorector Kenneth Ruud from UiT – The Arctic University of Norway. The chairman of the interim board is the Vice Principal at UiO, Knut Fægri.

What is the background for the establishment of the new company?
The reason for the establishment of a new company is based on the analysis and recommendations given by a committee headed by Prof. Morten Dæhlen. The recommendations are based on the developments taking place in recent years in both HPC and data services. Not least, the need for data storage has increased rapidly and is expected to increase even more. In this situation, it is beneficial for the universities to be more actively involved in the strategic decision-making and financing. The result is a model in which the Research Council and the universities finance the company directly.

What will be new in UNINETT Sigma2 compared to the current e-infrastructure organisation?
Since it is a new ownership structure, it is also necessary with a new structure of the company’s board. It has been discussed what kind of company to establish, but the interim board has decided that the best solution is to organise the company as a subsidiary under UNINETT. We hope to take with us the valuable experiences from the old Sigma. I believe that the users will not notice any changes due to the changes in the structure, however I hope that the new company will be noticed due to its user-oriented and forward-looking strategy.

How does the interim board think the changes in the organisation of the e-infrastructure will influence the collaboration between UNINETT Sigma2 and the universities?
We think that the re-organisation will have positive impact in terms of dedication and better involvement of the institutions. Regarding the operational level, the well-established concept of the Metacenter will continue to be important.

The interim board was appointed in the beginning of 2014. How has the work of the interim board been so far?
The interim board has held four meetings. The first meeting was about the choice of corporate form and the establishment of the company. Later, we have spent a lot of time on the agreements between the partners. Here we encountered a number of challenges in a situation where the universities act both as funders, as service providers and as customers through the users. We have also appointed a managing director of the new company.

Why did you decide to call the company Sigma2?
We think that the “brand” Sigma is valuable and we wanted to keep the positive associations linked to the name. It gives a signal of continuity, while the number 2 points to the future and to further development.

UNINETT Sigma2

The new company will once again become a subsidiary of UNINETT and headquartered in Trondheim. UNINETT is a limited company owned by the Ministry of Education and Research, whose aim is to create world-leading ICT infrastructure and services to bolster research and higher education in Norway.

Long term funding

The Research Council will increase its annual allocation for e-infrastructure from the current NOK 17 to NOK 25 million per year for the next ten years. The universities will contribute by covering costs equivalent to NOK 50 million each year. Additional contributions may come from rewarded applications for additional infrastructure funding and from research projects being obliged to pay for some of their services.

Board members
Petter Kongsnes, UNINETT AS (Chairman)
Anne Borg, Norwegian University of Science and Technology
Morten Dæhlen, University of Oslo
Kenneth Ruud, UiT
The Arctic University of Norway
Nathalie Reuter, University of Bergen
Heidi Ekrem, Mageli Lawfirm ANS
Ulf Gyllensten, Uppsala University, Sweden
The Managing Director of UNINETT Sigma2 is Gunnar Bøe.

The new Managing Director has his university degree (Siv.ing) in Computer Science from NTH (now NTNU) in 1993. However, he started his career with computers many years earlier when he was working with British Petroleum as an assistant to geologists and geophysicist doing a lot of programming in Fortran.

- It is with some amusement I remember how I was summoned to the IT-manager’s office to explain why I was using so much CPU resources on the VAX cluster. The reason was that I was helping to do reservoir calculations for one of the oil fields in the North Sea needing a lot of CPU hours. Working with BP really gave the taste for computer science and an expat job which later led to an almost 7 years stay in China, studying Chinese and later working with Ericsson.

The work in Ericsson included both building infrastructure for mobile data and several years at Ericsson’s research and development centre in Shanghai. The taste for Asia has led to several extended stays in the region, most lately a 4 month stay at the business school of the National University of Singapore. This was part of a master degree in technology management he is currently working on at NTNU/NHH.

- Infrastructure has also been the main focus for the almost 10 years I have worked in UNINETT. Most of the work has been related network infrastructure at campuses and network related services. This has been a very rewarding time where the great cooperation with the universities and university colleges to excel the ICT infrastructure has led to great recognition on European level and beyond.

