Vetenskapsrådet

SWEDISH SCIENCE CASES FOR E-INFRASTRUCTURE

VETENSKAPSRÅDET 2014

SWEDISH SCIENCE CASES FOR E-INFRASTRUCTURE

Henrik Grönbeck
Per-Olof Hulth
Lennart Johnsson
Matts Karlsson
Ulf Pettersson
Gunilla Svensson
Elizabeth Thomson

Editor: Anders Ynnerman

This corrected report was published 09/06/14. It replaces the earlier published version.

SWEDISH SCIENCE CASES FOR E-INFRASTRUCTURE

Rapporten kan beställas på www.vr.se

VETENSKAPSRÅDET
101 38 Stockholm

ISBN 978-91-7307-240-3

Editor: Anders Ynnerman
Original: Eprint
Tryck: DanagårdLiTHO, Motala, 2014

FOREWORD

The Swedish Research Council is a governmental agency with the responsibility to support basic research of the highest scientific quality in all academic disciplines. It is also part of the Council´s remit to evaluate research and assess its academic quality and success. The Council for Research Infrastructure (RFI), at the Swedish Research Council, has the overall responsibility to see to that Swedish scientists have access to research infrastructure of the highest quality. Specifically, RFI assesses the needs for research infrastructure in a regularly updated roadmap, launches calls and evaluates applications, participates in international collaborations and works with monitoring and assessments. Well-functioning e-infrastructures, such as digital communication, storage and computing capacity, together with human resources to aid in the usage of these infrastructures, are a prerequisite for most scientific disciplines today; both to support research projects and as a basis for other research infrastructures. The demand for existing e-infrastructures is very high and it is expected to increase even further. In addition, new services will also be required. With this investigation the Swedish Research Council has initiated a broad effort to map existing and future scientific needs for e-infrastructures. RFI has invited Professor Anders Ynnerman to lead the work and throughout the process he has been strongly supported by seven panels composed of distinguished scientists from different disciplines. The report at hand presents a diverse set of science cases that span a broad spectrum of existing research, and points to potential breakthroughs that can be made if sufficient supporting e-infrastructures are available. Such a report can never claim to cover all possible areas, but the effort has been to present a representative view of scientific needs as they are known today. The report was presented to RFI on November 2013 and will be used in the future strategic work of the council. On behalf of RFI I thank Professor Anders Ynnerman and the scientific panels for their excellent work.

Stockholm 2014-03-15

Juni Palmgren
Secretary General
The Council for Research Infrastructures
The Swedish Research Council

CONTENTS

1. EXECUTIVE SUMMARY

1.1 Motivation and Task

e-Science methods and tools are gaining an ever growing importance across a wide range of scientific disciplines, both in Sweden and globally. e-Science is a concept that builds upon the use of computer hardware and software as well as on human expertise to enable scientific discovery based on computation and exploration data from various sources, including simulations as well as experimental data, and databases. e-Science has its foundation in applied computer science and mathematics, and makes intense use of hardware infrastructures such as high performance computing, networking and visualization. The overall goal of this report is to provide the Swedish Research Council with executive information on the scientific requirements for future e-Science infrastructures in Sweden. The report does this by describing selected scientific cases that provide examples of the scientific results that can be obtained if the specified requirements are met. From an international perspective it is of utmost importance for a knowledge intensive society to provide a competitive and complete e-Infrastructure for research and development. The Swedish national e-Infrastructure must thus provide resources and services that enable Swedish scientists to compete at the leading edge of research and participate as front runners in international research collaborations and thus ensuring that the opportunities for groundbreaking research, as outlined in this report, are realized and that the e-Science paradigm spreads to new disciplines and contributes to the economy and well-being of our society.

1.2 The e-Science cases

The work on documenting the e-Science cases was conducted by seven panels:

- Climate and Environment
- Astrophysics, High Energy and Particle Physics
- Engineering Sciences
- Humanities, Social, Educational Science, and Epidemiology
- Life Sciences and Molecular Medicine
- Material Science, Chemistry and Nano Science
- Mathematics and Computer Science

The panels were chosen to represent areas of research that from an e-Science perspective are mature as well as areas where e-Science has just begun to make an impact. Each panel presents in their corresponding chapters several *Science Cases* that were selected to give a good understanding of the requirements that the potential scientific breakthroughs outlined will put on a national e-Infrastructure. The report does not intend to provide a complete listing even of key research efforts and associated possible breakthroughs.

1.3 Potential breakthroughs enabled by e-Infrastructure Investments

Each of the selected science cases in this report gives examples of exiting breakthroughs that can be made if internationally competitive e-Infrastructures are made available to the top Swedish researchers. enhancing their leadership capabilities. A comprehensive view of all these breakthroughs can be obtained only by careful study of all the chapters in the report. To provide a flavor of what can be expected and show the breadth and importance of applications depending on e-Infrastructures a few selected examples are given here:

- **Treating the earth as a system** – Use of sophisticated models and access to orders of magnitude larger resources than today will make it possible to address problems related to the human influence on the global climate, including coupling of the carbon cycle with the atmosphere, land and ocean. With the help of e-Science we are beginning to simulate the earth as a system and thereby enable more accurate and long term predictions for the development of our planet.
- **Pushing the boundaries of our knowledge of the universe** – In physics the boundaries of human knowledge are pushed in both macro and micro cosmos. By utilizing the e-Infrastructures it was possible to detect the Higgs particle, and there is a multitude of data and simulation driven discoveries still to be made involving physics beyond the standard model, finding explanations for such phenomena as the prevalence of matter over antimatter and the nature of dark matter in the universe.
- **Designing future materials and drugs** – In material science and biological simulation a paradigm shift is under way in which multi-scale methods can be used to bridge the gap between the atomic scale and macroscopic quantities. This will have tremendous impact on the way in which we can

predict and tailor material properties and how pharmaceutical drugs can be designed. Multi-scale simulations are, however, computationally very demanding and a massive increase of available computational resources is needed to enable breakthroughs.

- **Understanding complex diseases** – Through sequencing and massive data exploration and simulation it will be possible to reach an understanding of diseases linked to multiple human genes. This could lead to new diagnostic tools and treatments of many disabling diseases. The data handling in these projects is a major concern both from a storage and a processing perspective; the lack of dedicated resources for Swedish researchers limits the development in the field.
- **Improving the efficiency of fluid systems** – By accurate simulation of fluid flows the efficiency of various fluid systems involving turbulence and reactions can be improved, e.g. in the context of airplane wings, vehicle aerodynamics, wind turbines and internal combustion engines. Such insights could e.g. lead to lower airplane drag and improved engineering models. The complex turbulent flow structures that need to be accurately simulated will require both method development and access to large computational systems with high performance processors, internal networks and storage.
- **Mining of social media to understand political ideas and actions** – Collection and analysis of social media data in real time and linked to geographical coordinates would provide an entirely new method for observing origins, diffusion and disappearance of political ideas and actions. Apart from collection strategies and data handling, there is an urgent need for development of policies and legislation for data access and distribution spanning a variety of sources and uses to enable ground breaking research and provide insight for societal development. The importance of data access and distribution is an underestimated aspect of e-Science that needs immediate attention
- **Enabling seamless collaboration** – Secure, controlled sharing of resources combined with multi-modal and more intuitive, human centric interfaces will considerably extend sharing and collaboration within and across disciplines and further enable team formations based on competence and skills rather than geographic proximity for increased rate of innovation and productivity. Significantly improved human interfaces and enhanced security may help accelerate the integration of sensor systems related to health as well as vehicle systems into comprehensive information systems.

Underpinning all of the described breakthrough examples described above is a core of mathematics and computer science in which many of the tools needed to achieve breakthroughs are developed. The report provides, see chapter 4, an account of developments in mathematics and computer science, including software platforms, test beds and educational efforts, needed to provide the e-Science community with the needed tools and methods.

As can be seen from the list above e-Science is already deeply and widely embedded in Swedish research and there is an opportunity for Sweden to strengthen Swedish participation and enable leadership in the domains described in the report by provision of e-Infrastructures meeting the requirements detailed herein.

1.4 Major findings and conclusions

Based on the evidence produced by the panels and presented in panel chapters of this report 10 major findings have been made. These are:

1. Significantly enhanced resources and services will enable exciting breakthroughs in several disciplines and can be spearheaded by Swedish researchers.
2. Development of methods, tools and software within core disciplines is necessary to make breakthroughs.
3. Advanced and long-term user support and human infrastructures are keys to e-Science adoption.
4. The simulation paradigm dominates the current Swedish needs for e-Infrastructure. A complementary and more data centric aspect of e-Science should be promoted.
5. International participation depends on access to national infrastructure compatible with international infrastructures.
6. User communities must be actively engaged in the prioritization, design, deployment and operation of e-Infrastructures.
7. e-Social Science and e-Humanities are potentially very large users, but need active support like other communities new to e-Science.
8. e-Science methods and tools are in increasing demand and will be instrumental in increasing interaction between tool makers and tool users.
9. Secure and controlled access to data, software and other resources must be enhanced and simplified.
10. Improved co-ordination of the national e-Infrastructure and e-Science initiatives is needed.

Though the existing Swedish e-Infrastructure works well in most aspects and Swedish e-Science is in some areas at the international forefront the above key findings must be urgently addressed. The findings are elaborated upon in section 11.1 and they form the basis for the conclusions and recommendations provided.

From the panels projections of e-Infrastructure needs for Swedish competitiveness and leadership, it can be deduced that even if expected technology developments are taken into account a significant gap will quickly develop at the current level of investments in national e-Infrastructures.

The items deemed most critical are:

- Capacity computing – HPC systems tailored for throughput of many independent jobs or large jobs with extreme scalability.
- Storage – Large scale storage solutions that are integrated with database- and visualization services
- Software – Efforts to develop new software to address new problems and new approaches
- User support – The pool of human resources providing qualified assistance to users
- Data access policies – Access and distribution of data is key to many fields.

The evidence provided in the report is clearly showing that unless action is taken on these items the projected deliverables and breakthroughs described in the panel chapters will not be enabled.

1.5 Recommendation for e-Infrastructure investments

The conclusion of the report is that there is currently an opportunity for Sweden to take a leading role in the on-going transformative e-Science evolution process. This calls for further strengthening of Swedish e-Science in established and already internationally competitive areas as well as accelerated introduction of e-Science in new areas of e-Science with potentially high impact. In the spirit of this conclusion an agenda for the promotion of e-Science in Sweden should be put in place. It is recommended that the Swedish Research Council comissions a separate investigation providing recommendations for the definition and implementation of such an agenda.

1.6 Sammanfattning på svenska

Metoder och verktyg för e-vetenskap får allt större betydelse inom många olika vetenskapliga forskningsfält, både i Sverige och i världen. E-vetenskap bygger på användandet av datorers hårdvara och mjukvara, såväl som mänsklig expertis, för att möjliggöra vetenskapliga upptäcker, baserade på beräknings- och forskningsdata från olika källor. Här ingår såväl simuleringar som experimentell data och databaser. E-vetenskapen har sin grund i tillämpad datavetenskap och matematik. Här används i stor utsträckning hårdvaruinfrastrukturer som högpresterande beräkning, nätverk och visualisering. Den här rapportens övergripande mål är att förse Vetenskapsrådet med information om de vetenskapliga kraven på framtida e-vetenskapsinfrastrukturer i Sverige. Rapporten gör detta genom att beskriva utvalda vetenskapliga fall, vilka ger exempel på vetenskapliga resultat som kan uppnås om de specifika kraven uppfylls. Ur ett internationellt perspektiv är det av största vikt för ett kunskapsintensivt samhälle att tillhandahålla en konkurrenskraftig och fullständig e-infrastruktur för forskning och utveckling. Den svenska nationella e-infrastrukturen måste därför tillhandahålla resurser och tjänster som möjliggör för svenska forskare att konkurrera i forskningens framkant och inta ledande roller vid internationella forskningssamarbeten. Därigenom kan man säkerställa att de möjligheter till banbrytande forskning som skildras i denna rapport förverkligas, samt att e-vetenskapsparadigmet sprids till nya forskningsfält och därmed bidrar till samhällets ekonomi och välfärd.

I rapporten framställs tio betydande iakttagelser:

1. En betydande ökning av resurser och tjänster kommer att möjliggöra intressanta genombrott inom flera forskningsfält och dessa kan ledas av svenska forskare.
2. Utveckling av metoder, verktyg och mjukvara inom de huvudsakliga forskningsfälten är nödvändigt för att göra genombrott.
3. Avancerad och långsiktig användarsupport samt mänskliga infrastrukturer är centrala i införandet av e-vetenskap.
4. Simuleringsparadigmet dominerar nuvarande svenska behov av e-infrastruktur. En kompletterande och mer datacentrerad aspekt av e-vetenskap bör främjas.
5. Internationellt deltagande är beroende av tillgång till nationell infrastruktur som är kompatibel med internationella infrastrukturer.
6. Användargrupperna måste vara aktivt involverade i prioritering, utformning, spridning och hantering av e-infrastrukturer.

7. E-vetenskap inom samhällsvetenskap och humaniora har potentiellt väldigt frekventa användare, men behöver aktivt stöd, precis som andra användargrupper där e-vetenskap är nytt.
8. Efterfrågan på metoder och verktyg för e-vetenskap ökar stadigt och kommer att bidra till att öka interaktionen mellan tillverkare och användare av verktygen.
9. Säker och reglerad tillgång till data, mjukvara och andra resurser måste stärkas och förenklas.
10. Förbättrad samordning av nationell e-infrastruktur och e-vetenskapsinitiativ behövs.

Rapportens slutsats är att det för närvarande finns möjlighet för Sverige att inta en ledande roll i den pågående transformativa utvecklingsprocessen för e-vetenskap. Detta kräver att den svenska e-vetenskapen stärks inom etablerade områden där det redan finns internationell konkurrens. Det krävs också en påskyndad introduktion av e-vetenskap inom nya områden med potentiellt stor genomslagskraft. I enlighet med denna slutsats bör en agenda för främjandet av e-vetenskap i Sverige tas fram. Vetenskapsrådet rekommenderas beställa en särskild utredning som ger rekommendationer för definieringen och implementeringen av en sådan agenda.

2. PRELIMINARIES

2.1 Mandate and Scope

This report was commissioned by the Swedish Research Council with the overall aim being to provide the council with executive information on future requirements for e-Science infrastructures in Sweden, and on the science that motivates future investment in e-Infrastructures at the national level. The remit can be summarized as:

- To provide an overview of e-Science research in Sweden and present selected Science Cases representing the various research areas.
- To account for the current needs for different e-Science infrastructure services and to predict future requirements.
- To identify and describe the infrastructure-related challenges the e-Science areas are facing.
- To describe the potential scientific breakthroughs that an e-Infrastructure might enable if the challenges are met.

To deliver the requested information this report is based on several Science Cases, each presented in a separate chapter composed by seven independent panels. A Science Case is defined as a description of research that is expected to be enabled by use of a specific infrastructure. The science cases provide both examples of existing areas making use of national (and international) e-Infrastructures, as well as emerging areas of e-Science.

The main task is to report on the academic aspects of e-Science, but this report should also describe relevant interplay with industry, government agencies and other organizations. An important aspect of the mandate is that this report should provide foresight into future needs rather than evaluating the services provided by the existing Swedish e-Infrastructure, nor should it evaluate the organizational forms behind the e-Infrastructure.

2.2 Structure of work and presentation of panels

The primary source for this report is derived from work conducted during the spring of 2013 by seven expert panels covering key areas of e-Science. The intention is not to cover all aspects of Swedish e-Science and all potential research areas that could benefit from an e-Infrastructure, but rather to select a representative subset to serve as examples of the importance of e-Science and give examples of the impact that an e-Infrastructure can have, and to show what requirements the users of the infrastructure have already identified as well as how those requirements are expected to develop in the future.

The seven panels were appointed by the Swedish Research Council and each led by a designated chairperson. The chapters in this report have been composed by the panels under the guidance of the chairperson and should be seen as independent contributions, but in a format that enables comparison. These contributions form the basis for the conclusions presented in the common part of this report. The Swedish Research Council also appointed an overall chair for co-ordination of the work of the panels and to serve as the main editor of this report and the principal author of the general sections as well as the summaries of e-Science evidence and the projections for future needs of e-Infrastructures.

The panels were given the task of providing overviews of their research areas and selecting the science cases to be presented. In their work the panels were asked to consider and elaborate on the following aspects for the selected cases:

- Current use of e-Infrastructures - regionally, nationally, and internationally
- e-Infrastructure related challenges
- Potential breakthroughs that could be enabled through e-Infrastructures
- Projected needs and requirements for the full range of e-Infrastructure-related resources on 2 year, 5 year, and longer-term time scales until 2020.

The panels organized regular meetings during the spring of 2013 and the panel chairs held teleconferences with the overall chair on a monthly basis. Drafts of the panel contribution chapters were presented at an open e-Infrastructure workshop held on May 24th 2013. At this workshop researchers were invited to listen to presentations of the preliminary findings of the panels and to discuss the distributed draft of this report. Based on the outcome of the workshop revised panel contributions were developed and compiled into this report.

The final report was compiled and presented to the Council for Research Infrastructures (RFI), a council under the Swedish Research Council, in November 2013.

Each panel consisted of three to five members creating a balance of research areas, university affiliations, and gender. The members of the panels were selected by the Swedish Research Council based on recommendations by the panel chairs. The panels and their corresponding chairpersons are:

- Astrophysics, High Energy and Particle Physics (Chair: Per-Olof Hulth)
- Climate and Environment (Chair: Gunilla Svensson)
- Engineering Sciences (Chair: Matts Karlsson)
- Humanities, Social and Educational Science (Chair: Elizabeth Thomson)
- Life Sciences and Molecular Medicine (Chair: Ulf Pettersson)
- Material Science, Chemistry and Nano Science (Chair: Henrik Grönbeck)
- Mathematics and Computer Science (Chair: Lennart Johnsson)

Professor Anders Ynnerman of Linköping University was appointed overall chair for the work of the panels and has served as editor of the report and author of the non panel specific sections in the report.

Full lists of the panel members, together with descriptions of areas of expertise are provided at the end of each of the panel chapters.

3. INTRODUCTION

We live in the era of digital information. Over recent decades the world has gone from an information poor to an information rich, or even information overflow, society. New technological developments allow us to create new data through creative activities in performing arts, writing and daily activities, computation and measurement; and to store it, refine it, share and transmit it at an ever increasing rate. In research and development efforts the ability to access data has come to play a fundamental role in the exploration process as well as in the documentation and presentation stages. Research is increasingly carried out through team efforts, with teams formed based on merit with respect to the challenges being addressed and not on geographic proximity. Some challenges require expensive instruments, or other forms of resources, that can only be afforded through international collaborations. Other problems require access to data generated or housed in several locations, sometimes nationally sometimes globally. The increasing amount and new kinds of data, and the ability to draw upon many different forms of data sources, also enables researchers and developers to address a whole new range of problems. Data driven R&D is sometimes referred to as the "fourth paradigm" [HTT09] in the evolution of scientific discovery, moving from empirical studies to theoretical considerations, computer-based simulations and now to data-centric paradigms. e-Science embraces this evolution of scientific methodology and embodies the computation- and data-intensive approaches. The need for and use of e-Science methods and tools goes far beyond the academic use and, as shown in Fig. 3.1, the underpinning infrastructure can create an impact over the full range of applications found in modern society.

There are many different definitions of e-Science found in the literature. It is clear, however, that the most common interpretation of the concept involves aspects of computationally- and/or data-intensive science conducted on networked facilities enabling widespread collaboration. The term e-Science was coined by John Taylor, the Director General of the United Kingdom´s Office of Science and Technology, in 1999 and was used to describe a large funding initiative that started in November 2000. The initiative was reviewed in 2009 and the importance of a national strategy for e-Science and the enabling implementation and operations of effective e-Infrastructures is underlined in the conclusions provided in the report from that 2009 evaluation of the UK e-Science program which highlight four motivating facts for this:

- e-Science programmes world-wide have emerged from the exponentially increasing push of Information and Communication Technology (ICT) capacity, coupled with the pull of demand for transformative tools and methods to support the complexity, diversity, and integrative needs of modern scientific research.
- There is a strong need for a transformative infrastructure to facilitate transformative research.
- e-Infrastructures can accelerate knowledge creation that lies at the heart of innovation which is fundamental to economic and social well-being.
- e-Science serves as a pilot project for application of e-infrastructure in other sectors such as industry, commerce, learning, security, and crisis response.