Why did you accept the position as Managing Director of Sigma2?
I have been involved in building infrastructures for the benefit of research and education during the past 10 years and have found that very rewarding. The infrastructure for HPC and data services is essential for for research in many areas. To be part of this is something I’m really looking forward to.

What are your major goals for the new e-infrastructure company?
I would like to see more research disciplines starting to use the e-infrastructure and new services developed that will help research to make new progress. Another objective is to develop an even better cooperation with the universities for support and development. My experience from the work with network infrastructures shows that through cooperation we can take our results to higher levels than as individuals and in networking that made us among the best in Europe.

What are the current plans for Sigma2 in 2015?
We are currently working hard to put all new agreements in place with the Research Council and the universities. Hopefully this will be finished in the beginning of 2015. Another important start-up activity is to get a new board up and running. A major activity in 2015 will be the pre-study for the procurement of new equipment both for HPC and data. This study will prepare a lot of the material necessary for decisions about future operation models and the number of HPC sites. Another part of this study is of course to have close contact with the research communities in order get their requirements. Part of that will hopefully also put us in touch with new communities that can make use of the e-infrastructure.
A PORTAL FOR HPC-POWERED LANGUAGE TECHNOLOGY

At the University of Oslo [UiO], researchers from the Language Technology Group at the Department of Informatics and the University Center for Information Technology (USIT) are cooperating on creating the Language Analysis Portal, or LAP for short. The goal is to couple language technology (LT) and high-performance computing (HPC) in an easily accessible web interface that eliminates technological barriers to entry for non-expert users.

Language technology is the interdisciplinary science of making computers “understand”, at some level, human language. Application examples include machine translation, dialogue systems, intelligent information retrieval, and semantic search, and the field also goes by other names like natural language processing (NLP), computational linguistics (CL), or language engineering.

With a computing portal that provides a repository of LT tools, where single jobs and complex workflows can be configured and executed using a uniform graphical interface, and with seamless integration to a HPC backbone, we hope to enable new user groups to take advantage of large-scale language technology. The development of LAP forms part of the CLARINO infrastructure project [1], the Norwegian branch of the pan-European CLARIN initiative (Common Language Resources and Technology Infrastructure).
NEW COMPUTATIONAL SCIENCES
There are at least two important scientific trends that have motivated the LAP project: language technology becoming increasingly data-driven and computational, and the emergence of new user groups of LT within other [non-technical] scientific disciplines.

First, language technology has developed to be an increasingly data-driven discipline. Machine learning and statistical analysis is by now one of the main methodological pillars of the field. Data sets have grown to a size that earns the buzzword label Big Data. At the same time, machine learning problems encountered within LT are typically extremely high-dimensional, making parameter tuning a computationally costly affair, hungry for both compute cycles and memory. For many central research tasks, having access to HPC facilities is by now a prerequisite for being able to compete in the field. This situation creates a divide in more than one respect; while not all LT researchers have access to a large-scale computing cluster infrastructure, even those that in principle do, may not possess the necessary know-how for taking advantage of it.

Second, language technology is not the only discipline to have grown more computational and data-driven in recent years, of course. A similar trend is emerging within the humanities and social sciences (HSS), as witnessed by the surge of new “prefix disciplines” like the eHumanities, digital humanities, or computational social science. As part of this trend there is a growing interest among HSS scholars to use automatic analysis of large text-based data in their research, thereby making language technology relevant to a growing number of new user groups. At the same time, there is no denying that many common LT tools can appear rather daunting to use, requiring a lot of technical knowledge on the side of the user. Apart from the challenge of orienting oneself in the fragmented ecosystem of available tools, many potential users, especially from less technically oriented disciplines, might not be comfortable with command-line interfaces or having to wrestle with difficult and poorly documented installation procedures, or might lack the required knowledge about input/output formats or other dependencies.

HPC AND GALAXY
The objective of LAP is to maintain a large repository of LT tools that are easily accessible through a web portal, offering a uniform graphical interface. Any scholar registered in the federated identity management systems Feide or eduGAIN can log into the portal and create a Galaxy user. Each user has her own personal workspace, allowing data to be stored persistently across sessions. In addition to upload and storage facilities for user-provided data, the portal will in the future also give access to common, pre-existing language resources. It will include tools for content extraction and layout analysis (from common file formats and markup schemes), as well as a comprehensive repository of language analysis tools.

A core component of LAP is Galaxy [2], an open-source web-based platform for scientific workflow management initially developed for data-intensive research in genomics and bioinformatics. In these fields, Galaxy-based portals allow biologists with no programming skills to access and configure processing tools, conduct experiments, and share both the results and the processing steps associated with them. By adapting Galaxy to the context of language technology, the vision of LAP is to ensure the same kind of access and ease of use of language technology tools for researchers from the humanities and the social sciences (but also for researchers within the field of LT itself). In our context, one particularly relevant Galaxy-based computing portal from bioinformatics is the LifePortal [3] (formerly BioPortal), also developed and maintained at the University of Oslo. The success of using Galaxy in LifePortal was one of the main reasons why we chose to also build on this platform in LAP, also allowing us to draw on the already existing local expertise (for more background, both Galaxy and LifePortal have been discussed previously in META #1 2013: Portals to Computing.)

A central part of the Galaxy interface is a sophisticated workflow manager, enabling the user to specify and execute a series of computations. For example, starting with a PDF document uploaded by the user, she might further want to perform content extraction, sentence segmentation, word tokenization, part-of-speech tagging, syntactic parsing, and finally identification of subjective expressions with positive polarity – all carried out in a consecutive sequence. The output of each component provides the input to the next connected component(s) in the workflow, creating a potentially complex pipeline. Then, after the desired workflow has been specified; at the click of a few buttons, the resources and tools involved are configured and submitted to the Abel cluster at UiO.
The fact that the portal provides an easy-to-use front-end to an HPC cluster is a crucial feature. As mentioned above, language technology can be computationally quite expensive, often involving sub-problems where known best solutions have exponential worst-case complexity. At the same time, typical language analysis tasks can be trivially parallelized, as processing separate documents (and for many tasks, also individual sentences) constitute independent units of computation. The portal transparently submitting the sub-tasks of a workflow to the cluster — without the need for user knowledge about job scheduling, for example — means that users are enabled to perform analyses that might otherwise not be possible (and, of course, to obtain experimental results faster and over larger data sets). Galaxy also maintains so-called "histories" for each user, tracing the computational steps in a workflow. Histories, as well as workflows and data, can be shared with other users, a feature of great potential for the reproducibility of experimental results.

DATA INTERCHANGE

Through the simple and uniform Galaxy interface, a user can combine the various tools into workflows, have them executed on the cluster, and retrieve the results. One important challenge here, however, concerns how to make diverse LT tools work together seamlessly through workflows designed dynamically by the user. Unfortunately, existing LT tools are often based on different and mutually incompatible representation formats for encoding input and output data. If one wants to chain together different tools that use different encoding formats, some way of translating between one and the other is required. For this purpose LAP implements an interchange format that is used as a lingua franca to ensure interoperability between tools in the portal.

The LAP-internal interchange format is based on the so-called Linguistic Annotation Framework (LAF), an ISO-standardized model for representing linguistic annotations, and for scalability reasons we store these in a NoSQL database (MongoDB) rather than using a file based format. Upon completion of a workflow, the database records can be serialized into any number of file formats specified by the user. The database approach is also practical in terms of parallelization, in that different jobs can invoke separate parts of the database, eliminating the need for wrestling with the bookkeeping of splitting and merging files. Of course, having an interchange format does mean that each tool to be integrated in the portal needs a wrapper that makes use of the appropriate format converter to handle insertion into, and queries from, the database.

CURRENT STATE OF IMPLEMENTATION

A first prototype implementation of LAP is available [4] for testing, and the first production instance is scheduled to launch in mid-2015. We have designed and started to populate a repository of tools — called LAP Tree — with ready-to-run installations of processing components used by LAP.

Here, central design desiderata were replicability (versioning) and relocatability (in terms of file system locations), such that central installation on compute nodes is completely avoided. Instead, the LAP Tree is realized as a version-controlled repository, where individual tools can be checked out by an arbitrary user and into an arbitrary location, for immediate use through Galaxy. The mid-term goal is to enable users to request individual versions of tools directly in Galaxy, and then dynamically populate a user-specific instance of relevant parts of the LAP Tree.