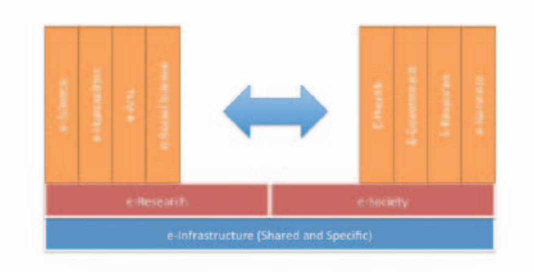


Figure 3.1: The tiered layers of applications of e-Infrastructures

It can already be seen that e-Science tools are used by researchers across all fields of academia, and support research collaboration across topical and geographical boundaries in a natural way. This means that e-Infrastructures have an increasingly important role in the national and international research infrastructure landscape and are seen as providing enabling services to other large scale infrastructures.

This report aims to describe the exciting fundamental changes that full deployment of the e-Science paradigm could bring to Swedish research by providing descriptions of scientific breakthroughs that could be enabled, and outline the foreseen challenges and demands that will be put on the future e-Infrastructure in Sweden. As is evident from the science cases in this report, as well as from global trends in e-Science and the rapid development of e-Science initiatives, it can be concluded that it is of the utmost importance that an internationally competitive and complete e-Infrastructure be provided to Swedish researchers. The national e-Infrastructure must provide services and resources that enable Swedish scientists to compete at the international forefront and participate as forerunners in international research collaborations and offer convenient and effective access to non-

Swedish e-Infrastructures to facilitate this objective. The science cases in this report highlight several research efforts that have a national as well as an international collaborative character, and which require that the Swedish e-Infrastructure be operated and evolve in ways compatible and consistent with other national and international e-Infrastructures.

3.1 e-Science and e-Infrastructures

e-Science has, to a large extent, emerged from the traditional high-end computing and data analysis communities, but is gradually spreading to all areas of science, technology, the humanities and even the arts. In this wider impact of e-Science it is recognized that e-Science does not necessarily depend on access to large-scale resources and it is important to be inclusive in the definition of e-Science activities as high quality science can be be conducted on a range of both small and large digital platforms, and frequently exploits different levels during different project phases. This report attempts to be inclusive of different kinds of usage and needs. As the scope of this report is limited to national e-Infrastructures, however, it has a bias towards the base of users, from numerous research institutions and disciplines, who have need of very large resources and associated services and/or need of access to unique international resources available through non-Swedish e-Infrastructures. Several examples are provided by the panels addressing the opportunities and needs in the natural sciences, life sciences, engineering, and humanities, social and educational sciences. An e-Infrastructure is, in the context of this report, thus taken to mean an infrastructure containing nationally available:

- digitally-based technology (hardware and software),
- resources (data, services, digital libraries),
- communications (protocols, access rights and networks), and
- people and organizational structures needed to support modern, internationally leading collaborative research be it in the arts and humanities or the sciences and engineering.

and the combination and interworking of all of these, as well as facilitating access to unique resources and services in other national or international e-Infrastructures. Note that this definition of the infrastructure also includes the human expertise required to enable and support the actual research efforts and to operate, maintain and evolve the hardware and software systems including any adaptation required by the research communities.

In the report entitled *Riding the wave* from 2010 [Woo+10] a wish list for the data aspects of a scientific e-Infrastructure is provided. Adding aspects of the computational demands and human resources to this list leads to e-Infrastructure services of the future being a techno-human ecosystem providing:

- seamless and reliable access to effectively unlimited state of the-art-computational and storage resources for simulation and data processing
- national (international) authentication, authorization and accounting systems
- high bandwidth and high availability secure networking
- storage hierarchies supporting easy and secure storage of data generated, guaranteeing data authenticity and preservation for long term storage of data
- data curation services integrated with the storage and computation services and with support for the generation of metadata
- a number of software and database solutions, often community specific, provided as services with abstraction layers to the underlying data, software and hardware resources
- a human infrastructure including experts in a range of e-Science topics available for short and long term consultancy
- an effective and transparent resource allocation mechanism

The implementation and operation of an e-Infrastructure providing high quality services is a challenging task and there are many obstacles faced by e-Infrastructure providers and funding agencies. Even though it is not the primary task of the panels to provide guidelines for the implementation and operation of the future e-Infrastructure, and the educational and training needs for the various disciplines for this major change in how research is pursued, some recommendations are provided in section 11.2 on how to move in the direction of an infrastructure that fulfills the objectives presented above.

An infrastructure is, by its very nature, a long term commitment, even if many components and technologies in the case of e-Infrastructure have only very limited life-times. Funding strategies need to be found to assure the longevity of e-Infrastructure and the considerable investments made in the resources it comprises and the human capital required for its evolution, support, operation and maintenance.

A holistic approach to the overall efficiency of the public research enterprise, of which e-Infrastructures are (or should be) an essential component, should be developed to ensure long term competitiveness and maximize societal benefits. Short-term tactical approaches should not jeopardize long-term success and benefits.

3.2 International Outlook on e-Infrastructures

Internationally, large research or research infrastructure efforts with significant e-Science components are emerging or being further developed. For example, in Europe the European Strategy Forum for Research Infrastructures (ESFRI) has listed 48 emerging major European research infrastructures, most of which are in need of large-scale data and computing services. On the international arena, the US has since long had a leading position within e-Infrastructures, but today countries like China and Brazil are investing heavily in e-Science research and often also e-Infrastructures to leap-frog the process of scientific progress and take a leading position in research.

On the European scale a number of major e-Infrastructure projects and initiatives are building or operating general-purpose service layers for e-Science. These entities form the European e-Infrastructure collaborations needed for European collaborative e-Science research, and they also provide or aim at providing very extensive resources that could not be built by a single country. Recently, the European e-Infrastructure Reflection Group (e-IRG) recommended that the European efforts are further developed into a "e-Infrastructures Commons" for knowledge, innovation and science in order to, e.g., meet the challenges of implementing the EU´s 2020 Strategy. The implementation of such a Commons requires well-defined roles among stakeholders and a high degree of collaboration and standardization. Also, to be able to provide leading-edge services in a sustainable way, constant innovation needs to be included. The Commons must be flexible and able to change to fulfill the needs by all users of European e-Infrastructure. A fundamental feature of the implementation is that an ecosystem of different organizations is needed, with clearly defined roles e.g. as user communities and providers of operational services, innovation and coordination.

The Swedish National Infrastructure for Computing (SNIC) provides a gateway to the main European e-Infrastructure initiatives related to large-scale computing, Partnership for Advanced Computing in Europe (PRACE) and the European Grid Infrastructure (EGI). SNIC also participates in emerging data infrastructures and collaborations such as the European Data Infrastructure (EUDAT) and the Research Data Alliance (RDA). At the Nordic level, SNIC takes part in the Nordic e-Infrastructure Collaboration (NeIC).

3.2.1 PRACE

The Partnership for Advanced Computing in Europe provides access and services including assistance in code porting on a competitive basis to high-end computational resources at a scale of about 10 times the largest Swedish national resources. Two of the 10 most powerful computers in the world (458,752 and 147,456 cores respectively) according to the Top500 ranking as of June 2013 are available through PRACE. Swedish researchers have successfully competed for access to PRACE resources. Proposals for access to PRACE resources are evaluated and ranked entirely based on merit by world class experts in the area of the proposal. Hence, success in the PRACE competition for access is a good measure of the international competitiveness of the applicants research. In the PRACE DECI calls issued twice a year about 100 million core-hours are typically allocated with SNIC contributing about 10% and swedish researchers hence having access on a competitive basis to about 90 million core-hours on a diversity of European major platforms. In 2012 a report presenting science cases for PRACE was published [G+12].

3.2.2 EGI

The European Grid Initiative (EGI) is operating a general pan-European distributed computing infrastructure building on the development and operational work that has been done by earlier international grid projects. EGI is based on a collaboration between National Grid Initiatives (NGIs) such as e.g. SNIC/SweGrid.

3.2.3 EUDAT and RDA

For computing, the European e-Infrastructure landscape has developed to form two major initiatives, PRACE and EGI. For data, the situation is less clear and several on-going projects exist. A main effort is made in the European Data Infrastructure (EUDAT), which is a three-year project that is developing a comprehensive picture of the data service requirements of the research communities in Europe and beyond and deliver a Collaborative Data Infrastructure (CDI) with the capacity and capability for meeting future researchers´ needs in a sustainable way. This will become increasingly important over the next decade as we face the challenges of massive expansion in the volume of data being generated and preserved and in the complexity of that data and the systems required to provide access to it. Another important effort is made within the global Research Data Alliance (RDA), RDA facilitates implementation of the technology, practice, and connections that make research data available across national borders and other barriers, and RDA aims to accelerate and facilitate research data sharing and exchange. The work of the Research Data Alliance is primarily undertaken through a number of open working groups.

3.2.4 Nordic e-Infrastructure collaboration

At the Nordic level, the Nordic eInfrastructure Collaboration (NeIC) was formed in 2012. NeIC is hosted by NordForsk and includes the former NDGF project with a focus on providing resources for the Nordic WLCG Tier-1. However, NeIC has a wider scope, facilitating Nordic e-Infrastructure collaboration and joint actions in a wider international setting. The establishment of NeIC was a result of a committee with a mandate from the Nordic Council of Ministers (NCM) to set up a plan for future Nordic e-Science, and currently a follow-up activity is producing an updated version of this plan.

3.3 Major Swedish e-Science Initiatives

In Sweden there are two major e-Science initiatives, both funded as governmental strategic research areas (SRAs). The Swedish e-Science Research Center (SeRC) is a collaboration between KTH, KI, SU and LiU. SeRC has formed e-Science communities around the selected application areas in the centre and identified core areas for the development of e-Science-related tools and for the access to expertise needed to break new ground in a range of application areas. The SeRC setup is illustrated in figure 3.2, showing the core and the selected application areas and their interplay.

The other SRA initiative in e-Science is eSSENCE collaborative research programme in e-Science between three Swedish universities: Uppsala University, Lund University and Umeå University. eSSENCE is organized in a similar way to SeRC, and is focusing on both communitybuilding efforts in application areas and coordination of efforts to develop e-Science tools and methods. There are also other e-Science initiatives in Sweden but these are on a smaller scale and with less available funding, such as the Chalmers e-Science initiative.

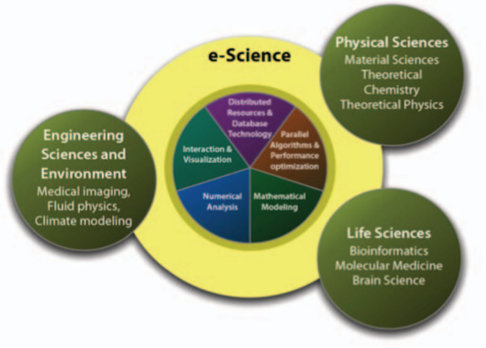


Figure 3.2: The SeRC model for structuring of e-Science. A core of methods and tools is in place, with e-Science communities formed around application areas.

What is striking in the Swedish setup for e-Science is that the funded centres are not directly coupled to the national non-domain-specific e-Infrastructure, which is primarily operated by the Swedish National Infrastructure for Computing (SNIC). It should also be noted that a small number of researchers in the Swedish e-Science initiatives successfully compete for access to unique or large-scale international resources, or participate in international collaborations with access to such resources. This is also the case for several internationally competitive Swedish research groups not being part of the above mentioned e-Science initiatives. Furthermore it should also be noted that there are domain specific national infrastructures used within and outside of the major e-Science initiatives. Examples include the Bioinformatics Infrastructure for Life Sciences (BILS) and Systems biology Infrastructure for Life Sciences (SILS).

3.4 Swedish National non-domain specific e-Infrastructure Today

It is beyond the scope of this report to give a full account of the current Swedish e-Infrastructure and, as described in the mandate and remit, the task is not to evaluate the function of existing e-Infrastructures but rather to project the future needs based on the cases presented. It is, however, important to have an overview and understanding of the current non-domain-specific e-Infrastructure to appreciate some of the comments made by the panels and to calibrate the assessments of needs that are, in some ways, based on the current resources and services provided and their past evolution. Therefore this section contains a brief overview of the Swedish system.

3.4.1 Computation and Storage

The coordination and development of the persistent national computing resources and data storage for academic research in Sweden is managed by the Swedish National Infrastructure for Computing (SNIC), which was formed in 2002 and has since supported and coordinated the resources and services of six already established computing centres at major Swedish universities, from north to south, Umeå University (HPC2N), Uppsala University (UPPMAX), Royal Institute of Technology (PDC), Linköping University (NSC), Chalmers University of Technology (C3SE), and Lund University (Lunarc). Initially SNIC was hosted by the Swedish Research Council but, as of 2012, Uppsala University has been given the task to operate the SNIC metacenter. The total level of the Swedish Research Council's funding for HPC and storage was 123 MSEK in 2013. The funding was generally channeled through SNIC and included base level funding of national SNIC resources, participation in PRACE (Partnership for Advanced Computing in Europe), as well as Nordic collaboration, through NeIC, regarding managing and storing data from the LHC.

The computing resources available through SNIC cover a full range of facilities, from what have traditionally been called "supercomputers" to commodity clusters with standard interconnects. The SNIC resources are made available to Swedish users both via traditional login access and via grid interfaces through the Swedish National Grid Initiative (NGI), which is integrated within SNIC. Access to SNIC resources are allocated via its committee, the Swedish National Allocations Committee (SNAC), which solicits applications for computer time and storage resources on a regular basis.

By the end of 2012 there were nearly 100 000 processor cores in the SNIC set-up. SNIC had a large number of systems in production (15, of which two were funded by KAW and the host institutions). Most of the systems are, however, roughly equivalent in type (Infiniband cluster) but differ primarily in CPU and Infiniband capabilities due to the different dates of of purchase. Several of the systems are heterogeneous in the sense that they contain nodes with different (thin, thick) memory configurations.

The three main systems in SNIC are presently:

- abisko (15,264 cores, HPC2N, cluster)
- triolith (19,200 cores, NSC, cluster)
- lindgren (36,384 cores, PDC, Cray)

These three systems together comprise more than 70% of the cores in SNIC.

SNIC centres also offer storage solutions and there exists approximately 1 PB of so-called "center storage" in SNIC, that is spread across the six centres.

This can be used by HPC-users (or other local scientific usage), for example for home directories, storing temporary data, etc. This is disk storage and only part of the data is backed up. According to plans the center storage will be expanded by at least three sites during 2013.

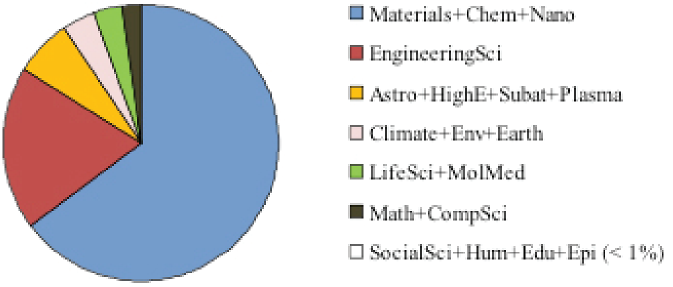


Figure 3.3: SNAC allocations for 2012 for different research areas. It should be noted that some large users of dedicated hardware such as, climate research are not included in the SNAC application procedure.

In the SNIC co-ordinated project, Swestore, approximately 1.8 PB of so-called "nationally accessible storage" is provided, physically spread across the six centres. This can be used for making data available publicly or under agreed access restrictions. The media are disk or tape. For all data there is a copy, so the usable part is effectively 0.9 PB. The accounting of usage per discipline is not available yet but most demands are coming from climate/environment and life sciences. SNIC also hosts storage for the CERN LHC experiments amounting to approximately 550 TB.

3.4.2 Data archives and Curation

The survival of digital scientific information depends on a hierarchy of constantly shifting technologies - hardware, storage media, operating systems, applications, software, security, privacy and other forms of access restrictions and database solutions. It also relies heavily on establishing and maintaining the relevant metadata structures.

The Swedish Research Council is currently funding a number of infrastructures, both nationally and globally, that aim to collect, curate and disseminate data and metadata within their specific scientific fields. On a national level, Environment and Climate Data Sweden (ECDS) and Svensk Nationell Datatjänst (SND) are providing metadata and catalogues with pointers to data sets within the fields of environment/climate and social science/humanities/medicine/health respectively. ECDS, SND and more recently also the infrastructure Bioinformatics in Life Science (BILS) also provide assistance to researchers wishing to manage, curate, archive and disseminate data properly. In addition, an increasing number of databases are funded with infrastructure grants from the Swedish Research Council. On the European and global arena the Swedish Research Council funds Swedish participation in a number of infrastructures that work with collection, curation, and dissemination of data. Some examples are The European Life-Science Infrastructure for Biological Information (ELIXIR), The Global Biodiversity Information Facility (GBIF), The Integrated Carbon Observation System (ICOS) and The European Social Survey (ESS). Although promising, the present initiatives are incomplete and do not cover all scientific areas and it can be concluded that much work still remains before researchers can easily share their data with each other in an easy and safe way. It should also be noted that, as many challenging research questions become even more multidisciplinary, the need for cross-disciplinary access to data increases and, with that, the need for support for different forms of access in a variety of formats.

Dissemination of data and open access to scientific results has become a high priority for the Swedish government (and many other other governments as well as the European Commission) and the Swedish Research Council has been asked to complete two major assignments. The first task concerns a national policy for open access to scientific results, including both publications and data, where the Swedish Research Council will work with other stakeholders, aiming to propose a national policy in 2014. The second assignment concerns implementation of an infrastructure for registry-based research, in order to facilitate access to registry-based data and provide expert advice on the associated legal issues. Access, analysis and storage of data that is based on individuals is heavily regulated in order to protect the personal integrity of the subjects. Finding a way to share such data remains a major challenge.

In addition to data being collected or generated through public funding, many other organizations also collect and make available highly valuable data for research purposes. Such data often comes with various degrees of access restrictions that need to be managed through proper e-Infrastructure tools and processes.

3.4.3 Networking

Networking is, as described above, one of the key components of an e-Infrastructure. The Swedish University Network (SUNET) was formed in the early 1980´s as a research and development project for the Swedish computer scientists and paved the way for the Internet in Sweden. Today SUNET serves mainly the affiliated universities by providing an infrastructure for national and international data communications together with a variety of data services. SUNET is the facility through which Swedish researchers access international and other national e-Infrastructures and their associated resources and services. It is also the facility through which various forms of electronic collaboration, such as videoconferencing, take place.

Since 2001 the Swedish Research Council has held the overall responsibility for SUNET, but it is mainly the SUNET-affiliated organizations that finance its operation. Some of the activities are, however, funded by the Ministry of Education and Research through the Swedish Research Council.

SUNET has been one of the leaders in academic networking world-wide and has inspired the roll-out of several internet operators. The most recent implementation of SUNET is named OptoSUNET and delivers redundant and diverse connections at 10 Gbps to all major Swedish universities. OptoSUNET is a also well connected to international networks and provides Swedish researchers with access to international collaborations.

Figure 3.4: The OptoSUNET network consists of leased fibre, and is divided into three separate systems: north, west and south. Each of the systems have two separate networks called "red" and "green" such that each site has at least two access paths. Image: SUNET, Börje Josefsson.

In addition to normal network access with routers, OptoSUNET can provide point-to-point connections without routers to transmit large amounts of data directly between two points through a wavelength service, which is a service of increasing importance to large e-Science projects. This service can be provided both nationally and internationally through the NORDUnet collaboration. NORDUnet connects directly to the pan-European GEANT and the US Internet2 networks. SUNET also provides a dedicated 10 Gbps link from Stockholm to Frankfurt as part of the PRACE infrastructure and in support of the LHC project. SUNET also provides a 10 Gbps connection from the Onsala Space Observatory to SURFnet (the Dutch equivalent of SUNET) via NORDUnet, to a supercomputer in Dwingeloo serving as a collator of telescope data.

3.4.4 Visualization

Even though visualization is a natural component in the workflow in many areas of e-Science, there are currently no organized visualization services provided on the national level in the Swedish e-Infrastructure. There is, however, some limited access to visualization servers and experts provided by some of the SNIC centres. Within the e-Science programs there are research groups providing possibilities for collaboration with visualization experts to the e-Science communities.