Integration of computational portals with national allocation practices is a bit of a balancing act, where ease of access and use must be weighed against the principles of fairness and transparency in the use of shared computing resources. LAP is integrated with standard Notur project management and accounting and effectively provides the technical infrastructure for different Notur projects (i.e., resource allocations) to use core hours from their account.

We have met with the UiO LifePortal community and the Notur Resource Allocation Committee to understand the parameters of existing usage constraints and to design a LAP-specific allocation policy. For the current trial period, a bulk allocation to LAP is sub-allocated according to the following principles: (a) First-time LAP users, who typically do not have access to an existing Notur project, are granted a standard quota of 10,000 cpu hours per six-months period; (b) users or groups of users can apply for LAP-internal sub-projects, typically for allocations of up to 100,000 cpu hours per period; and (c) users or groups of users can independently acquire Notur allocations and use these transparently for LAP computation.

Links
[1] https://clarin.b.uib.no
[3] https://lifeportal.uio.no
[4] https://lap.hpc.uio.no
The Research Council of Norway has adopted its first policy on open access to research data from publicly funded projects. The guidelines apply to all data generated by projects funded by the Research Council – with a few exceptions. The main principle in the Research Council’s new policy is that research data in general must be accessible to relevant users, on equal terms, and at the lowest possible cost.

“This is an important step towards achieving a research system that makes the most effective use of the large amounts of data collected as part of research projects,” says Arvid Hallén, Director General of the Research Council.

The study has given The Research Council insight into how researchers in Norway share and store their research data, necessary for developing a policy on the sharing and archiving of publicly funded research data.

Norwegian researchers are willing to share research data. The findings of a study carried out by DAMVAD Norway show that most Norwegian researchers are in favour of making research data accessible to other scientists. Nearly three of four researchers are willing to share their data. Eight of ten believe that open access to research data enhances research.
RECOMMENDATIONS, NOT REQUIREMENTS
The guidelines in the policy describe how the data should be stored, made accessible to others, and shared. Key concepts in this context are embargo periods, provision of licenses for use of research data, and the need for a long-term plan for data management.

The policy is formulated as a set of recommendations – not requirements. The Research Council’s various funding instruments will be adapted to accommodate the principles in the policy. One possible measure is to introduce a requirement that research projects incorporate data management plans. The Research Council would use these plans as a basis for accepting data accessing costs as part of the operational expenses of funded projects. In addition, the Research Council will focus on encouraging the establishment of well-designed infrastructure for data storage and data management, in part through the National Financing Initiative for Research Infrastructure.

The policy describes how the Research Council will use its instruments to promote more open access to research data, providing a basis for accepting data accessing costs as part of the operational expenses of funded projects. It also focuses on encouraging the establishment of well-designed infrastructure for data storage and data management, in part through the National Financing Initiative for Research Infrastructure.

The guidelines in the policy are intended to assist researchers when planning their research projects.

KEY PRINCIPLE
Better access to research data will enhance the quality of research in that results can be validated and verified in a more effective manner and data sets can be used in new ways and in combination with other data sets. Open access to research data will also help to avoid unnecessary duplication of efforts and pave the way for more interdisciplinary research.

“The results of publicly funded research comprise a public good that is valuable to the research community as well as to society at large. This is set out as a key principle in the new policy. The policy also describes how the Research Council will use its instruments to promote more open access to research data,” says Mr Hallén.

EXCEPTIONS: PROTECTION OF PERSONAL PRIVACY, COMMERCIAL FACTORS AND SECURITY CONCERNS
Some data sets may not be made openly accessible as a matter of course. The Research Council’s policy therefore establishes clear exceptions. Data sets must not be made openly accessible if doing so could pose a threat to personal or national security or conflict with the applicable statutory framework regarding the protection of personal privacy or other legal provisions.

Data that have commercial value and are generated in projects in which a company is the contractual partner with the Research Council may be exempted from the general principle of open access. In these cases, it is recommended that the data are made available after a certain period of time, preferably after three or five years. Data that are highly impractical or costly to make accessible may also be exempted from the general principle of open access.