3.4.5 User Support for SNIC resources

The computer centres are providing standard help desk support and limited application expert (AE) support using staff based at the centres. It is clear, however, that many users and groups of users of the e-Infrastructure are providing their own user support by hosting and hiring research engineers (RE) with the task of supporting the users of the local and national e-Infrastructure. Nationally there has also been an increase in the number of AEs affiliated with SNIC during 2012. The recruitment of the new AEs was synchronized with the e-Science centres (SeRC, eSSENCE, Chalmers e-Science Centre). There are currently 14 FTEs working as AEs at the SNIC centres.

There is also work on establishing a mechanism where researchers can submit requests for application support that can be evaluated with the help of the Resource Allocation Committee. As such, application support - in the form of person months - can be allocated as a resource, like cycles and bytes. Calls for advanced user support can then be synchronized with calls for applications for computing time and data/storage. This allows users to apply for any combination of computing time, storage and support.

References

[G+12] M Guest, G Aloisio, R Kenway, et al. *The scientific case for HPC in Europe 2012-2020*. Technical report. PRACE, October 2012. http://www. prace-ri. eu/PRACEThe-Scientific-Case-for-HPC, 2012 (cited on page 19).

[HTT09] Tony Hey, Stewart Transley, and Kristin Tolle. *FOURTH PARADIGM*. Microsoft Research, 2009 (cited on page 15).

[Woo+10] J Wood et al. "Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data-European Commission". In: *European Union* (2010) (cited on page 17).

4. COMPUTER SCIENCE AND MATHEMATICS

- *Testbeds are critical for many aspects of CS&Math R&D, some of which cannot be "emulated" through a local set-up, some must be connected/integrated with real-time data collection systems.*
- *To maximize the benefits of CS&Math research as well as to enhance productivity of CS&Math research there must be a professional, sustained pathway from research outcomes to community software and tool´s*
- *Education and training is the primary pathway through which CS&Math research outcomes are adopted by a broad range of user communities. A successful E&T effort requires significant and sustained resources.*
- *There is a strong need for community platforms that provide infrastructure to manage security, privacy and proprietary concerns to enable public R&D to undertake collaborative activities with commercial and other entities whose business is not public.*

In this chapter the findings of the panel on computer science and mathematics are presented. As this is an underpinning and enabling science the presented science cases emphasize the contributions that this field makes to other disciplines as well as the e-Infrastructure needs for CS&Math R&D.

Computer Science and Mathematics develop methods, models, theory, formal languages, algorithms and a wide range of software for systems configuration, operation, management, software development, debugging and optimization and for use by other disciplines and human-computer and computer-computer interaction. Application software is typically developed and maintained in cooperation with users of e-Infrastructure and its various components, or entirely by the users.

Computer Science is closely related to engineering in that it addresses the application of scientific knowledge and discoveries to technologies and products useful in human endeavours, including research. Technology, in particular computer and information technologies, have become pervasive in today´s society and accounts for a measurable fraction of mature economies (more than 20% of GDP growth and more than 3% of GDP, a significant job growth with more than 2.5 jobs created for every job lost, with 75% of the impact being due to traditional businesses according to several studies).

Computer and information technologies are changing at an unparalleled exponential rate with the capabilities having increased more than 100-fold per decade for the last 40 years at constant or even slightly decreasing cost . The exceptional evolution also encompasses digital capture of data leading to a data explosion due not only to an increasing fraction of data capture being digital, but the resolution in space and time increasing rapidly. Human created content is now largely in digital form with the use of video of increasing resolution becoming mainstream in many contexts. High resolution digital imaging has enabled molecular electron microscopy imaging to reach resolutions close to 1 Angstrom and the cost of gene sequencing to approach USD 1,000 for the human genome (with a single sequencer today outputting about 1 TB/day) from $3 billion for the first sequencing of the human genome just a few years ago (see chapter 8). Similarly, in astronomy the digital evolution has enabled captured data to grow to Petabytes of data yearly for some instruments (see chapter 6). The digital revolution has also resulted in the ability of companies to collect data, in some cases Petabytes of data annually, about their products and services and about consumer behaviour and, through data mining and analysis techniques, enhance their products and services and better reach and serve their consumers. The widespread deployment of video cameras for surveillance, the widespread use of social networks, and digital communication surveillance have led to large demands for storage, processing and methods for generating knowledge from a diverse set of sources, Communication systems, including global networks, have experienced a growth rate of their capabilities that even surpasses that of computer servers, and have a rapidly increasing reach. Today´s communication networks enable effective collaboration and sharing of resources and data regardless of the locations of the participants in the developed world and, increasingly, also in the developing world, and collection of data (continuously) from almost any location. Computer networks are fundamental to e-Infrastructures and integral to the science cases presented in this report with some depending on dedicated high-speed networks for access to unique resources, sharing of data, or creation of virtual instruments.

With the exponential changes in the capabilities of the e-Infrastructure, including its expanding reach, new problems come within reach and disciplines that traditionally have been small or modest users of computer and information technologies have been presented digital “opportunities” of an unprecedented magnitude, for instance through the emergence of internet social networks harnessing vast amounts of data (for example 72 hours of video are uploaded to YouTube every minute, and there are about 0.5 billion registered Twitter users). These vast amounts of digital data not only require vast amounts of storage, but place heavy demands on communication networks for the sharing of data and on processing capacity, with processing needs for many applications growing much faster than the size of the data sets. The sharing of data provides the opportunity to generate new insight and improved decision-making by using data from many disparate sources, as exemplified by IBM's Dr Watson which offers diagnoses and treatment plans based on vast amounts of medical data and specific patient information largely from unstructured data.

The dramatic changes in capabilities and related opportunities in most cases have required new approaches (methods, models, algorithms, software, etc.) for their effective exploitation. Computer Science and Mathematics have been major contributors to the required changes and, of the experienced gain in application capabilities in science and engineering, advances in methods, algorithms, software etc. have contributed as much as hardware gains. Visualization is likely to grow in importance to generate knowledge from large and complex data sets. Much of the digital information generated today is created outside of public research but is, nevertheless, of great interest and importance for research. The security and privacy challenges associated with medical and health information are well known. In many industries such as, for example, insurance, finance, pharmaceuticals and energy exploration, much information is considered proprietary but some of it can be shared with research communities in restricted ways. Methods and tools for tiered, secure, access to information and software in open environments are also required by many research communities. These challenges represent significant opportunities for the Computer Science and Mathematics communities.

The success of the microprocessor as a commodity, inexpensive processor, not only made computers affordable even for private use in developed economies, but also came to dominate high performance and large scale computing by employing hundreds to tens of thousands of commodity microprocessors linked together into a single system. Parallel computing came to dominate computing in the sciences and engineering about 25 years ago and required new programming paradigms, software, tools and sometimes methods and algorithms and a change in how computational scientists and engineers were educated and trained. The dominating technology for processors and memory for the last 40 years or so, CMOS, has limitations that, during the last decade, have had a dramatic impact on microprocessor design and software. Multi-core processors have been the industry´s response to CMOS technology limitations and the demand for exponentially increased capabilities, driven by historic trends and "Moore's law". Multi-core processors have become the norm in every device, from cell phones to tablets to laptops and servers for large-scale computers. The key limitation causing this change was the ever increasing power consumption of microprocessors, and the challenges in removing the associated heat. For the last decade, successive microprocessor generations have been designed with heat dissipation targets being flat.

The multi-core era has triggered a new change in programming paradigm and associated tools. The increased ability to fit more transistors on a chip, together with improved design methodologies and tools, has also contributed to improved specialized designs, like Graphic Processing Units (GPUs) becoming less specialized and taking on a role as cost-effective "accelerators" for floating-point intensive applications. This change has again impacted programming methods, paradigms, tools and algorithms. The power consumption of processors that drove the change to multi-core processors has become one of the primary challenges for the computer industry, including the memory industry and increasingly also for the computer networks industry.

The electrical energy consumption of the computing and information technology industries has surpassed that of the aviation industry and is expected to reach 5-6% of all electricity consumption by the end of the decade. The life-time cost of electricity for operation and cooling of computer systems has surpassed the capital cost and large data centre electricity consumption has approached that of heavy industries such as steel and aluminium. As an example, the Facebook data center in the north of Sweden is planned for a load of up to 120MW, corresponding to the average power consumption of about 40,000 single family Swedish homes using electricity for heating. This trend is not sustainable. There are several proposed approaches towards reduced energy consumption, including operating transistors at near or below "threshold" level, architectural changes leading to simpler, less energy consuming cores including heterogeneous processors with several functional units/cores optimized for different tasks, dynamic power and performance management as a function of the workload ("energy proportional computing"), and using more cores operating at lower frequencies since the power consumption for CMOS grows faster than linearly with clock rate. These changes are expected to have an even more profound impact on programming methodologies, algorithms and software than the changes caused by the emergence of clusters and then multi-core processors.

The consensus is that computer systems for the foreseeable future will become more complex rather than simpler, and that large scale systems will have tens of millions of cores or more, that systems will be heterogeneous and dynamically managed for power consumption during application execution, and that there may be a divergence in architecture to more effectively address the needs of different workloads. The use of e-Infrastructures is likely to play a key role for data driven research, similar to what has been the case for simulation based research for many years [JH1]. E-Infrastructure, by its very nature, is distributed and the trend to integrated/federated global infrastructures is likely to continue. With this vision we have selected three science cases to exemplify the needs of research in Computer Science and Mathematics: Community Platforms, Method and algorithm development, and Testbeds. In addition, we highlight the educational needs.

4.1 Community Platforms – Sharing of data and software

Today many research projects not only result in novel findings that are well-documented in scientific publications, but also often in data and software that, to a lesser extent, are archived for use in future research. In some cases, data can only be released and used under certain conditions and, hence, are not well suited for being provided as supplementary material together with the publications. In other cases, the resulting data and software may be far more comprehensive than can be covered by a single publication. Currently, each research project has to find its own solution by, for example, maintaining data and software repositories, but once the funding terminates or key personnel leave, there is an obvious risk that large parts of the results cannot be taken advantage of in future projects.

Research community platforms, allowing for data and software storage and sharing, could be a long-term solution to this problem. Such platforms would also promote the integration and exploitation in "end-to-end" usage scenarios in community settings to accelerate future research projects. They would also contribute to meeting the directive on the re-use of public sector information, which regulates how public sector bodies in the EU should make their information available for re-use. As the amount of collected digital information is increasing at a very rapid rate, the storage cost therefore has been increasing despite the exponentially decreasing cost per unit of data stored. The management and effective use of data is growing in significance from both a value and a cost point of view.

Another aspect of data sharing is access to "raw" and/or derived data that have some proprietary or privacy aspects associated with them. It is important that methodologies and tools be developed for the research communities to be able to work with and share such data in an efficient way and minimize the number of, and effort associated with, case-by-case solutions.

Accessibility: One of the most basic requirements of a research community platform is to allow access over the Internet. In order to allow for tiered access to data and software that wholly or in part should not be publicly accessible, for example when certain conditions of use apply, authentication and authorization will be required. To ease the access to stored resources, methodologies and tools should be developed and made available to the research communities to minimize the amount of duplicate work, manual processing and to support audit of use and enforcement of policies.

Sustainability: Research community platforms should guarantee sustainability, meaning that all data and software that have been made available and archived through the platform should be stored in such a way that they can be accessed and executed within the foreseeable future. This not only places requirements on continuously updating storage media, but also that data and software are kept in standardized formats that can be processed in the future, for example the Open Document Format, and ported to new formats and standards whenever desirable or necessary.

Security: In order to allow for authorized access to data and software for authenticated individuals, security protocols have to be implemented and maintained, and policies updated and enforced. Highest standards for securely transmitting and storing both data and software must be employed to enable handling of highly sensitive information, for example personal data and software of commercial interest. Research community platforms should provide means to securely communicate and share results within "members only" groups sometimes referred to as Virtual Organizations (VOs), for example providing access only to project members, which may include non-academic partners.

Integration: Research community platforms should support integration of data from several sources and in different formats, as well as integration of software components that have been archived under terms allowing use for the intended purpose. To allow for such integration, platforms should require, and support the development of, standardized documentation of programming interfaces and data protocols. The platforms should also allow archiving of results from analysis of the data in standard ways, for example using Predictive Model Markup Language.

Curation: Tools for checking the validity and conformance of data with metadata specifications should be provided within the platforms. The platforms should also provide tools and support for documenting and annotating archived data and software.

Foreseen benefits: Potential major innovations through access to more data, in particular in areas that are especially valuable for the major research and societal challenges and which tend to be multidisciplinary in nature. This should result in significantly enhanced productivity of the research communities (not just Computer Science and Mathematics).

4.2 Model, Method and Algorithm Development – Interactivity

Supercomputing resources are, today, critical for method development in computer science and mathematics (such as numerical methods for differential equations, linear algebra, optimization, statistics, data mining, optimal control, resource efficient algorithms, programming models and tools), visualization and data analysis (computer graphics, machine learning, computer vision, etc.), and for applications in mechanics, biology, physics, chemistry, image analysis, speech recognition, and others as exemplified by the other panels.

Science and Engineering research is increasingly data driven, with huge amounts of data being generated by instruments for the study of fundamental physics (such as the Large Hadron Collider, discussed in chapter 6), Astronomy (for example the Square Kilometer Array, discussed in chapter 6), Chemistry (for example in molecular imaging), Life Sciences (for example gene sequencing and proteomics, discussed in Chapter 8), Geophysics (as in hydrocarbon exploration), and product lifetime information in, for example, the auto and aircraft industries. Retail businesses, banks and the internet companies today have repositories with up to tens of Petabytes of data which they use for various planning purposes and understanding of human behaviour. Storage and analysis of such vast amounts of data represent significant challenges and opportunities for the Computer Science and Mathematics communities with the simultaneous use of disparate sources of information potentially leading to significant discoveries, and increasingly being used in security and criminal investigations, but also in solving complex problems, such medical diagnosis of non-trivial cases as illustrated by IBM´s DrWatson. Interactive visualization and computational steering, even in distributed settings over large distances, is emerging from research to "production" (for example VisIt , Paraview) and opening up new possibilities in computational mathematical modelling, method and algorithm development as well as new modes of human-machine interaction. Interactive visualization, computational steering, and interactive model development will be of key importance in enhancing research in many areas. Interactivity at this level places new demands on the e-Infrastructure (hardware and software), with an integrated environment for communication, computation, and data, including possibly data acquisition and assimilation. Computational Steering: Interactive, large-scale, computer simulations with real-time (remote) visualization in which parameters of the simulation model (such as the geometrical model, constitutive parameters, etc.), data and algorithms can be accessed and modified during the simulation session will require a significant change in how e-Infrastructure resources are designed, operated and supported. This interactive workflow model of instant feedback is a significant shift from today´s use of shared large resources where the feedback loop between user and simulation or analysis of results may be hours, days or even weeks. The envisioned interactive workflow models create opportunities for new "fast science". A concrete example of such interactive use of large-scale data analysis is in real-time imaging of proton treatment in which backscattering can be used for imaging and control of the beams so that the target is hit with maximum energy and with the minimum amount of damage to surrounding tissue, hence improving both treatment outcome and quality.

Interactive Model Development: An e-Infrastructure facilitating simultaneous access to disparate data sources, computation for analysis and simulation, visualization, validation and uncertainty quantification can lead to new discoveries and enhance quality of outcomes, as well as reducing the time between observations and results. The benefits of such integration have been demonstrated by both increased accuracy and reduced time in severe weather prediction enabling enhanced emergency preparedness. Data assimilation (integration of observational data with simulation/modelling data), validation, and uncertainty quantification are becoming increasingly realistic and enable computer simulations to become an effective tool in tackling many of the great challenges to our society, such as the development of new technology for sustainable development (for example in environmental risk analysis), e-health, epidemiology (such as in studying the rapid spread of viruses like "bird flu"), and emergency management.

Example in Medicine: Visualization (for example in medical imaging) is today extensively used in clinical practice as a tool for decision support in diagnosis and treatment planning. With today´s technology for simulation in biomedicine, completely new possibilities open up, for example in the preoperative assessment of surgery where different postoperative scenarios can be evaluated. This approach has been used to assess the risk for rupture of aneurysm surgery [Ceb+05], for example. With the continued exponential growth in capabilities expected of computing systems, together with improved models and algorithms, interactive computational assessment in clinical practice should be possible for a wide range of medical treatments including more wide spread use in reconstructive surgery.

The technology used in preoperative assessment is similar to that of virtual prototyping in the vehicle industry, or risk analysis in construction engineering. Virtual prototyping of, for example, prostheses (from hand prostheses to mechanical heart valves), based on patient specific data, would enable highly customized individual patient prostheses.

The ability to interactively test different scenarios in a preoperative virtual surgery setting would offer a break-through technology with the potential to improve clinical diagnosis and treatment, improving quality of outcomes and possibly saving lives. Such use would put very high demands on validation and uncertainty quantification of simulation models. The e-Infrastructure would need to supply possibly extensive computational resources, on interactive timescales, for simulation, data analysis and visualization in an integrated environment in which patient-specific information is assimilated into the simulation and enabled for user interaction (steering) by the medical professional and, possibly, by patients.

4.2.1 Potential breakthroughs

Interactive use of distributed resources, including supercomputers, databases with a variety of relevant content for multidisciplinary problems, unstructured data, and rich human-machine interaction media (such as 3D, haptics, audio and speech). These achievements should, in turn, enhance the scope and rate of innovations in other disciplines as well as increasing productivity across a broad range of disciplines making use of e-Infrastructure.

4.3 Scientific Test Beds

The Computer Science and Mathematics Communities require test beds covering not only different processor and system architectures but also storage system architectures, visualization systems, computer and communication networks and distributed environments for the development of models, methods, algorithms, programming paradigms and languages, software including operating systems, compilers, libraries and program development and optimization tools, system administration, storage and resource management tools and tools that support various forms of user interaction with the e-Infrastructure (job submission and monitoring, result retrieval, interactive execution and debugging, domain specific portals, etc.). Different tools require different forms of test beds: some a diversity of processor architectures, other different forms of I/O systems, or different forms of display (visualization) systems, or communication technologies. A particular challenge is scalability that may require extensive test beds, and methodologies, tools, algorithms and software for distributed environments, such as clouds and grids, and for combining disparate resources into an integrated or federated environment. Distributed environments is common in many industries with virtual machine technologies being used for dynamic resource allocation in the distributed environments, with the Internet companies being the most visible in exploiting this technology. But, distributed environments are increasingly also used for research e-Infrastructures with some dynamic provisioning of resources being used at least experimentally. PRACE, EGI, and XSEDE are examples of distributed e-Infrastructures for research.

Current and future computing systems, in particular high-performance computing systems, have complex architectures that include a high and growing degree of parallelism that is managed through "hybrid" programming models such as mixed OpenMP, hardware accelerator control and MPI (Message Passing Interface) code, increasing heterogeneity, and an increasingly complex memory hierarchy [Dur+13]. It is fairly common for applications to achieve less than 10% of theoretical peak floating-point performance. In order to obtain reasonable performance, codes must be heavily optimized. However, manual tuning of programs is neither possible nor economically reasonable. Also, the increasing degree of specialization in both application domain and computing system sciences makes it ever more difficult to combine the necessary domain and system skills necessary for developing powerful algorithm adaptations. High-level programming approaches such as PGAS (Partitioned Global Address Space) languages, domainspecific languages, skeleton programming frameworks and advanced software tools need to address this issue. The development of such tools requires access to a diversity of architectures covering the essential characteristics of the platforms available to users of the e-Infrastructure. Though some development can be carried out in a production environment, some aspects of the development required may jeopardize the stability of the production environment and other development simply cannot be done at all in a production environment, such as operating and file system development and testing, and many aspects of resource management systems. Dedicated test beds are, therefore, essential.

Though many applications may be dominated by the effective use of the processors in the system, others may be limited by the I/O system and associated file and hierarchical storage management systems. Several of today´s storage and file systems do not scale well, and have severe performance problems in handling large numbers of small files, which are common in, for example, gene sequencing. New storage system architectures are required and correspondingly optimized file systems. The integration of storage systems into clusters and MPPs (Massively Parallel Processors) can have a significant impact on performance and cost. Development of storage and file system technologies and tools for monitoring and optimization will require dedicated test beds.

The increased emphasis on energy efficiency has caused both the HPC industry and the Internet industry to investigate the use of commodity processors other than "x86" processors for server platforms, specifically processors dominant in the embedded and mobile markets and that may only require 10%, or even less than 1% of the power of a typical "x86" processor. The impact of the difference in architecture and capabilities on applications, and proper instrumentation and measurements of delivered energy efficiency cannot be carried out in a production environment and hence also require test beds. Such instrumentation and measurements are an important step in moving the industry forward with regard to effective use of platforms from an energy perspective, in much the same way as the HPC community´s demands for more information about efficiency of codes led the processor and platform industries to m ake more performance information available. Significant steps by the industry were the making available of register information related to performance, and the introduction of additional counters, initially reluctantly, but now commonly available and which form the basis for performance monitoring tools such as PAPI .

Cloud computing has been adopted by several research communities, in part because the service model offered by computing centres serving academic research is perceived as cumbersome and not well adapted to their needs. A cloud computing model should be part of an e-Infrastructure for research and should be supported through the development and provision of tools that make the sustained and shared use of the cloud computing model easy and efficient. The development of proper tools and software environments for managing software and data require proper test beds.

Development of computational tools for solving important problems in science and technology (as well as in social sciences and finance etc.) adapted for modern computing systems requires the development and analysis of computational models and fundamental algorithms (discretization, optimization, numerical linear algebra etc.), and research in the area of programming systems (including programming languages, libraries and software frameworks, compilers, runtime systems, parallelization, performance analysis and modelling tools, data management, visualization tools, and middleware). If the results are to come to wide use it is necessary to bridge the widening gap between application-level software and modern HPC architectures and computing environments.

Generally, for research and development of algorithms and tool infrastructures it is important to have access to "components" and systems early in the technology development, ideally before the technology is ready for production use, with direct (interactive), frequent, and possibly exclusive access. While the time required for a single job for development, testing, debugging and benchmarking is often quite short, there is an immediate need for the tests and many may need to be run without interference from other users. The need for such early access is also recognized by the manufacturers who often seek community engagement for enhancing the software environment and revising the architecture to pave the way for a successful product, an approach taken by, for example IBM for the Cell processor, Intel for the Xeon Phi, and NVIDIA. It is important that the e-Infrastructure providers engage in this activity since they have knowledge of a wide range of applications and need to develop the knowledge, management tools, and user environment, and need to train staff to deploy and support the new technology. This need has been realized by several of the PRACE partners who, through future technologies work packages, have engaged in technology assessments with regard to performance, energy efficiency and software impact.

Example: Research on tools for auto-tuning. As a concrete example, the algorithm engineering and optimization task is increasingly delegated to automated performance-tuning tools that combine design space exploration and machine learning techniques with machine-specific code generation to automatically adapt and optimize the code for a given algorithm. These techniques can, for example, be integrated in domain-specific library generators (such as ATLAS , FFTW or SPIRAL), compilers (such as CAPStuner), and software composition systems (such as the PEPPHER framework [Ben+11]), thereby making the automated tuning machinery accessible to application-level programmers and to user-defined program structures. Research on programming environments, tools and compiler techniques that are needed for efficiently mapping HPC applications to modern HPC computing systems has thus become increasingly important in the last 10 years. Several Swedish universities have recently been involved in national and international research projects, such as the FP7 projects PEPPHER (www.peppher.eu) and ENCORE (www.encore-project.eu), which are concerned with the design and implementation principles of programming environments and tools that allow the level of abstraction for programmers to be raised and, at the same time, support automated performance tuning and performance-portability.

Example: Numerical Linear Algebra. At Umeå University´s Computer Science department, there is an ongoing, long-term effort for developing, analysing and implementing scalable and robust algorithms and software, aimed at massively parallel computer systems. This work requires access to modern parallel computer systems, and is important for retaining efficiency in numerical linear algebra libraries as computer architectures evolve.

Example: Computer vision algorithm research. An example in the algorithm development field comes from the computer vision research group at Linköping University. In their development of new computer vision algorithms they are regularly faced with the problem that very high computational capacity is needed, for example for testing or for comparative experiments. Without a high-level computational framework with the necessary software components, using a central supercomputer is not an option. Another issue in computer vision is the huge data sets that exceed the capacity of the institutional file servers. For example the uncompressed raw data from 6 color cameras produced at a frame rate of 16fps over several hours should preferably be stored centrally before processing; some of these could well serve as benchmarks for comparative experiments so they should be made available to other research groups via a web server. However, some institutions put a 50 GB limit on content for their web servers.

Example: PDE solving environments There are several high level PDE solving environments, for example Deal II [BH07], FEniCS [LMW12]. These enable efficient code development for testing of different computational models for partial differential equations as in, for example, computational treatment of surface tension in two-phase flow [ZKK12]. These environments include many different options for discretization, and allow easy access to linear algebra libraries. In this example both the PDE environment and the linear algebra libraries are infrastructure, and need to be kept up-to date to run efficiently on modern heterogeneous large scale computer systems.

4.3.1 Potential breakthroughs

An improvement in energy efficiency of computing systems by an order of magnitude beyond what can be expected from "Moore´s Law" which, if realized, will bring the energy costs for computing systems and their cooling down to a modest fraction of the total cost of ownership as opposed to the current situation. Test beds should also enable tools for improved usability and programming productivity and avoid a potential setback in these regards due to expected changes in the architecture of computer and storage systems.

4.4 Education and Training

In mature usage scenarios, including hardware, algorithms, software, data representations and metadata, standards may have evolved to a point that a "black box" scenario applies but the very nature of e-Infrastructure for research makes such a scenario the exception rather than the norm. Hardware, methodologies, models, algorithms, software, data representations, communication modes and capabilities are subject to constant change, sometimes causing entire paradigm shifts rather than just incremental changes. This change mandates a concerted effort on behalf of the infrastructure to continuously educate and (re)train its user communities on best practices, in concert with those communities, and to facilitate the integration of new methodologies, algorithms and programming paradigms into standard curricula. The effective use of an e-Infrastructure for research involves a substantial number of practical issues that are not a natural part of the academic curricula and hence the normal processes through which research results get integrated into common curricula are not sufficient.

There are several target audiences for education and training with regard to e-Infrastructure and there are several mechanisms that can be used for each audience.

Train-the-trainers: Clearly, attempting to educate and train all current and potential users of the e-Infrastructure would require a massive effort, not only for the delivery and support of the training effort of the many highly relevant aspects of the e-Infrastructure, but also for producing and maintaining training material. Realizing this objective is not possible without an effective "train-the-trainer" program. E-Infrastructure resource providers should collaborate and take the lead in this type of effort. The e-Infrastructure resource providers are in the best position to take on the responsibility for this type of effort since they have the primary knowledge of the e-Infrastructure and its main interrelated parts and dependencies. An effective implementation of a train-the-trainers program requires close collaboration with disciplines using the e-Infrastructure or which could, potentially, benefit from it.

Education and Training of Active researchers: Education and training of active researchers should be organized differently from training of trainers. Active researchers are likely to already have experience of the e-Infrastructure or use of resources similar to the ones available in the e-Infrastructure, or have good knowledge of models, methods, algorithms, software, data, etc. within the domain of research, but they may need to learn how to best use the e-Infrastructure and its associated services, or learn of alternative or new approaches better suited to the specific e-Infrastructure.

Integration into academic curricula: For maximum impact, knowledge of the e-Infrastructure resources (hardware, software, data, services), how to access and use them effectively and associated methodologies, models, algorithms, programming systems etc., should be integrated into standard academic curricula, adapted as needed to the many disciplines using the e-Infrastructure. Though academic disciplines are responsible for their curricula, e-Infrastructure resource and service providers should facilitate this integration. For example, education in parallel programming was initially predominantly offered by academic computing centres through "Summer Schools" or code porting workshops but is now, at least for commonly used parallel programming paradigms, integrated into many standard academic courses in many disciplines at the undergraduate level. Change is, however, a constant in e-Infrastructure, and new material continuously needs to be introduced into academic curricula to ensure that the next generation of e-Infrastructure users are well prepared to effectively use and enhance it.

There are many known and well established approaches to effective education and training with their own advantages and disadvantages. Most are relevant for education and training the e-Infrastructure´s effective use and evolution.

Seasonal Schools ("Summer Schools"): These are common and have been found effective in many disciplines including the education and training of users and developers of e-Infrastructures. Seasonal Schools have been used very successfully by, for example, EGI, PRACE, and XSEDE in regards to e-Infrastructure and by HPC centers with, for example, PDC having trained over 1,000 researchers in its 2-week Summer School, entitled "Introduction to HPC".

Domain specific workshops: These enable an exposure to more advanced material with a focus on material that is of specific interest to a community, or the effective use of particular resources, software, data sets, or code porting techniques, etc. Like Seasonal Schools, domain-specific workshops have been successfully used in regards to e-Infrastructure by EGI, PRACE, XSEDE and others and by individual resource provides, such as HPC centres. Web based education and training: Though seasonal schools and workshops are very effective, and offer intangible benefits such as potential new collaborations (possibly across disciplines), they also have their limitations with regard to accessibility by being available only once or a few times a year and at a few locations and with attendance limitations. Therefore, web-based training should be made available, both in an entirely self-learning mode and in an assisted mode.

Collaboration: Knowledgeable and effective trainers are in short supply and developing quality training materials is very time consuming. Typically the needs are common to many e-Infrastructures, resource providers and communities and, hence, development of training materials and the delivery of training in the form of Seasonal Schools, Workshops, and the Web can be carried out as cooperative efforts. For instance, PRACE and XSEDE cooperate in both these aspects and PRACE is establishing a few advanced training centres focused on the development of training materials and cooperative delivery of training at sites other than the centre.

Outreach: An e-Infrastructure serving the Swedish research communities is a significant resource and investment. Therefore, an outreach effort should be associated with it that not only ensures it is used to its full potential for research, but that its potential to attract new users, the next generation researchers that will move research forward and enhance the society, and generate public good is exploited. To serve such purposes the XSEDE "Training, Education and Outreach Service" (TEOS) works with campuses and organizations to help instil digital services into the practices of faculty, staff and students. Activities with colleges and departments include: providing support for incorporation of computational science resources and methods into the curriculum; recruiting and training Campus Champions to raise local awareness of digital services; helping campuses and organizations to enhance their cyberinfrastructure resources; and visiting sites to provide digital services needed by the institution´s personnel. To deeply engage the community in sustained use of computational services, the TEOS team also works closely with individuals to understand and address their particular needs and challenges. These individuals may be faculty, researchers, students, IT staff, or administrators. Deep engagement activities include training, workshops, internships, fellowships, and consulting sessions designed to address the unique needs, perspectives, and practices of diverse communities. PRACE also has extensive outreach efforts and, in addition, has established a PRACE Young Investigator Award to stimulate the interest of young researchers in the innovative use of its infrastructure. These efforts are viewed as effective.

4.5 Recommendations

Infrastructure by its very nature is long-term, even if many components and technologies in the case of e-Infrastructure have very limited life-times. Funding strategies need to be found to ensure the longevity of e-Infrastructure and that appropriate returns are seen from the considerable investments made in the resources it houses and the human capital required for its evolution, support, operation and maintenance.

A holistic approach to the overall efficiency of the public research enterprise, of which e-Infrastructures are, or should be, an essential component, should be developed to ensure long term competitiveness and maximize societal benefits. Short-term tactical approaches should not jeopardize long-term success and benefits.

- A Swedish platform(s) for information and software, most likely different platforms for different communities, with most also accessible by the public at large, should be developed as part of the e-Infrastructure.
- Platforms for secure, tiered access to resources (software, data, etc.) requiring access privileges for privacy, commercial, security or other reasons should be developed and be part of the e-Infrastructure.
- Interactive access should be developed to much of the e-Infrastructure for improved productivity and enhanced research outcomes.
- Support for integrating and maintaining diverse resources into environments, as needed by the research communities should be an integral part of the e-Infrastructure.
- Support for turning research outcomes (data, software, etc.) into entities usable by the research communities when needed (and if there isn't any or sufficient commercial uptake) should be a service included in the e-Infrastructure.
- Support for evolving and improving the e-Infrastructure in the form of testbeds necessary to develop, test and validate the next generation of e-Infrastructure technologies (hardware, software, environments, etc.) and services should be an integral part of the e-Infrastructure.
- Education, Training and Outreach should be an integral part of the e-Infrastructure to maximize its impact and benefits.

Panel Members

- Professor Lennart Johnsson, (Chair), Computer Science, Mathematics and Electrical and Computer Engineering, University of Houston. Expertise: Computational and data Grids, high-performance scientific computation, parallel algorithms.

- Professor Henrik Boström, Computer and Systems Sciences, Stockholm University. Expertise: Machine learning, data mining, knowledge discovery, predictive analytics.
- Professor Johan Hoffman, Computer Science and Communications, Royal Insitute of Technology. Expertise: high performance scientific computing, computational mathematical modeling, uncertainty quantification and data assimilation.
- Professor Christoph Kessler, Computer and Information Science, Linköping University. Expertise: parallell computing, modeling and performance optimization, compiler technology.
- Professor Gunilla Kreiss, Information Technology, Uppsala University Expertise: numerical methods for partial differential equations, hyperbolic systems, numerical simulations.

References

[BH07] W. Bangerth and R. Hartmann. "Deal.II - a General Purpose Object Oriented Finite Element Library". In: *ACM Trans. Math. Softw* (2007) (cited on page 33).

[Ben+11] Siegfried Benkner et al. "PEPPHER: Efficient and Productive Usage of Hybrid Computing Systems". In: *IEEE Micro* (2011) (cited on page 33).

[Ceb+05] JR Cebral et al. "Characterization of Cerebral Aneurysm for Assessing Risk of Rupture Using Patient-Specific Computational Hemodynamics Models". In: *AJNR Am J Neuroradiolgy* (2005) (cited on page 30).

[Dur+13] M. Duranton et al. *The HiPEAC vision for Advanced Computing in Horizon 2020*. 2013. URL: www.hipeac.net/roadmap (cited on page 31).

[LMW12] A. Logg, K-A. Mardal, and G. Wells. *Automated solution of Differential Equations by the Finite Element Method*. 2012 (cited on page 33).

[ZKK12] Zahedi, Kronbichler, and Kreiss. "Spurious currents in finite element based level set methods for two-phase flow". In: *International Journal for Numerical Methods in Fluids* (2012) (cited on page 33).

5. CLIMATE, ENVIRONMENT AND EARTH SCIENCES

- *To enable Swedish scientist to keep up with the forefront of research based on Earth System Models, HPC resource increases of the order of factor five are needed for the next five years.*
- *Efficient storage with access and analysis capabilities to the huge output data that the community are producing are as important as CPU hours.*
- *The need for organized and distributed databases for mainly observational data involves basically all established and new user communities.*

Research and development within climate, Earth and environmental science are of highest importance for future planning and policy making. Many societal challenges, on both shorter and longer time-horizons, are related to climate and climate change and its impact on, for example, weather, floods and droughts, other natural hazards, biodiversity, ecosystem services, and food safety. Other areas of major societal importance are energy, raw materials and mitigating the effects of major natural hazards such as earthquakes and volcanic eruptions. The climate system involves complex interactions of the atmosphere, the oceans, the cryosphere, the biosphere, and the solid Earth. To be able to provide insight into the response of this complex system to changing conditions, transition from traditional disciplinary research to Earth System Science is a necessity and is currently ongoing. This shift in research tradition demands new facilities for substantial advances in the field. This chapter outlines science cases within these fields that will enable a smoother transition that in turn will result in reliable research results to support society.

Climate, Earth and environmental science encompass a wide range of disciplines from the study of the atmosphere, the oceans and the biosphere to issues related to the solid part of the planet. They are all part of Earth system sciences or geosciences. In this chapter we consider, in addition to climate science, atmospheric science, biodiversity, biogeochemistry, glaciology, geology, geophysics, hydrology, and oceanography. Today, e-infrastructure is an essential component of the research practice in all these disciplines, with varying degrees. They contain fields where e-infrastructure has been an integral part of the research practice for a very long time. This is the case for weather and climate science in which extensive datasets are communicated worldwide and used together with extensive computer-based numerical models. These fields have been and will continue to be at the forefront of use of HPC resources. A number of large institutes and university or network-based centres have already provided substantial e-Infrastructure in their spheres of interest. Common to these types of initiatives is that they are also closely tied to similar initiatives in the European or global arena. As a result, only a part of the current e-Infrastructure usage involves the national infrastructure today provided by SNIC. The science cases presented here do not distinguish between SNIC, agency or university financed infrastructure.

Weather prediction was the first non-military scientific problem addressed in computational science beginning in the 1940´s and the first real-time weather forecast was performed in Sweden in 1953. Skill in numerical weather prediction (NWP) is based on accurate knowledge of the initial state while climate projections address how a change in external forcing changes the characteristics of the climate system, in terms of mean conditions, variability and the likelihood of extreme events, and is more of a boundary value rather than initial value problem. The time scale of climate change is also such that the models need to consider not only the atmosphere, but also the ocean, soil, and such Earth system components as vegetation and carbon cycle.

Today, research and development of numerical weather prediction models is rather limited in Sweden, and is mainly carried out at SMHI with some university collaboration. Two Strategic Research Areas (SRAs) in climate modelling are currently funded, both in collaboration with the Rossby Centre, SMHI: the Bolin Centre for Climate Research (hosted by Stockholm University) and MERGE (hosted by Lund University. Climate research at Stockholm University, SMHI and KTH has, for almost a decade, had access to substantial dedicated HPC resource funded by the KAW foundation. It clearly shows how competitive Sweden can be since it has made it possible for Sweden to participate in the internationally coordinated global climate model simulations that are assessed in the IPCC report published in September 2013 [SDP13]. This experience, together with several recently recruited international senior scientists through SRA funds, has paved the way for substantial scientific accomplishments in Sweden. The success is, however, critically dependent on further investments in e-infrastructure of considerable size.

Another line of research where Swedish scientists are sizeable users of e-Infrastructure is computer-based numerical simulations for process understanding. Besides the scientific understanding, which is highly relevant by itself, they are also used systematically to provide valuable insight into how to describe the effects of small-scale processes(such as clouds, turbulence, aerosol particles, ice-sheets, ocean waves, geophysical, hydrological and biogeochemical processes) in large-scale models. Knowledge about these is imperative for improved NWP, climate models and projections. This area engages a large group of HPC users in Sweden, at the Universities of Lund, Gothenburg, Stockholm and Uppsala, as well as KTH, the Swedish Institute of Space Physics and SMHI. Observational data of the Earth´s properties and changes over time is fundamental for climate, environment and Earth sciences. The volume of these data is growing exponentially due to technical improvements and increased spatial/temporal resolution, and are now of the order of Petabytes. Examples are satellite-based observations, meteorological data, greenhouse gas data, seismic and magnetic data, and DNA data for fauna and flora. Data management and interpretation tools are essential for scientists to access, integrate and synthesize information from these large and often distributed data sets. Substantial breakthroughs are envisioned from research based on combinations of a range of observational data. For this to happen, procedures for open, interoperable and sustainable e-infrastructure are fundamental. Portals for discovery, evaluation and access to distributed data sets, including web services and applications, are envisioned.

Usability of these large data pools will be greatly facilitated by provision of standard analysis tools, potentially multiplying the user communities. Accessibility of such portals and services based on huge amounts of data require a perspective for future development of network speed. Metadata, data specifications and data quality descriptions for the data sets within the e-infrastructure should follow international standards to enhance interoperability with other national and international portals. Already initiated and ongoing national/international efforts are, for example, ESGF, ICOS, ECDS, Geodataportalen, INSPIRE, Centre for Genetic Identification, EPOS, Swedish Life Watch and the Global Biodiversity Information Facility. The building of such databases, and research based on them, are both growing rapidly across all research fields within climate, environment and earth sciences, and the demands on e-infrastructure in terms of storage, application experts and network speeds are steadily increasing.

e-Science challenges

Climate, environment and earth sciences have witnessed an impressive development during the past years. The field continues to evolve and current challenges that rely on e-Infrastructure are:

1. The need to move from current climate models towards Earth System Models.

2. The need for very high-resolution models to understand, quantify and predict extreme events, to better assess the impact of climate change on society and economy and improved weather forecast systems.
3. The need for comprehensive and complex datasets effectively accessible for various science communities for process understanding, model evaluation and advanced high-resolution assimilation of multitude of parameters and model development.
4. The need for improved mapping and understanding of biodiversity and biota changes through whole ecosystem modelling.