Public access to data is a national strategic ambition in Norway. NorStore has developed a framework for handling, documenting and archiving research data and integrated it with its existing services in order to make the data publication process as simple and efficient as possible.

The ‘NorStore research data archive’ provides long-term archiving, metadata provisioning, data sharing, content integrity and a framework for curation of legacy research data. Search and download of data is accessible to anyone, while ingestion of data is available to eligible users who register and authenticate themselves using the national federated credentials (Feide) or OpenID. Published data sets are registered with a unique data object identifier (DOI) that is associated with the scientific publication. Two type of data licenses are offered, Norwegian License for Open Government Data (NLOD) and Creative Commons version 4 (CCv4).

The aforementioned Data Archive service (currently being evaluated as a pilot service) and the NorStore resource allocation service, dubbed ‘project area’, researchers are capable of both storing, processing and archiving large and complex data sets within in a single national e-infrastructure. It enables researchers to implement both short and long-term data management requirements, typically set out in a data plan.

Written by Andreas O. Jaunsen.
The next call, period 2015.1, for researchers requiring national e-infrastructure resources in Notur or NorStore, will soon be announced. UNINETT Sigma has developed a new application submission tool that aims to make the application process for researchers more targeted and efficient, pulling necessary information from other supporting systems. Starting with the announcement of the call, the new tool will be accessible to all users and hopefully contribute to a simpler, more user-friendly and efficient application process.

**NOTUR AND NORSTORE: NEW PROJECT APPLICATION TOOL**

**UNIFIED APPLICATION FORM**
The application form itself has been restructured, simplified and unified. Instead of the previous separate application forms, there is now only a single form for both Notur and NorStore. The new application system is integrated with the existing systems for quota and accounting. Existing Notur and NorStore projects may therefore continue to use their allotted account numbers, whereas new project leaders will be asked to submit contact details before entering the application forms. Authentication is based on Feide or OpenID and it is therefore necessary that our existing users have communicated their FeideID/OpenID to the administration.

**OLD APPLICATIONS STILL AVAILABLE**
Old applications submitted by researchers in previous periods will be available directly from the new application forms. From the 2015.2-announcement and onwards, you will be able to update the existing application to reflect any changes in the project application, for example the expected need for resources.

Contact: sigma@uninett.no
The current allocations by the Research Council of Norway of the High-Performance and Storage facilities of the Notur and NorStore projects expire 30 March 2015. Projects who wish to continue using the facilities after this date, must apply to this call. Applicants that already applied for access for the period 2015.1 in a previous call need not apply.

The deadline for applying for applications for period 2015.1 is 23 January. Earliest project start is 1 April 2015.

ADVANCED USER SUPPORT
It is also possible to apply for advanced user support. For this call the administration will accept applications for well defined and time-limited advanced user support proposals (Notur and NorStore). In early 2015 UNINETT Sigma2 will evaluate the advanced user support arrangement with the intention to revise the framework and conditions to ensure effective and efficient use of the resources and funding.

For more information about the call for resources, please visit www.notur.no and www.norstore.no
**Upcoming Events**

**NeIC 2015. The 2nd Nordic e-infrastructure Collaboration Conference**
5 - 8 May, 2015
Hanasaari, Espoo, Finland
neic.nordforsk.org/neic-conference-2015

**PRACEdays 15**
PRACE Scientific and Industrial Conference
26 - 28 May 2015
Dublin, Ireland
www.prace-ri.eu/pracedays15

**ICCS 2015**
International conference on computational science
1 - 3 June, 2015
Reykjavik, Iceland
www.iccs-meeting.org/iccs2015

**ICS 2015**
29th International Conference on Supercomputing
8 -11 June, 2015
California, USA
www.cs.ucr.edu/~ics15

**ISC High Performance**
12 -16 July, 2015
Frankfurt, Germany
www.isc-hpc.com

---

**SUBSCRIBE to the META magazine?**

To see earlier editions and to subscribe to the paper version of the META magazine, please visit:
www.notur.no/publications/magazine