The Panel has chosen to exemplify these needs in four science cases structured by e-science practice. Within each of these areas, a few more concrete examples are given. The general conclusions of the panel are summarized in section 5.5.

5.1 Fundamental understanding of weather and the climate system, sensitivity and predictability

Society requires weather forecasts and climate scenarios for decision-making. They rely on computer based numerical prediction models that have shown to be extremely useful, and the skill and usability has continuously improved. Nevertheless, the complexity of the problem and uncertainties in models limits their usability. For example, seasonal forecasts of sea-ice in the Arctic for shipping purposes and reliable information for regions such as Scandinavia on climate change for the next few decades would potentially be very useful. Substantial breakthroughs in quality and usability of weather prediction and climate simulations call for drastically increased model resolution, increased model complexity, and a larger number of ensemble integrations for characterization of robustness and uncertainty. Swedish scientists in the context of the international climate community are presently pursuing development options in all three directions.

5.1.1 Decadal predictability and extreme events

Modelling the climate system is challenging because of the myriad of interacting complex processes which are acting on a variety of spatial and temporal scales. This behaviour comes from the inherently non-linear governing equations describing the physical processes, and the interaction between the components, for example atmosphere and ocean. There is evidence that a number of important climate processes on seasonal and decadal time-scales, such as blockings and tropical storms, are better represented when the resolution is increased. To assess the predictability on seasonal to decadal timescales, and the added value on regional scales, fully coupled climate models need to be applied with sufficiently high resolution for a large number of simulations of the recent past. The intrinsic uncertainty of the initial state requires an ensemble of identical simulations with slightly perturbed initial conditions which results in a large number of coupled climate simulations. High resolution and large ensembles facilitate the sampling of extreme events such as damaging wind storms and extensive heat waves. For the decadal prediction core experiment of CMIP5 (Coupled Model Intercomparison Project, Phase 5. The experimental basis for IPCC AR5), a total of 500 model years with the coupled ECEARTH model [Haz+10] at a resolution of approximately 120 km was completed by Stockholm University and the Rossby Centre at a cost of roughly 250000 core-hours [1] and about 47 Tbyte storage. Similar experiments with moderately increased model resolution and larger ensemble size require a factor of 180 increase in HPC resources. The highest resolution possible today, keeping efficient scaling (T511-ORCA025), requires more than 3000 times the resources spent for the CMIP5 core experiment and about 120 times more storage demands.

5.1.2 Ocean circulation and sea ice

Ocean circulation modelling is computationally intensive and these models currently often run on coarse grids which limits their usability. Central ocean processes that are currently not resolved in most global models are eddies and turbulent processes that contribute fundamentally to the mean circulation. They also transport heat, carbon, and other key properties both vertically and horizontally, important both for the state of the ocean and also the biota and ocean acidification. Major ocean currents are critically governed by small-scale topographic features of the ocean bottom, such as narrow straits and deep sills; these flows also need very high resolution. In current climate models, these features are essentially unresolved and experimental results with higher resolution show strong effects on, for example, the strength of the Gulf Stream and polar ocean circulations. The Arctic and the Baltic Sea have the additional challenge of sea ice. State-of-the-art global ocean models can currently be run down to 1/12 resolution (5-10 km). Such a run for 50 years requires 25 Tflop/s for a period of 2 months. Further increases in resolution are required to resolve other important processes

1 Numbers based on simulations on Ekman.

and, since observations are still scarce even though they have increased immensely during the last decade, many ensembles are needed to understand the sensitivity to forcing and small scale process descriptions.

5.1.3 Earth System Modelling

Understanding of climate evolution on time-scales from several decades to centennial and longer requires that models also must include sophisticated representations of non-physical processes, including those related to the carbon cycle and changes in land use and vegetation. This involves representing biology, biogeochemical and chemical processes in the ocean and atmosphere and on land as well as human land use. Some of these processes are, in a simplified way, already included in some climate simulations. The results indicate strong impacts in some regions and significant feedback mechanisms on the global scale that can further accelerate climate change. One example of a recent issue concerns the tight coupling between the carbon and nitrogen cycles.

The processes are not fully understood but are potentially important, in the fully coupled system, since nitrogen limitations affect the natural carbon sinks and thus have a bearing on the urgency of climate mitigation. Changes in vegetation, due to climate change can, in turn, affect trace gas emissions and consequently the abundance of particles in the atmosphere with implications for clouds and radiation. The effect of the inclusion of these processes on the computational demands are, depending on the complexity, in the range of a factor of 5 to 20 increase.

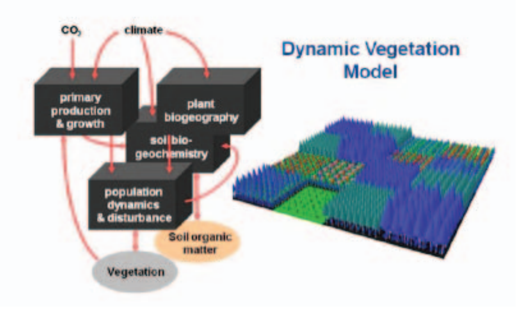


Figure 5.1: Schematic illustrating model components and grid structure in a dynamic vegetation model. Image: Ben Smith.

5.1.4 Potential breakthroughs

- Cloud-resolving climate-change projections, likely dramatically constraining the model climate sensitivity uncertainty.
- Substantial increase in ocean circulation mode resolution to understand the ocean´s role in climate variability and change on both shorter and longer time-scales.
- Earth system models that include land and ocean biogeochemistry, to assess the complex interactions between the physical climate and the living planet and their respective roles in climate change.

5.2 Process-resolving numerical simulations for improved understanding of the Earth system

Many small-scale processes with an influence on the climate are not resolved in global climate models. They have effects on the flow, which determines the climate, so they need to be parameterized. Examples of these are cloud microphysics, convection, turbulence, air-sea exchange, mountain-waves, intertia-gravity wave and tropopause folds. The parameterizations have traditionally been based on theoretical insights and empirical data, however, we are currently experiencing an increase in the usage of computer simulations to further improve the situation. Depending on the scale of the process of interest, techniques such as direct numerical simulations (DNS), large eddy simulations (LES), cloud resolving models (CRM) or high resolution mesoscale NWP models are used. Common to all these techniques is that they are based on the same equation system, the Navier Stokes equations, and solved on a numerical grid defined by the scale of the fluid motions being studied. The great advantage with computer-based simulations, provided sufficient HPC resources, is that they can provide a complete picture in parameter space for a whole range of conditions, which is something extremely difficult and expensive to provide through observations in nature.

5.2.1 Gravity waves and stratosphere-troposphere coupling

Large-scale models are generally based on the hydrostatic assumption that makes the representation of waves and convection particularly inaccurate. Non-hydrostatic mesoscale modelling, in conjunction with observations, is key to study of the influence of such processes, especially in polar regions where direct observations are sparse and hard to obtain. Current modelling of specific observed events, using atmospheric radar measurements in Arctic Sweden and in Antarctica, is carried out using SNIC resources. A single simulation of 8 days for a 2000 x 2000 km^2 domain requires about 40000 core hours [2] and the demand for storage is about 1 Tb. To obtain statistically relevant results, lengthy (~ 2 years) simulations over larger domains are needed, meaning millions of core hours and petabytes of storage. Such a simulation would provide, for example, statistics of trace-gas and moisture transport between the stratosphere and troposphere as well as between the polar regions and lower latitudes leading to improved parameterization so that the effects of frontal systems, turbulence and cloud-cover would be captured in large-scale models.

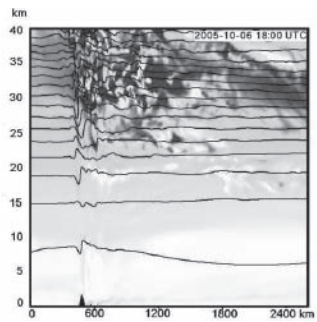


Figure 5.2: Simulated turbulent vortices in the stratosphere which are the result of mountain waves over the Antarctic Peninsula. These contribute to the general circulation of the atmosphere but occur at heights and scale sizes where direct observations are not possible. Image: Sheila Kirkwood.

5.2.2 Simulations of clouds

Detailed studies of clouds are performed using CRM, LES and DNS techniques to study fundamental processes and their interaction. Observation of all aspects of clouds that are important for their formation and radiative properties are difficult and comprehensive datasets are only available for short field experiments. Thus, process modelling, often conducted in internationally coordinated frameworks, is used to complement these observational cases in a most elegant way and has lead to many insights and improvements of climate and NWP models. Currently some outstanding issues regarding low-level mixed-phase clouds, that is clouds with co-exist-

[2] Numbers based on simulations using Weather Research and Forecasting Model (WRF) on Abisko.

ing water droplets and ice crystals in the planetary boundary layer, are being tackled with LES models. These models, for example MIMICA developed at Stockholm University, resolve the majority of the turbulent motions and also provide an advanced description of the cloud microphysics and aerosol-cloud droplet/ice crystal interactions. These clouds are frequently observed and are long-lived in the Arctic but are inherently poorly described in NWP and climate models. LES modelling is expected to substantially improve the situation.

5.2.3 Air-sea interaction with special attention to green-house gas transport between air and water

In order to be able to study and quantify the impact of the green-house gas increase the interaction between the atmosphere and the ocean (and lakes) needs to be understood and modelled more thoroughly than presently. The uncertainties in current knowledge of, and parameterization of this exchange, are due to a complex interaction between fluid motion underneath the water surface and the molecular diffusion in the very thin molecular boundary layer just beneath the water surface. Gas exchange is currently modelled as a function of the wind speed ten meters above the water surface, based primarily on empirical evidence. DNS can be used to study the processes controlling the exchange and compared with empirical and theoretical models proposed so far and how they depend on waves and surfactants. The computational domain is of approximately a few metres horizontally and a few decimetres vertically. Even such a limited computational domain requires tens of millions of computational cells and requires months of computations using up to 500 cores.

5.2.4 Potential breakthroughs

- Substantial improvement in understanding of scale interaction on a range of problems relevant for the Earth system.
- New parameterizations for large-scale models applicable over large parameter space based on a combination of process-resolving numerical simulations and observations.

5.3 Assimilation techniques for weather, climate and greenhouse gas assessment

Data assimilation in NWP aims at determining the initial condition of a system from which its future evolution can then be calculated with a model. In the assimilation process, measurements from a variety of sources are combined with a short model forecast valid at the same time. Two main approaches exist: ensemble Kalman filters and variational techniques. The assimilation process itself also includes modelling and inverse modelling translating between the prediction model and measured quantities. For NWP at scales capturing large weather systems, assimilation techniques have reached a high degree of maturity. Due to the complexities involved, the assimilation process in NWP is computationally more demanding than the subsequent forecast started after the assimilation. For high-resolution NWP, there is a potential accuracy gain from assimilating more satellite-based observations of, for example, cloud properties and winds, but these are still unexplored due to the large requirement on HPC resources.

Keeping track of greenhouse gas (GHG) concentrations and changes is necessary to be able to understand climate change. Quantification and spatial-temporal attribution of greenhouse gas exchanges, particularly over land, represents an important and challenging scientific goal with fundamental policy implications. Biogeochemical exchange between the vegetated Earth surface and oceans and the atmosphere currently significantly dampens the rate of increase of the greenhouse effect and climate change. Current methodologies have large associated uncertainties with respect to the diagnosed magnitude or even direction of GHG sources/sinks as well as their distribution at broad spatial scales. The most promising approach for achieving a future reduction in these uncertainties is through inverse modelling/data assimilation techniques.

5.3.1 Inverse modelling of carbon dioxide

By combining observational data from multiple sources with atmospheric modelling to account for transport, dispersion and chemical transformation of emitted gases, and bottom-up ecosystem modelling, either providing a-priori estimates of local GHG exchanges for use in inversions or being part of the inversion procedure, the model process parameters can be constrained [Sch+07]. Key ingredients in this context are the measurements of GHG concentrations and fluxes as well as ancillary measurements from distributed GHG monitoring networks such as ICOS. The Carbon Portal facility of ICOS, which Sweden will host at Lund University, is responsible for initiating and coordinating joint research community efforts to generate elaborate products, such as maps and time series of GHG variations across Europe and the northern Hemisphere, derived from by ICOS data. This generates an immediate need for supercomputing resources to support inverse modelling. Both MERGE and Bolin Centre scientists are active in international research community efforts for carbon cycle modelling, such as the multi-model ensemble studies contributing to ICOS. The demand for supercomputing resources, especially HPC, for this type of research will build up from close to zero today to a substantial fraction of current national resources over a period of 2-5 years, in parallel with the development of methodology and the accumulation of observational time-series. An important breakthrough will be the establishment of "quasi-operational" inversions for CO2 as well as research mode inversions for other GHGs than CO2, such as CH4 or N2O, which are currently much less established. In addition to HPC, this vision relies on the development of visualization and analysis tools tuned to users´ (both science and societal) needs.

5.3.2 Assimilation in next generation high-resolution regional NWP models

With the advent of more powerful computer technology, NWP models allow prediction of smaller atmospheric features, such as convective clouds, precipitation extremes and hurricane-force winds, which can have high socio-economic impact. Currently, phenomena on the order of few kilometres are becoming resolved by NWP models. Due to the intrinsic non-linearity and short predictability of these features, such as convective clouds, new data assimilation methods need to be developed. New high-resolution observations, such as radar and satellite, provide information on moisture and clouds that possibly could be used in the initialization but the short predictability requires a reliable estimate of the forecast error-of-the-day, for example provided by ensemble forecasts making the whole process very expensive to execute. Numerous data assimilation methods have been developed for the HARMONIE NWP system, such as 4D-Var, hybrid ensemble variational methods, 4D-ensvar and particle filters. Presently available resources allow only pre-studies of these new methods and their accuracy cannot be determined. A typical investigation of one assimilation method required about 750000 core hours [3] for each of a reasonable number of ensemble members, perhaps 20.

[3] Numbers based on experiments on Byvind.

5.3.3 Hydrological modelling for future resilience in society

Hydrological modelling is highly dependent on assimilating facts about the geography, society and measured data. The modelling is dependent on open databases, such as GEOSS, GMES or national portals, efficient processes, and tools for quality assurance of data and results. In particular, there is currently a large requirement for comparative studies within and between research teams around the world to understand processes and their interactions. HYPE, developed at SMHI, is currently used for hydrology forecasts, water-quality analysis and climate-effect studies for Sweden and Europe, and is currently developed for the Arctic region, India and parts of Africa. The predictive ability of the model is highly dependent on different types of data, including climate, soil, land use, and agricultural information. With the increasing availability of observational data, there is a potential to improve the models substantially if the hydrological research community can take advantage of it. In particular, this means additional tools for data management, application experts and computational resources to speed up the process. These resources are crucial for progress in the area and, in particular, for running ensembles of large models over different conditions to estimate uncertainties of the simulations, test hydrological hypotheses over large areas with many measurement stations, and building simulation engines that can give rapid decision support in a crisis situation.

5.3.4 Potential breakthroughs

- A new generation of high-resolution weather forecast models providing statistically significant information on smaller spatial scales.
- Observationally constrained global long-term greenhouse gas concentration and emission databases.
- Hydrology models with substantially enhanced capabilities and expandable to new regions.

5.4 Environmental Genomics, DNA Barcoding and Ecosystem Modelling

Charting the diversity of Earth has been exceedingly slow because it has been so time-consuming and expensive to identify and describe organisms. More than 250 years after the start of taxonomy, only 10-20% of the planet´s macroscopic diversity is known. The study of microbial diversity is largely restricted to organisms that can be grown in the lab. Among multicellular organisms, precious little is known about most decomposers and parasites, including the majority of fungi and many insects. As a result, biologists have so far been limited to the study of isolated bits and pieces in the complex fabric of natural ecosystems.

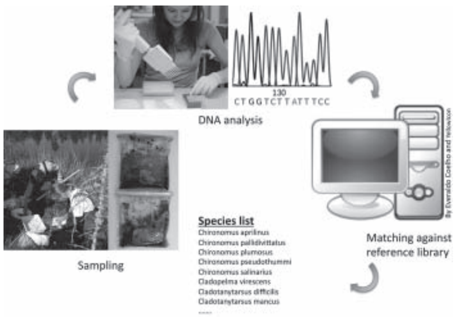


Figure 5.3: Workflow for genetic species identification of biological samples. Image: Johan Bodegård.

New DNA sequencing techniques are now changing all of this. About a decade ago, microbiologists began sequencing random DNA from environmental samples, such as sea water, leading to the development of the field now known as environmental genomics or metagenomics. About the same time, macrobiologists started systematically building reference libraries of selected DNA markers, "DNA barcodes", that could be used for genetic identification of species. A large number of scientists are now contributing to the established global reference library for genetic identification of species, the Barcode of Life Database (BOLD). Web platforms for taxonomic annotation of DNA sequences from environmental samples, such as the UNITE system developed by mycologists, bridge environmental genomics and DNA barcoding.

Environmental genomics and large-scale DNA barcoding are likely to result in several important breakthroughs in coming years, most importantly in biomonitoring and ecosystem research, and Swedish scientists are well poised to play a prominent role in the field. The field relies heavily on e-Infrastructure, including access to medium- to large-scale storage for local databases or mirrors of international databases, and bioinformatics resources (software and hardware) for data processing and analysis. Many future projects will also develop web platforms allowing humans or machines to access, analyse, annotate, and contribute to the available data, and would benefit from an open and flexible environment.

5.4.1 Biomonitoring

Society spends large sums on biomonitoring for purposes such as assessment of environmental change, or detection of invasive species. Genetic identification has been extensively tested as an alternative to the traditional morphology-based approach used today. Despite requiring very little training, genetic identification has been found to be cheaper, faster, more reliable, and vastly more accurate than traditional methods. In addition, it typically allows a much higher proportion of the specimens to be identified to species. For instance, immature life stages typically make up the bulk of benthic biomonitoring samples. While morphology-based identification of immature organisms often stops at the family or genus level, genetic techniques allow species-level identification of the entire material. An important breakthrough will occur when next-generation sequencing of bulk samples, such as soil, water or trap samples can reliably replace traditional identification methods.

5.4.2 Community Ecology and "Biomics"

Genetic identification opens up a whole array of new research possibilities for ecologists. Key interactions among species, which used to take hours of painstaking observation to establish, can now be identified simply by examining a specimen for DNA traces of other organisms. These DNA traces are then matched against a complete DNA reference library for the local flora and fauna to identify the salient interactions. Recent research shows that genetic species-level identification can be applied to gut contents even after the food has been fully degraded and processed, making it possible to identify host species of parasitoids, prey items of carnivores, or host plants of herbivores even when the life history of the species is completely unknown. The entire food web around the spruce budworm, a major forest pest in North America, has been inferred in this way. The availability of material collected by the internationally unique Swedish Taxonomy Initiative (Svenska artprojektet) makes it possible to establish a DNA reference library for the complete Swedish flora and fauna of some 60,000 species relatively easily. Together with our strong ecological research tradition, this could make Swedish researchers world leading in studies of whole ecosystems, or "biomics".

5.4.3 Ecosystem modelling

Ecosystem modelling has traditionally been based on small datasets collected through timeconsuming manual efforts and analysed separately. During the last decade, there have been substantial international efforts to standardize data collection protocols, push the datasets onto the web, and develop and implement the controlled vocabularies, ontologies, and analytical tools needed to allow analyses across datasets. The Global Biodiversity Information Facility (GBIF) represents the largest and oldest of many national, regional or international initiatives pushing for these changes. Several recent trends are now accelerating this development. For instance, biodiversity researchers are becoming increasingly sophisticated in making use of "citizen scientists" in data collection efforts. Biodiversity observations by citizen scientists now account for the majority of the 400 million biodiversity occurrence records provided by GBIF to the research community. New sensor technology and camera systems have facilitated the collection of a large number of biodiversity observations, for example tracking data documenting the movement of individual animals. Swedish scientists are internationally prominent, for example, in the field of animal movement and continental-scale modelling of vegetation, carbon cycle and greenhouse gas exchange [Smi+11]. Significant improvements in the ability to model ecosystems and their resilience, understand past trends and predict future changes are expected as well as novel research by combining biotic and abiotic data. The e-Infrastructure required includes computing resources and storage and web services to assisting Swedish scientists in contributing data to international data portals, and to mine existing data for their own research projects.

5.4.4 Potential breakthroughs

Methods to characterize large ecosystems and improved mapping of biodiversity. Ecosystem modelling of larger and more complex systems.

5.5 e-Infrastructure Requirements

The northerly location of Sweden, with complex and difficult weather patterns, rapid climate change, and particular flora and fauna, presents both unique challenges and opportunities in research within climate, environment and Earth sciences. The scientific challenges and possible breakthroughs are of very high societal importance; the economic benefit of quantifying the certainty and impact of forecasts on all time-scales is enormous. The Swedish contribution to these highly international research fields is well established and is expanding.

The selected science cases demonstrate either already heavy dependence on substantial einfrastructure resources or are foreseen to rapidly develop sizeable needs. It is important, in this context, to take a more holistic view of e-infrastructure. It is not only the access to HPC hardware and mass-storage facilities that is crucial, the human component or "human cloud", in terms of system managers, and scientific computing and application experts, are just as important, especially for databases and more advanced use of visualization techniques.

Projecting into the future, the problems that can be addressed in three of the science cases above (Sections 5.1-5.3) are strongly constrained by computing power and that substantially larger resources would be efficiently used and would lead to much advanced knowledge and breakthroughs. Being at the end of a CMIP cycle, the preparations for the next are already started with the main experimental period between 2015 and 2019, to be in line with the planning of IPCC AR6. For Sweden to be able to participate at the same level as in CMIP5, with the expected resolution increase (T511-ORCA025), it will require more than 3000 times the computational resources expended for the CMIP5 core experiment and about 120 times more storage. The decadal prediction experimental suite performed in Sweden will, alone, then add up to 750 million core hours and 5.6 Pbyte of data. This is without taking into account increasing model complexity, model development experiments, tuning, spin-up and control simulations. The needs for ESM simulations (see section 5.1) are thus substantial and require some additional constraints discussed in detail in a recent European assessment [G+12]. In the 2020 perspective, critical issues concerning the future design of supercomputers will have immense consequences for this field. The simulation tools discussed in the two following cases (sections 5.2 and 5.3) are more diverse, as is the scientific community performing these. To keep up with the international forefront, a doubling of the available resources every two years is a necessity. Considering the research initiated by the SRAs and the general expanding interests of using these tools to advance the field, the five year outlook requires much more than that, at least 10 times the capacity today.

Common to all the research discussed in this chapter is the usage of databases with observational and model data to be used by national and international communities. The dissemination of ESM outputs (Petabyte range) for analysis by a community of scientists over a long period, needs to be resolved. Observational data usually requires less storage but they are less regular and are more challenging to organize in databases. The e-infrastructure needs here are, naturally, adequately sized mass-storage resources but adequate training and proper access to application experts is also extremely important. This is especially true for the biology community (Section 5.4), with less previous experience of e-Infrastructure.

Panel Members

- Professor Gunilla Svensson, (Chair) Department of Meteorology, Stockholm University, Stockholm. Expertise: Numerical modeling of weather and climate from the process level to global climate models.
- Johan Bodegård. Director, Swedish Species Information Center, Swedish Museum of Natural History, Stockholm. Expertise: Biodiversity policy development and implementation as well as natural resource management.
- Professor Sheila Kirkwood, Polar Atmospheric Research, Swedish Institute of Space Physics, Kiruna. Expertise: Studies of chemistry and dynamics of the polar atmosphere, using ground-based radar, satellite remote sensing and modeling.
- Professor Ben Smith, Dept of Physical Geography and Ecosystem Science, Lund University. Expertise: Terrestrial ecosystem biogeochemistry, vegetation dynamics, ecosystem and Earth system modelling.
- Lars Kristian Stölen, Head of department of Geodata, Geological Survey of Sweden. Expertise: Spatial data infrastructures.

References

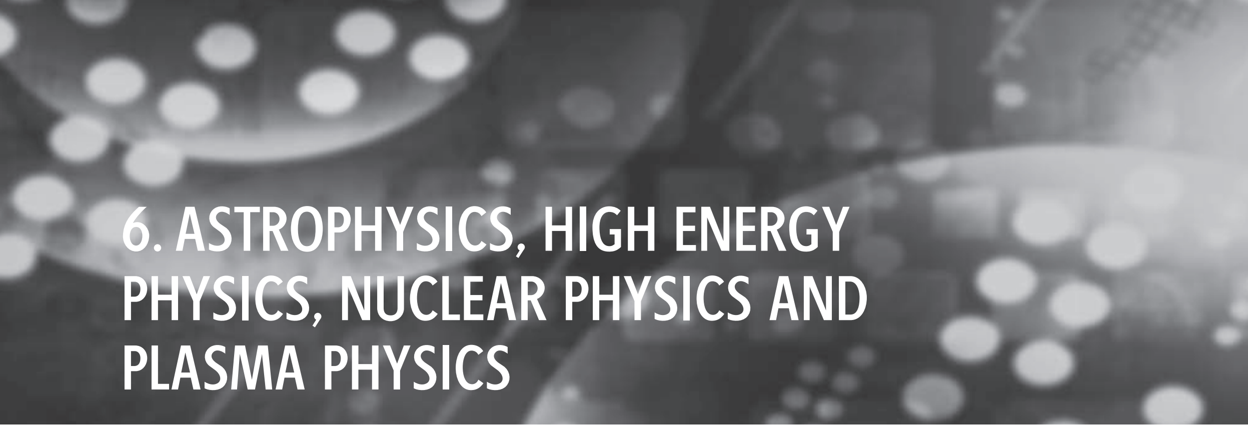
[G+12] M Guest, G Aloisio, R Kenway, et al. *The scientific case for HPC in Europe 2012-2020*. Technical report. PRACE, October 2012. http:// www. prace-ri. eu/PRACEThe-Scientific-Case-for-HPC, 2012 (cited on page 49).

[Haz+10] Wilco Hazeleger et al. "EC-Earth: A Seamless Earth-System Prediction Approach in Action." In: *Bulletin of the American Meteorological Society* 91.10 (2010) (cited on page 42).

[Sch+07] M Scholze et al. "Propagating uncertainty through prognostic carbon cycle data assimilation system simulations". In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 112.D17 (2007) (cited on page 45).

[Smi+11] Benjamin Smith et al. "A model of the coupled dynamics of climate, vegetation and terrestrial ecosystem biogeochemistry for regional applications". In: *Tellus* A 63.1 (2011), pages 87–106 (cited on page 48).

[SDP13] Thomas F Stocker, Q Dahe, and Gian-Kasper Plattner. "Climate Change 2013: The Physical Science Basis". In: *Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Summary for Policymakers (IPCC, 2013)* (2013) (cited on page 40).



6. ASTROPHYSICS, HIGH ENERGY PHYSICS, NUCLEAR PHYSICS AND PLASMA PHYSICS

- *The current HPC in Sweden corresponds to 45 M-CPU-H/y and the demand will, in five years, rise by a factor of five.*
- *Increased user support for optimizing software for multi-core systems is needed.*
- *The network speed needed for observable astronomy will reach 40 Gbit/s in 2020.*
- *Breakthroughs thanks to e-Science could be the identification of the "Higgs" particle, a working fusion reactor, and the identification of the Universe's dark matter*

The topics for this panel range from the smallest constituents of matter (quarks), with a size of 10^{-18} metres, up to the largest structures in the known universe, at 10^{-26} metres. Several fundamental scientific questions are addressed: What are the strange dark matter and the even more strange dark energy in the Universe? What are the sources of the Ultra High Energy cosmic rays? Why did matter and antimatter not completely annihilate during the "Big Bang", allowing us to exist today? How does the strong interaction force confine quarks and gluons into protons and neutrons and generate more than 95% of their mass? What is the role of atomic nuclei in shaping the evolution of the universe? Is the particle recently discovered at CERN the standard model Higgs or is it the first sign of an extension of the standard model? How can we understand colliding extreme objects like neutron stars and black holes? Can we control the same nuclear processes as occur in the sun to create energy in a fusion reactor?

In modern astrophysics very large and complicated instruments, often operated internationally, are used for observations, including radio, optical and neutrino telescopes, detectors like ATLAS and ALICE at the LHC at CERN and satellite instruments like Fermi and PAMELA. The demands for HPC, data storage and high speed data links have, historically, been very high in these fields both for processing and analysing data from the different instruments as well as for detailed computer simulation of processes, for example gas dynamics and plasma behaviour in the future fusion reactor, ITER. The time scales for experiments in these fields are normally very long and the instruments very expensive and complicated. The recent discovery of the Higgs particle at CERN in 2012, for instance, was preceded by more than 40 years of developing, building and running detectors and accelerators. Without an extraordinary development of e-Science in parallel the discovery would not have been possible.

The demand for HPC is, for some areas, in principle unlimited. With a given increase in the calculation capacity the resolution in, for example, a simulation of a process can be increased to consume that increased resource by just changing some parameters. The gain in computer capacity may also open up the possibility to ask new questions not possible to ask today due to limited capacity.

e-Science challenges

During recent years all fields within this panel have shown impressive progress. Still, there are many demanding questions, which need a continuously improved e-Science. A few examples of these challenges are:

1. A description of the nuclear forces from first principles will require significantly increased e-Science resources, in particular computing power.
2. Particle Physics searches for signals beyond the Standard Model in ever-increasing volumes of experimental data, storage and processing of which require complex distributed e-Infrastructures of proportionally increasing capacities.
3. In the field of plasma physics, increased e-Science resources will enable integrated modelling of the full ITER device, which may lead to more efficient use of ITER and more compact and economic fusion reactors in the future.

This panel is focussed on the situation in Astrophysics and Astronomy, Astroparticle Physics, High Energy Physics, Nuclear Physics and Plasma Physics. A summary of the requirements from this field is given in section 6.6.

6.1 Astrophysics and Astronomy

Astrophysics and astronomy address questions about the physics of everything beyond the Earth. In essence it is science driven by ever more powerful observations but for the interpretation of these observations we increasingly rely on advanced computer simulations. Both the observational and the computational work make use of e-Science facilities.

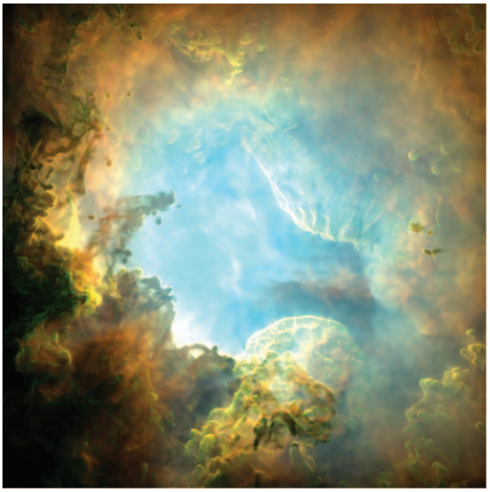


Figure 6.1: Snapshot from a three-dimensional (512^3) simulation of the dynamic growth of an ionized region around a young, high-mass star. Image: Will Henney, Jane Arthur and Garrelt Mellema.

6.1.1 Astrophysical Simulations

Large astrophysical simulations are currently performed by researchers at all the major astronomy departments in Sweden. Such numerical simulations are essential to interpret astronomical observations because the relevant physics cannot be studied experimentally in the laboratory. Simulated astronomical systems range from atmospheres of stars (including the Sun at NORDITA, the largest single user), whole stars, circumstellar environments, groups of stars, the interstellar medium and structure formation in the universe. At present the combined use of national resources (different SNIC facilities) is about 25 million core hours per year. Use of European facilities (PRACE) amounts to an additional 25 million core hours per year. The use of core-hours for astrophysical simulations in Sweden is approximately doubling every two years, with this rate of growth being limited by supply and not demand. Computational astrophysics is an internationally competitive field where the facilities to which researchers have access is vital in determining their success. For Sweden to not fall behind it is vital to keep up with international trends. On the time scale of 5 - 10 years this implies access to systems approaching Exascale performance. Since Sweden by itself will not be able to afford the most powerful systems, continued and expanded access to larger resources in a European context (such as PRACE) is an important goal for astrophysical computing.

To make optimum use of available core-hours, and to give Swedish researchers a competitive advantage, improving the scalability and parallelization of codes will be increasingly important in the future. The establishment of national "competence centres", staffed by computer scientists to work with users to improve codes is therefore an important goal. Developing and providing 3D visualization tools is another important need. Concerning data storage requirements, at present the results from HPC simulations are usually copied to local computers for post-processing, analysis and storage. This model cannot be maintained as the size of simulations increases. It is already common to produce ˜ 1 Petabyte of simulation output data, but this cannot be stored locally. In the future awards of computing time for astrophysical simulations should be made jointly with allocations of medium time-scale (a few months) storage resources and sufficient core-hours for post-processing of these results.

6.1.2 High Bandwidth for Observational Astronomy

The majority of future data reduction needs for observational astronomy are likely to be provided either at university/national facility level or at international level (at data centres for ESO survey telescopes, GAIA, SKA etc.). One specific requirement on national infrastructure is, however, that of high bandwidth through Sweden via SUNET and onward to NORDUNET and to the European academic networks. The radio telescopes at Onsala Space Observatory currently send peak data rates of 10 Gbit/s to collation centres in the Netherlands; this is projected to increase to 40 Gbit/s by 2020. Providing this bandwidth through a mixture of dedicated optical pathways and Bandwidth on Demand is an attractive option to accommodate the different radio telescopes and their needs. A major science case requiring such high bit rates is studies of transient sources in the sky. The time dimension of astronomy observations is, at all wavelengths, expected to be a major growth area of astronomy in the coming decade.

6.1.3 Potential breakthroughs

- The possible detection of natural (or artificial) radio waves from extrasolar planets (from observations).
- Detection of pulsed emission from pulsars in black hole-neutron star binaries providing the most sensitive tests possible of general relativity and competing theories of gravity (relying on a combination of observations and simulations).
- An explanation for the various types of solar activity such as the solar cycle and the potentially dangerous major outbursts known as coronal mass ejections (from simulations).

- Characterization of the earliest phases of galaxy formation in the Universe (relying on a combination of observations and simulations).

6.1.4 Summary of Requirements

The highest priority is that astrophysical numerical simulations have expanded access to large computational resources at national and European level so that Sweden remains competitive. Access to national competence centres to improve code generated by users is another important goal. A common future requirement, both for synthetic data from simulations and final results from telescopes, is the ability to store on few-month long time scales 1-5 Petabytes of data together with modest computational resources to analyse and compare these data products. By the end of the decade large bandwidths (~ 40 Gb/s) over the Swedish and international networks will be needed for both transfer of raw astronomy data from Sweden and downloading large databases. The needs for tiered data reduction capabilities in Sweden for observational astronomy projects depend critically on which projects are supported and their operational models.

6.2 Astroparticle Physics

The field of Astroparticle Physics studies elementary particles, like photons and neutrinos, of cosmic origin to learn about astrophysical processes in space. The field addresses some fundamental questions in science: the unknown sources of cosmic rays (first detected in 1912) and the unknown "Dark Matter" in the Universe (first observed in the nineteen-thirties). The detectors used by Swedish scientists are the Fermi and PAMELA satellites, large Cherenkov detectors situated on the Earth´s surface, like HESS in Namibia, the planned large Cherenkov Array (CTA) (location not yet decided) and the huge IceCube neutrino detector deployed in the ice at the South Pole. IceCube has recently detected neutrino events of energies up to the PeV scale, which might be the key to opening the field of neutrino astronomy after more than 20 years of development.

6.2.1 Potential breakthroughs

- Identification of sources of High Energy cosmic rays
- Indirect observation of Dark Matter using neutrinos and/or photons
- Understanding the reasons for the prevalence of matter over antimatter.

6.2.2 e-Infrastructure Requirements

In the coming two years the total requirement in Sweden for Astroparticle Physics will be 13 Mcoreh/y and in five years is projected to be about 20 Mcoreh/y. The simulation of IceCube events is expected to require much higher quality, demanding more HPC power. The IceCube software is moving towards a more intense use of GPUs. Global fits of dark matter parameters are also expected to start using HPC.

6.3 High Energy Physics (HEP)

HEP researchers in Sweden contribute to and use international accelerator facilities, most notably, the LHC at CERN. The ultimate scientific goal is to describe the origin of the Universe in terms of its fundamental building blocks: elementary particles and forces. To achieve this accelerator experiments collect and study Petabytes of data. In order to preserve and process this data, the Worldwide LHC Computing Grid (WLCG) e-Infrastructure has been built using contributions from all of the countries involved, including Sweden.

6.3.1 LHC experiments: ATLAS and ALICE

ATLAS and ALICE are the two LHC experiments with Swedish participation (KTH, LU, SU, UU). ATLAS is a discovery detector capable of exploring all aspects of particle physics, while ALICE focuses on heavy ion collisions and quark-gluon plasma phenomena. The WLCG storage and computing e-Infrastructure is an integral part of every experiment. It is organized as follows: data from CERN (level Tier 0) are copied to 11 regional centres (Tier 1), and on to national sites (Tier 2) and Universities (Tier 3). The Grid technology makes this structure transparent: scientists in Sweden access data and services located abroad, and vice versa. Sweden is a part of the WLCG, via an MoU signed with CERN, contributing resources to the Nordic Tier1, and providing a Tier2 centre and Tier 3 facilities. Access and utilization of this e-Infrastructure relies on Grid software, adequate local computing and storage facilities, and good network facilities.

6.3.2 Potential breakthroughs

If LHC data is to be analysed to its full potential, it will lead to major breakthroughs in the field, such as:

- Understanding of the primordial state of matter, the origins of mass, reasons for the prevalence of matter over antimatter, the nature of dark matter;

- The possibility to discover physics beyond the Standard Model of particle physics.

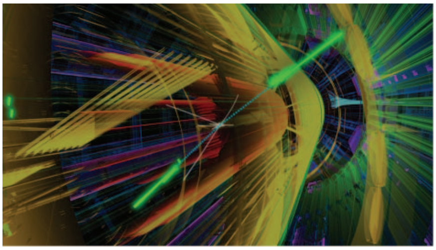


Figure 6.2: Higgs candidate decay into two photons as observed by the ATLAS detector in 2011. Image: ATLAS Experiment © 2014 CERN

6.3.3 e-Infrastructure Requirements

The key HEP e-Infrastructure requirement is continuous provision of and access to the WLCG services, which is necessary for getting unlimited and prompt access to the LHC data. Specifically, it requires: continuous in-kind contribution from Sweden to the WLCG, existence of the necessary Grid tools to access it, and proper operational support. In addition, regional and local facilities for quick turn-around are essential for the Universities to stay competitive. Such facilities must be easily accessible from the WLCG, and yet local users must have the highest priority on them.

Current and projected usage of Tier 1 and Tier 2 resources in Sweden is summarized in the table below. Tier 3 centres at SU, UU and the upcoming one in LU are also a part of the WLCG, though they are used exclusively by Swedish scientists. Unfortunately, the required Tier 3 capacity is not yet available to the LHC physicists. The required network connectivity for all facilities is 10 Gbps. The WLCG e-Infrastructure also requires Grid software (ARC), application software, and human expertise (system and software experts). The LHC will be upgraded to its design parameters by spring 2015, meaning that the WLCG e-Infrastructure capacity in Sweden will have to increase significantly, see the table below. Software will also need to be upgraded. By 2018, when the next upgrade will take place, the needs will increase further. A conservative estimate is to expect linear growth in demand. After the upgrade, the computing, and especially storage needs, will increase by an order of magnitude in comparison with today.

In addition to hardware and software maintenance, user support is necessary in the areas of using and improving Grid tools, introducing new features in response to user requests, and providing efficient software utilities and services for scientists.

Table 6.1: Existing and anticipated future needs for High Energy Physics

Resource	National and Regional				International			
	2013	**2015**	**2018**	**2020**	**2013**	**2015**	**2018**	**2020**
Computing, Mcoreh/y	13	21	34	43	857	1270	1890	2303
Disk, TB	2000	4000	7000	9000	122000	168000	237000	283000
Tape, TB	2000	3000	5000	6000	95000	160000	258000	323000

6.4 Nuclear Physics

The primary goal of research in nuclear physics is to understand the origin, evolution and structure of matter that interacts via the strong force and which constitutes almost 100% of the visible matter in the universe. The main activities of the Swedish nuclear physics groups are hadron physics, experimental nuclear structure physics and nuclear theory. Experiments are performed at international accelerator facilities and will, in the future, be concentrated at the Facility for Antiproton and Ion Research (FAIR) in Germany. FAIR is expected to start operations in 2018 and will host several scientific experiments.

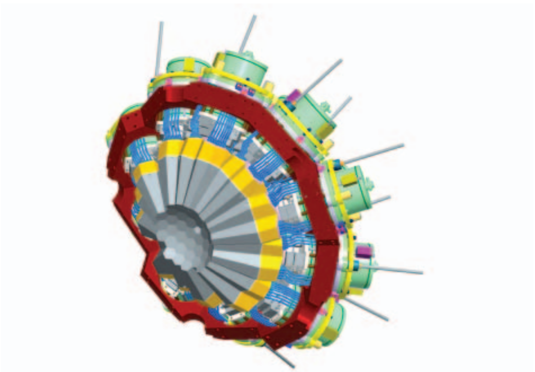


Figure 6.3: The AGATA gamma-ray spectrometer.
Picture courtesy of STFC Daresbury Laboratory.

6.4.1 Hadron Physics

PANDA will be one of the major experiments at FAIR. An understanding of non-perturbative quantum chromodynamics (QCD) is an unsolved problem within the Standard Model. For this, PANDA will study the transition region between perturbative and non-perturbative QCD, for which the ability to handle large data sets is essential. So far, batch farms at GSI and Bochum have been used for the generation of Monte Carlo data and for physics analysis. The feasibility of enabling the PANDA-Grid has been tested. In future, the researchers would like to explore new ways of massive parallel data analysis and large-scale data-mining procedures and techniques.

6.4.2 Experimental Nuclear Structure Physics

This research is focused on investigations of the structure of nuclei far from stability, using high-intensity stable and radioactive ion beams. Topics addressed are, for example, the limits of stability of the elements, shell structure far from stability, neutron skins and neutron matter, neutron-proton interactions, tests of the Standard Model using exotic nuclei, stellar nucleosynthesis and abundances of elements. The NUSTAR (NUclear STructure, Astrophysics and Reactions) science pillar at FAIR has strong involvement from researchers in the Swedish groups which focus on the HISPEC/DESPEC (High-Resolution Spectroscopy/Decay Spectroscopy) and R3B (Reactions with Relativistic Radioactive Beams) experiments. The major instrument of HISPEC is AGATA (Advanced Gamma Tracking Array). The first phase of AGATA is already operational and will also be used at other major accelerator facilities like SPIRAL2 in France. The main use of HPC in this context is for data analysis and simulation.

6.4.3 Nuclear Theory

The focus of nuclear theory is the development of a comprehensive description of all nuclei (stable and unstable) and their reactions, with quantified uncertainties, thus establishing the predictive capabilities of the underlying theory. The development of advanced theory requires large scale computing facilities and close collaboration between nuclear theorists, experimentalists, computer scientists and applied mathematicians. The current use of HPC is 35 Mcoreh/y, most of which is allocated through international resources.

6.4.4 Potential breakthroughs

Nuclear physics is an example of what we would call multi-scale physics. For example, the theory of nuclear forces links quarks and hadrons, and the modelling of neutron stars involves length scales that cover several orders of magnitude. Therefore, modern nuclear physics aims for developments in the connections between the Standard Model (and possible extensions), quantum many-body physics, and the cosmos. Examples of science questions with potential breakthroughs include:

- Can we obtain an improved understanding of non-perturbative QCD?
- Nuclear forces from first principles: is it possible to link lattice QCD with effective field theories in order to better understand the nuclear forces to be used in a many-body theory and the origins of symmetry breaking in nuclei?
- The structure of nuclei at and beyond the driplines: what are the limits of the existence of nuclei?
- Which is the relevant nuclear physics for understanding stellar matter?

6.4.5 e-Infrastucture Requirements

The projected needs of HPC resources in Sweden are 30 Mcoreh/y in two years, 150 Mcoreh/y in five years, 350 Mcoreh/y in ten years. The demands for data storage are 1 Petabyte/y in two years, 5 Petabytes/y in five years, and 15 Petabytes/y in ten years. The required network bandwidths are 5 Gbit/sec in two years, 20 Gbit/sec in five years and 40 Gbit/sec in ten years. The experimental groups will extensively use the GRID infrastructure for data analysis and MC simulations. International access to the HPC resources is essential. In many of the applications, high speed temporary storage and shared memory computers are needed. The availability of support by application experts would be very useful.

6.5 Plasma physics

Swedish research in plasma physics covers a broad range of areas including fusion research, laser-plasma interactions and space physics. In magnetic fusion research, hot plasma is confined by magnetic fields in suitable devices. Plasma physics is playing an important role in determining the performance and hence the size and economy of fusion devices. In laser-plasma research, plasmas are generated via the interaction of a high-intensity laser and a matter target. Such targets can be either solids and gases, and may, depending on the target choice and geometry, be a compact source of very fast particles (for future medical applications) or act as an alternative fusion drive (inertial confinement fusion). In Space Physics research, space plasma processes, such as the interaction between the solar wind and planets, are investigated.

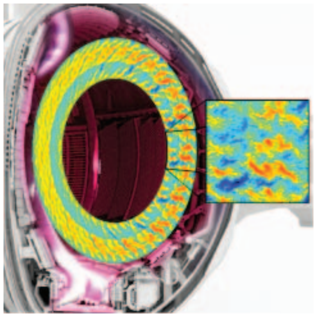


Figure 6.4: Micro-turbulent structures from a gyrokinetic plasma simulation of the JET device, performed at Chalmers University. Image: EFDA and Chalmers University.

6.5.1 Fusion plasma physics

The plasma physics governing the performance of a fusion reactor covers an extreme range of temporal and spatial scales. The Swedish modelling activities are strongly integrated with the large fusion experiments JET (Joint European Torus) and ITER (under construction), a link that is expected to grow stronger over the coming years in preparation for ITER operation. The most computationally demanding activities are currently the study of plasma turbulence. Large fusion plasma physics simulations are performed at Chalmers, KTH and Uppsala University and laser-generated plasma simulations are performed at Chalmers and Umeå University, in connection with experimental activities at Lund University. The main tool used for studying gyrokinetic turbulence is the GENE code which can efficiently use ~ 10^4–10^5 cores on most current devices.

6.5.2 Potential breakthroughs

- Integrated modelling of the full ITER device.

6.5.3 e-Infrastructure Requirements

Currently the largest fusion research user of HPC is the Plasma Physics and Fusion Energy group at Chalmers, which has a large project on gyrokinetic simulations of plasma turbulence using facilities from SNIC/PDC. Typical runs use ~ 6000 cores and 200-300 kCPUhours/run with a yearly allocation of about 10 MCPUh. For the case of laser-plasma interactions, the relatively memory intense computations make systems such as Akka and Abisko ideal, while systems like Lindgren are more difficult to utilize in an effective way. The Chalmers Fusion research group also has access to the international

HPC resources HPC-FF and Helios (www.iferc.org). The aim is to move the turbulence simulations from flux-tube runs to global runs which currently is not possible due to resource limitations. It requires more memory per node than is currently available at Lindgren and would increase the runtime per run to 500 kCPUhours on about 10 kcores. We estimate our needs within the next 2 years to be of the order of 10-20 MCPUh/year. Moving to global simulations increases this estimate to ~ 100 MCPUh/year on the 5 year outlook. This is in order to stay competitive at the research frontier. On a longer timescale, exascale computing is needed. As most of the storage required for fusion research is associated with the experiments we mainly have smaller requirements, on the order of 100 TBytes at most. Storage requirements will likely grow significantly as ITER moves from construction to operation. A Swedish (or Nordic) gateway node would require some dedicated resources to install, maintain and operate. Plasma physics will have an increased need for applications experts in several areas. Application experts will be needed for workflow integration and to improve performance on new systems and on the main resources.

6.6 Summary of e-Infrastructure Requirements

The importance of a strong and developed e-science infrastructure in Sweden is critical for the projects in this panel. Without that Sweden will have limited opportunities to make significant contributions to these fields. The current total use of HPC, for this panel, sums up to roughly 45 Mcoreh/y in Sweden, 8 Mcoreh/y in the Nordic region and 40 Mcoreh/y from outside the Nordic countries. In two years the demand for HPC inside Sweden is expected to be more than 130 Mcoreh/y and in five years more than 350 Mcoreh/y. Looking further into the future, exascale computing is needed. There is an increasing need for access to GPU clusters. Shared memory computers with several 100 TBytes of high-speed temporary storage are also requested for use in, for example, astrophysics and nuclear physics simulations. The demand for data storage varies between users and ranges from a few Terabytes to several Petabytes per year The high speed networks needed today range from 1 Gbit/s to 10 Gbit/s and will increase in the future. The bandwidth transmission needs are expected to increase steadily to 40 Gbit/s by 2020 in the case of Onsala, for example.

Several users have strongly underlined the need for more user support. The scientists are normally not computer specialists and need professional support for, for example, parallelization of the software to use many cores in an efficient way. One possibility would be to establish special "competence centres" to achieve this. Close interaction with HPC centre support staff is needed to port software to new platforms. User support is also needed in the area of using and improving Grid tools, introducing new features in response to user requests, and designing user-friendly interfaces to the e-Infrastructure.

Panel Members

- Professor Per-Olof Hulth (Chair), Department of Physics Stockholm University. Expertise: Neutrino physics.
- Professor John Conway, Chalmers University of Technology, Expertise: Observational radio astronomy.
- Professor Kerstin Jon-And, Department of Physics Stockholm University. Expertise: Experimental particle physicist, working on the ATLAS experiment.
- Professor Garrelt Mellema, Department of Astronomy Stockholm University. Expertise: Astrophysical simulations.
- Professor Hans Nordman, Department of Earth and Space Sciences, Chalmers University of Technology. Expertise: Turbulence and transport in fusion plasmas.
- Professor Johan Nyberg, Department of Physics and Astronomy, Uppsala University. Expertise: Experimental nuclear structure physics.
- Docent Oxana Smirnova, Department of Physics University of Lund. Expertise: Experimental particle physics, e-infrastructure.

7. MATERIALS, CHEMISTRY AND NANO-SCIENCE

- *Materials, Chemistry and Nano-science constitute the largest community of HPC users in Sweden and worldwide.*
- *There is a need for an increase in available computer power of orders of magnitude.*
- *It will become possible to move into a paradigm of predictive simulation of materials and molecular design.*
- *New understanding will emerge from large-scale simulations of materials, nano-systems and chemical phenomena.*
- *Accurate and realistic calculations of complex and dynamic systems will become possible.*
- *There is a need for development of methods, models and software.*
- *There is a need for application experts at different levels.*

Human history is linked to materials used by society, giving rise to eras such as the bronzeand iron-ages. The ability to master and develop materials, molecules, and chemical processes has accelerated over the past 150 years and we are now at a stage where novel materials and molecules are discovered continuously. Advanced materials with tailored properties are crucial to meet many of the challenges that society is facing. Atomic-scale calculations and simulations already play a major role in this development by providing information and understanding that is available only at this scale.

Computer simulations have revolutionized the areas of materials, chemistry and nano-science and have provided, for example, mechanistic understanding of chemical reactions such as ammonia synthesis, explained many of the intriguing hydration properties of water and predicted the ionic conductivity in improved solid-oxide fuel cell materials. For even more complex systems, molecular simulations have been instrumental in revealing how addition of nano-particles to polymers may modify their properties, how ion transport in mixed solid polymer electrolytes can be used to improve battery capacity, and how clays interact with water and can swell or adsorb small molecules. Similarly, modern inkjet printer technology would not have been possible without extensive simulations of polymer emulsions in water.

It is molecular and atomic interactions that determine the functionalities of materials. Thus, it is not surprising that the major share of computational work within materials science, chemistry and nano-science is performed at this level. The algorithms are demanding, and access to high-performance computing (HPC) resources is still a serious bottleneck. Perhaps more so than ever as it has been realized that, with some orders of magnitude more computational power, not only incremental improvements would be possible but, instead, accurate calculations of complex and dynamical systems would become reality. This new computational paradigm will gradually enable simulations to attain experimental accuracy and deliver knowledge quickly and at a comparably low cost. Simulations will have the possibility to predict properties of materials that are difficult to engineer for high-throughput screening.

Another aspect of computational atomic scale modelling is that it has grown into an integrated part of virtually all scientific projects. In fact, it is the possibility to perform theoretical simulations that, in many cases, enables experimental characterization at the atomic level. Several spectroscopic and diffraction techniques critically rely on access to calculations for unambiguous interpretation. Progress in computational materials science is needed for full utilization of the currently developing infrastructures for new types of experimental materials characterization, including those at MAX IV and ESS.

Modelling at the atomic scale is generally performed along two lines, either from first principles by solving quantum mechanical equations, or by use of classical potentials (force-fields). Both approaches are widely used and find applications for different phenomena. First principles calculations for materials and molecules have witnessed a tremendous development during the past 15 years. This progress has occurred thanks to the development of new theoretical methods which have resulted in an improved accuracy that allows for predictive simulations, and thanks to the development of computational software that has implemented these methods. Similarly, classical potentials have proven indispensable in unravelling the structure and dynamics of liquids, large biomolecules and disordered materials. These are systems where the sampling of possible molecular conformations is a primary challenge. For both approaches, an important success factor has been the expansion of computational resources in Sweden that has occurred during the past decade.

The areas of atomic scale calculations have a long tradition of HPC usage in Sweden. The users generally belong to established groups with local support regarding the computational methods, and they use a range of computational resources from local clusters to large international facilities. Research groups within materials, chemistry and nano-science are the largest HPC users in Sweden and worldwide, and many Swedish groups are amongst the international leaders. This puts high demands on Sweden to provide competitive resources. An expansion of the HPC infrastructure resources available to this community in a 5-10 year perspective is crucial. On the international scene, the US Materials Genome Initiative launched in 2012 is an impressive new venture in materials e-Science, with a first year budget of $100 million. In Sweden, the materials, chemistry and nano-science areas constitute important cornerstones in the national e-Science research initiatives, eSSENCE and SeRC.

This panel has focused on methodologies that are based on the atomic scale, although it should be noted that there are important areas that use coarse-grained or continuum descriptions of matter. To receive additional input to this report, the panel distributed a questionnaire to all researchers with large-scale projects at SNIC-funded centres and Swedish researchers with PRACE contracts. The questionnaire asked for scientific grand challenges, how new types of e-infrastructure could promote scientific breakthroughs and what kind of support is needed.

e-Science challenges

The development of atomic scale modelling has been rapid during recent years. However there remain general fundamental challenges, linked to the use of HPC, that need attention. Some of these are:

1. Development of underlying methods and implementations of first principles quantum mechanical techniques for ground and excited state properties. This is needed in order to reach chemical accuracy for all types of materials.
2. Development of underlying methods and implementations of advanced force-field techniques (and statistical-mechanical simulation techniques) to reach high accuracy and robustness in computational predictions.
3. Development of seamless multi-scale methods that bridge atomic scale information with macroscopic quantities.
4. Development of new software and algorithms as well as existing software in order to make use of the capability and capacity offered by new hardware.

The panel has chosen to exemplify how progress in these technological challenges, together with an increase in computational resources, would advance four different areas, namely inorganic materials, nano-structured materials, liquids and molecules, and Soft- and biomaterials. Within each area, a few representative examples will be highlighted. Some general conclusions will be made in section 7.5.

7.1 Inorganic materials

Research in the area of inorganic materials covers a broad spectrum of fundamental and applied problems. Currently atomic scale calculations can – in principle – be carried out for virtually all types of materials including metals, intermetallic alloys, steels, semiconductors, multifunctional ceramics, transition metal oxides, nano-structured composite materials, piezoelectrics and spintronic materials. Likewise, calculations can – in principle – be performed for phenomena related to some of the most central applications of our society, namely ion transport, catalysis, nuclear energetics, coatings and cutting tools. However, the challenge is to make these calculations more accurate both with respect to the system description and the computational method. With increasing quality and reliability, such simulations will be able to provide unprecedented understanding and, equally importantly, successful predictions. The materials field is of high national importance, as more than 70% of Swedish export revenue is based on materials and materials-based products.

Investigations of materials often start with a solution to the first principles quantum mechanical problem. In the framework of the Density Functional Theory (DFT), for example, this can presently be done for systems in the size range of 1-2 nm, or for infinitely large but periodic systems. The methods are under constant development, however, and it can be foreseen that the field will, within only a few years, be dominated by first principles-based molecular dynamics for finite temperature simulations and calculations where interactions between electrons are treated using methods beyond currently used approximations. Also higher-level, quantum- mechanical methods will continue to develop, allowing accurate treatment of ground and excited states for large systems and their interactions with, for example, electromagnetic radiation. This requires an increase in available computer power by several orders of magnitude as the computational requirements of the new techniques scale much faster than the size of the system.

Moreover, first principles calculations are becoming a solid basis for multi-scale modelling by providing parameters that are difficult or impossible to determine experimentally. Such data can be used in statistical mechanics simulations or thermochemical simulations in order to reach extended length and/or time scales. This enables studies of, for example, crack propagations, melting, or solid-solid phase transitions. For simulations of mechanical properties and microstructure (properties on the scale of micrometres), it is common to substitute the atomistic description with a continuum treatment. In the near future development and extensive use of multi-scale schemes can be anticipated, which will bridge first principles quantum mechanical descriptions for atoms and electrons with device engineering and design.

7.1.1 Potential breakthroughs

- With increased computer power new types of materials systems and materials phenomena will be modelled, for example amorphous materials, hard materials and materials with complex magnetic or electronic behaviour. Simulations will be brought close to reality, addressing the most urgent needs of engineers.
- With new developments in theoretical tools and computer algorithms, simulations will be carried out under realistic external conditions, for example at technologically relevant temperatures and pressures, or under stress. This will significantly increase the reliability of the predicted materials parameters.
- With an increased accuracy of computer simulations, it will be possible to build reliable databases of materials parameters with efficient search tools. This will greatly accelerate the design of new materials.
- With seamlessly integrated multi-scale computational tools, new pathways for materials discovery will be possible.

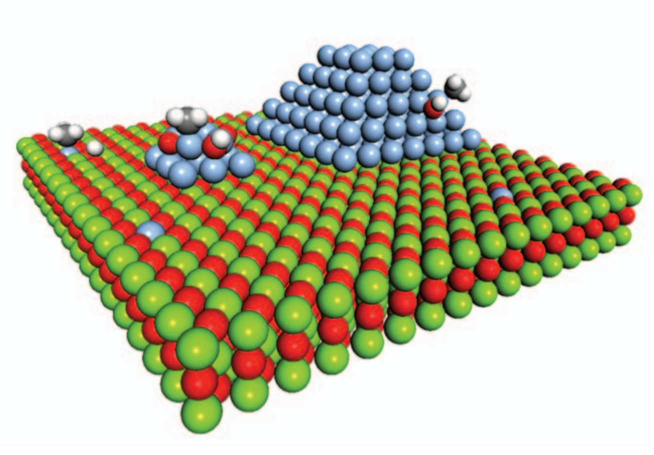


Figure 7.1: Schematic representation of supported catalytic particles. Image: Henrik Grönbeck.

7.2 Nano-Structured Materials

The science and technology of nanostructured materials is, in many respects, an emerging field. Even if nano-scaled materials have been present in the form of, for example, DNA-molecules or catalytic materials, it is advances during recent decades that have enabled design of materials with control at the nanoscale. Such materials are key components in the development of sustainable energy solutions where materials for efficient harvesting of solar and thermoelectric energy are only two examples. Nano-scaled materials are also present within the semiconductor industry where miniaturization is now moving towards atomic or molecular devices.

Catalytic materials are used to produce 90-95% of all chemicals used by society. In many real applications, the sizes of the active catalytic particles are in the order of 1-4 nm. This is a range that poses additional challenges for calculations because of the need to explore a multitude of structural conformations. At the same time, their surface chemistry is often complex. Within ten years this area is expected to enter a new paradigm where first principles quantum mechanical calculations can be performed for experimentally and technologically relevant system sizes. The possibility to access this length scale with accurate calculations is close to a dream scenario and will open up an avenue for faster progress within nano-science where catalysis is only one example.

Interfaces and grain boundaries are critical for many materials properties at the macroscopic scale since they affect mechanical, electrical, optical and thermal properties. It is clear that the importance of interfaces is further magnified for nano-structured materials. One field where atomistic control and first principles calculations of interfaces have become crucial is the semiconductor industry. As devices are miniaturized and ultimately approach the scale of the silicon lattice, the precise atomic configurations become critical. In particular, properties are affected by the exact position of dopants. Atomistic calculations of interfaces and grain boundaries are currently severely limited by the system sizes that can be modelled, with the consequence that further approximations are necessary and the reliability of the calculations is compromised.

The description of the detailed electronic structure and their time-dependence is important both for non-adiabatic chemical reactions at elevated temperatures and, in particular, for modelling of spectroscopic and optical properties. Photo-stimulated processes, such as artificial photosynthesis are believed to be a cornerstone in sustainable energy solutions. Methods for such calculations are, however, computationally very demanding and are, at present, limited to very small systems.

7.2.1 Potential breakthroughs

- When first principles quantum mechanical simulations of system sizes of 2-3 nm can be performed routinely, direct comparisons with experiments will be possible.
- When calculations of free energy barriers over nm-sized particles are possible, accurate design of novel catalysts will be possible.
- When computational resources are available that allow for atomistic simulations beyond current approximations, design of sparse materials will be possible.
- Calculations of atomic structures, electronic states and life-times for nano-structures will enable the design of new functional materials with tailored properties.

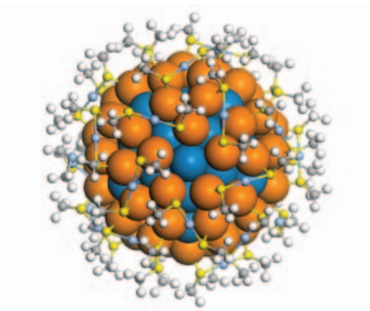


Figure 7.2: Structure of Au144(SCH3)60 proposed from quantum mechanical calculations. The particle which is produced in solution has a core of 114 Au atoms, arranged into three concentric shells. The core is protected by 30 CH3SAu-SCH3 complexes. Image: Henrik Grönbeck.

7.3 Liquids and molecules

Pure water is considered a complex liquid, its building blocks (the molecules) being both polar and polarizable, and in constant motion at different time-scales. Ions in aqueous solution play a central role in geochemistry, electrochemistry, biochemistry, and environmental chemistry and have attracted the interest of the scientific community for a long time. However, the local structure of water and of small ions in solution is still not known. Experiments alone do not provide sufficient detail to unravel the structure of complex liquids. Computer simulations are already now used in many such experiments to aid the interpretation. However, to fully assess the structure, dynamics and energetics of water and simple ionic solutions, long-duration molecular dynamics simulations of several thousand particles would be necessary, and (simultaneously) with an accuracy 5-10 times better than is currently possible with classical force-fields or DFT-based ab initio MD interactions. Ideally, the quantum mechanical effects in the nuclear motion should also be incorporated. Such detail will not be possible within the foreseeable future, but access to a few orders of magnitude more computer power, and accurate many-body classical force-fields, will take science a significant step forward in this field. This will allow the understanding and prediction of transport mechanisms, redox reactions, pH, electrochemistry and a range of industrially important processes will come within reach.

A related, computationally challenging area, where increased computer power will make a significant change, is solid/liquid interfaces. Finding the structures, energy barriers and reaction pathways at such interfaces will yield information that is difficult to obtain with any other methods, and which will have a direct impact on the understanding of, for example, electrochemical reactions at electrode surfaces, corrosion, heterogeneous catalysis, geochemical phenomena and nanoparticles in solution. To unravel the chemistry of such complex multi-component systems is an important target for the future.

A large increase in computer power, together with the use and design of hybrid methods that combine, for example, molecular mechanics, molecular dynamics, DFT and higher-level quantum mechanical methods, will make it possible to simulate complicated molecules and chemical reactions with increased accuracy. Predicting reaction mechanisms and rates, together with their pressure and temperature dependencies, can then become a reality.

Access to a few or many orders of magnitude more computer power, will take science a significant step forward in this field, and allow researchers to:

- Predict and understand structure, transport properties, pH and reaction rates of complex molecular and ionic solutions.
- Predict molecular and electronic spectra of complex systems.
- Predict and understand redox chemistry in solution and near interfaces.
- Perform high-level quantum-chemical simulations based on long molecular dynamics trajectories for thousands of molecules to calculate statistically reliably liquid properties.
- Treat arbitrary dynamical systems with a uniform and integrated, high-level, quantummechanically-based molecular dynamics approach for many atoms and long durations, thereby opening up new challenging systems for simulation studies.

7.4 Soft- and Biomaterials

Soft materials and biological macromolecules are characterized by extreme conformational complexity. While there are experimental methods, in particular crystallography, that can determine the structure of biomolecules, other soft materials only have average order properties rather than a well-defined structure. Even for structured biomolecules such as proteins, functions are determined by room-temperature motions and interactions with other molecules. Computer simulations have been revolutionary in this field. The combination of conformational complexity and motion means that most methods rely on classical force fields together with statistical mechanics to sample different states. Both molecular dynamics simulations and Monte-Carlo sampling techniques can achieve this for tens of thousands to millions of atoms. In the last decade there has also been a strong emerging trend of coarse-grained approximations with particles larger than atoms. In life science, molecular modelling of biological macromolecules has become a cornerstone of structural biology and is also used to understand bioenergetics, photosynthesis, nerve signals, diseases due to mutations that affect the protein structure in our DNA, and as a general "computational microscope" to study atomistic dynamics.

Even when coarse-grained models can be used, or simple atomistic force-fields are sufficient, soft-matter simulations and bio-materials modelling are limited to length scales of ~ 10 nm and timescales of around a microsecond. With a continued expansion of the Swedish computing infrastructure, this scenario is likely to change significantly within the next decade or two. By use of massively parallel supercomputers, researchers are only 2-3 orders of magnitude away from standard simulations on millisecond scales, where simulations will start to overlap with direct microscopy measurements (rather than indirect spectroscopy). In fact, some of these types of simulations are already possible on new special-purpose application-specific integrated circuit hardware.

Also in the soft matter and biomaterials field, the possibility to go to larger systems, longer time scales and more elaborate interaction models will provide more accurate results. One example from the area of biomolecular simulations is the calculation of free energies, in particular for small molecules binding to proteins or DNA. This is the central focus for the pharmaceutical industry, which generally uses simplified docking methods to predict compounds that should be tested experimentally. Today, simulations are neither fast enough nor accurate enough to replace the traditional screening techniques. However, there is extensive research on the develop-

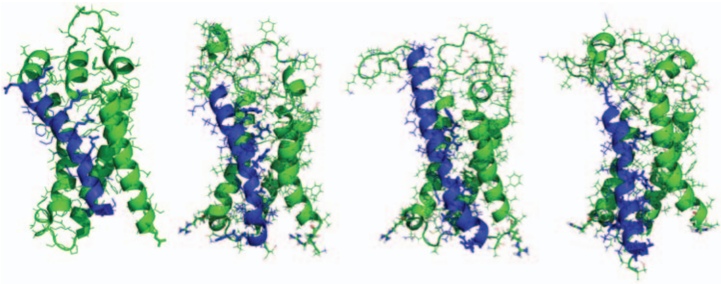


Figure 7.3: The cycle of states a voltage-sensor membrane protein in our cells goes through when a nerve impulse arrives and alters the potential over the cell membrane. Short-lived conformations like these cannot be captured directly in experiments but, by using experimental data in combination with state-of-the-art simulations, it is becoming possible to derive computational models. Image: Erik Lindahl.

ment of hybrid techniques where classical simulations are combined with polarization or first principles quantum mechanical methods for a small part of the system. With computers offering 3-5 orders of magnitude in performance, it is expected that such techniques will be able to predict free energies with similar accuracy to current high-throughput experiments. These calculations can easily be performed in parallel for tens of thousands of molecules and are expected to replace these approximate screening techniques. Combined with bioinformatics and biotechnology, this will enable highly efficient computational design of future drugs, artificial enzymes, and even vaccines.

7.4.1 Potential breakthroughs

- Within five to ten years, researchers expect to treat millisecond timescales, and systems with millions of atoms with force fields, which will overlap directly with experimental techniques. Researchers will increasingly predict results before control experiments are performed.
- Multi-resolution techniques combining first principles and force-field models will be common, and enable researchers to study very complex and large systems, but still achieve first principle accuracy. This will enable soft materials design.
- Next-generation hardware will enable high-throughput studies of binding, and improved accuracy with polarization and multi-scale modelling, which increasingly will make it possible to feed high-throughput sequence data into chemical simulations.

7.5 e-Infrastructure requirements

We are only at the beginning of a paradigm where atomic scale simulations constitute an integrated part of research and development within materials science, chemistry and nanoscience. This is a world-wide development and, in order to maintain and increase the impact of Swedish research within these areas, it is vital that Swedish researchers have easy and abundant access to large-scale computing resources.

The areas covered by this panel have several common scientific challenges which, if solved, will allow for simulation-based investigations of novel materials and phenomena with high strategic potential for future technological applications. It is foreseen that the traditional theoretical approach directed primarily towards an explanation of existing experimental findings will enter a new dimension where predictive computational materials and molecular design will also be a reality. As a consequence, an increasing number of users with new, powerful modeling applications can be anticipated.

The areas covered by the panel have common needs that should be addressed in order to target the described challenges:

1. In a shorter perspective (five-ten years), the single most important issue for the fields of Materials Science, Chemistry and nano-science is a significant increase in available computational resources.
2. There is a need for an application expert infrastructure at different levels, from basic user support to advanced integrated long-term support.
3. A long-term need is assistance with migration to new hardware architectures

Panel Members

- Professor Henrik Grönbeck, (Chair) Department of Applied Physics, Chalmers University of Technology. Expertise: First principles calculations of surfaces and nanoparticles.
- Professor Igor A. Abrikosov, Department of Physics, Chemistry and Biology, Linköping University. Expertise: Condensed matter theory, first principles computer simulations of materials properties and knowledge based materials design.

- Professor Kersti Hermansson, Department of Chemistry - Ångström, Uppsala University. Expertise: Computational chemistry methods for condensed-matter, multi-scale modeling and intermolecular interactions.
- Professor Erik Lindahl, Department of Biochemistry and Biophysics, Stockholm University. Expertise: Classical molecular dynamics simulations, bioinformatics and membrane protein applications.

8. LIFE SCIENCES AND MOLECULAR MEDICINE

- *Due to technology development and general usefulness, the needs in all areas will increase substantially, with more and larger projects and many new users.*
- *The current needs have, across the entire field, reached the ceiling of the current national e-Science infrastructures in Sweden. There is thus a need for more storage, computing and support-capacity.*
- *Improvements in computer technology are needed to cope with the massive increase in data production.*
- *There will be an increasing demand for service functions that enable efficient use of generated big data sets.*
- *A national e-Science infrastructure must be developed which supports secure analysis, storage and access to data with clear rules and routines which guarantee that all activities follow current legislation.*
- *There is a need to develop tools and practices for handling sensitive medical information.*

This chapter deals with the need for e-Science infrastructure in life science research. For convenience the discussion is focused on four sectors where there is an outstanding need for computational resources. These are: genomics, systems biology, structural biology, and imaging. They by no means represent the entirety life science field but should serve as guiding examples.In addition we discuss the growing need for e-infrastructure in clinical/molecular medicine.

With the birth of the revolutionizing tools for DNA sequencing, bioinformatics evolved as a scientific discipline. Sanger´s method for DNA sequencing was published in 1977 and in the mid-eighties the first automated sequencing instruments entered the market, generating immense amounts of data that could no longer be handled manually. At an early stage a focus was placed on collecting sequence information comprising complete genomes and the test cases were the chromosomes of viruses like OX174, SV40, and phage lambda. Thereby the field of genomics was born. A landmark was the start of the human genome project (HGP) in 1990. A first rough draft was announced in 2000, fifteen years ahead of schedule, and a more definite reference sequence of a human genome was published in 2003. This was a historical moment in the sense that humanity, for the first time, got a glimpse of the blueprint of the human body. Although it was neither complete nor 100% accurate has it since been of immense importance for further research in a great variety of fields, even reaching into the humanities. A consequence of several technological improvements is that the increase in production rate of data is monumental and is likely to grow further since even faster and more cost-effective technologies are seen on the horizon. Moreover the emerging possibilities to study genes and gene expression at the single-cell level will generate very large data sets.

Sequence information is descriptive and the biological conclusions that could be drawn directly from the human genome sequence were few. Understanding of gene function requires more sophisticated tools than gathering nucleic acid sequence information. So called reverse genetics has been rewarding but is a slow process. At the end of the day one would like to know, not only how individual genes function, but rather how the genes in an organism function in concert. Having access to vast data sets obtained by different experimental routes is at the heart of an emerging discipline known as systems biology. The objective of systems biology is to obtain a quantitative description of a biological system, and this quantitative description is typically in the form of a mathematical model. It is apparent that this discipline requires extraordinary computing capabilities

Structural Biology is central for understanding of how the components in a living organism are constructed at the atomic level and how they interact. The discipline involves a broad range of experimental and computational techniques. Our overview deals with the computational and data handling needs in protein crystallography, NMR spectroscopy, cryo-EM and computational biology. Structural biologists have, since the early days of the discipline, been heavy users of computer technology as these play an essential role in transforming massive amounts of collected crystallographic data to 3D structures. The need for a good e-infrastructure is very clear and also related to storage and archiving of information. Efforts are ongoing to establish international standards for handling of the data.

Imaging, based on classical tools like light microscopy and medical X-ray examinations has not, in the past, been in need of sophisticated e-infrastructures. However, this changed with the introduction of modalities such as Computer Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) resulting in increased volumes of imaging data. The needs for advanced e-Infrastructure are related to the production of images as well as post-processing, storage and transfer of information between sites. The problems are not only technical but also related to strict requirements for protection of the integrity of the patient who has been subjected to the examinations.

The progress in life science research is changing the face of medicine, which is becoming increasingly more evidence-based. Disease processes are becoming understood at the molecular level and the knowledge is gradually translated into new modalities for diagnosis and treatment. **Molecular medicine** is the designation of this branch of medicine where knowledge from the molecular biosciences is translated into medical practice. The expectations of this branch are overwhelming. Accordingly the progress in molecular diagnosis has been monumental whereas the advances in treatment are less impressive due to the enormous time lag between discovery and the introduction of a new drug into clinical practice. The costs of clinical trials are also prohibitive and a surprisingly small number of new original pharmaceuticals enter the market annually. Much of the medical research is based on analysis of samples from patients and their clinical records. The management of such data is controlled by legislation in Sweden, under the personal data act. Informed consent from the donors, as well as an approval from a regional ethical committee, are strictly required. Samples and data are also regarded as personal and must be handled according to what has been included in the informed consent. How this type of data can be transferred, used and stored is central to the e-Infrastructure of medical research in the future.

e-Science challenges

1. To cope with the rapid accumulation of vast data sets that need to be processed, stored and in some cases archived.
2. To establish the functional consequences of the millions of genomic variants which are present in human genomes.
3. To develop technologies which enable analysis of gene structure and gene expression at the single cell level.
4. To develop computational methods allowing studies of gene-gene and gene-environment interactions.
5. To develop predictive computational tools for drug design.
6. To further develop advanced imaging techniques and post processing algorithms for personalized patient management and treatment.
7. To further develop imaging techniques and post processing algorithms in functional MRI to gain new insights into the wiring of the brain.
8. To provide computational tools for the rapid structure determination of large macromolecular complexes/machineries using X-ray crystallography, electron microscopy and NMR.
9. To design programs which will facilitate interpretation of genomic information in a medical context.

10. To develop practices for handling sensitive medical information and to develop standardized concepts for collecting phenotypic information from patients and controls.
11. To respond to the demand for service functions that enable efficient use of big data sets.

8.1 Genomics

Genome sequencing is today being used in a variety of fields generating extremely large data sets. Sequencing has evolved dramatically since the publication of Sanger´s method in 1977. Thee early sequencing methods which were based on electrophoretic separation of polynucleotides on gels, were laborious and costly and could only be used for reading sequences from small genomes like those of viruses. In 1987 an American company, Applied Biosystems, introduced an automated version of the Sanger-method enabling larger sequencing tasks. In the late 1980s a debate started as to whether time was then ripe to sequence the whole human genome. Although it was anticipated that superior sequencing technologies would be introduced during the course of the HGP, the two drafts of the human genome which were announced in the year 2000 were obtained by the automated Sanger technology. Since the completion of the human genome project, several dramatic improvements in sequencing have been made all of which are non-gelbased. Several competing instruments are currently on the market having different advantages. A common feature is that they are capable of producing gigabase amounts of sequence daily. The improvement in speed has been accompanied by a dramatic reduction in cost. The current goal is to sequence a complete human human genome with high accuracy within a week and at a cost of less than 1000 US dollars. This goal will be reached in the near future. The introduction of vastly more powerful and cost-efficient sequencing technologies makes it possible to address new types of biological questions, such as studies of population structures, identification of mutations in cancer cells, identification of pathogens in crude samples, and comparisons of entire microbiota.

The new molecular technologies have changed the face of genetics and a key challenge is to identify genetic variants in patient genomes which are associated with common diseases like hypertension, heart disease, asthma, and schizophrenia. The information contained in genomes is not restricted to the nucleotides but includes epigenetic modifications of the nucleotides which now can be studied.

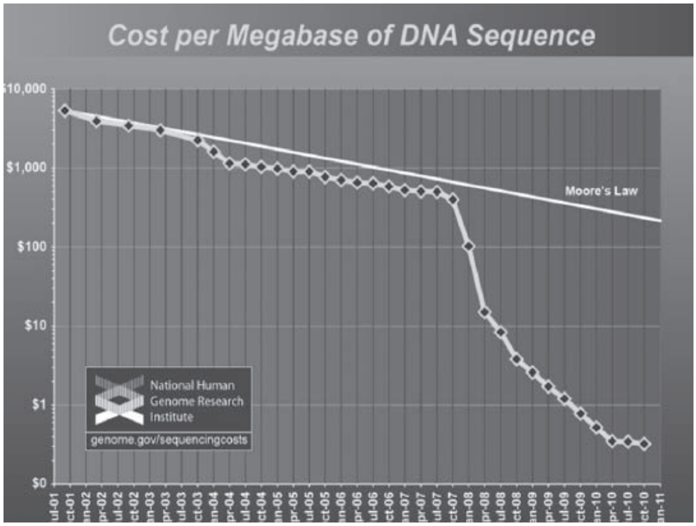


Figure 8.1: Sequencing has progressed much faster than computer performance. Image: National Human Genome Research Institute.

The utilization of sequencing methodology has, in recent years, increased drastically in Sweden. With the development of the genomics platforms at the National Genomics Infrastructure hosted by the Science for Life Laboratory (SciLifeLab), and a few smaller facilities, the data output, and with it the need for both computing and data storage, has increased dramatically. In the last three years, the number of projects run at the Stockholm platform has increased from a few hundred to several thousand and the number of computer core hours has increased accordingly.

While it was previously possible to carry out the computational tasks associated with the bioinformatics analyses in house, this is now impossible. The datasets are so huge that computing clusters and vast amounts of memory are needed. In addition, large amounts of storage space are needed. Internationally, several genome centres have built up their own computing platforms, but in Sweden this has not been possible. The researchers are thus now dependent on the available national infrastructure provided by SNIC (the Swedish National Infrastructure for Computing). The general idea is that SNIC supplies and supports the hardware, while different subject-specific organizations provide software and assistance for users. Of the SNIC platforms, the genomics field currently mostly uses UPPMAX, due to their close association with SciLifeLab and the transfer of specific competence in bioinformatics from ScLlifeLab to UPPMAX. The expansion of the use of genomics to many more fields and users, makes it necessary to expand organizations such as BILS (Bioinformatics Infrastructure for Life Sciences) and WABI (Wallenberg Bioinformatics Infrastructure).

There is a general agreement that the sequence data output will continue to increase exponentially through both technology improvements and through the addition of more users and projects. It is clear that the existing platforms are already beginning to be insufficient to cope with all the projects as waiting times are getting longer.

8.1.1 Potential breakthroughs

- Genomic as well as epigenetic information will be collected from thousands of individuals, from different tissues and in some cases from single cells
- Mechanisms behind so called complex diseases, which depend on more than one gene variant, will be unravelled opening new possibilities for diagnosis and treatment of many disabling diseases in the western world
- The role of microbial populations for human health will be understood
- Genome information will become a routine clinical tool guiding physicians in diagnosis as well as in therapeutic choices.

8.1.2 e-Science Requirements

The infrastructure platforms that are in place at the moment seem to be working well in most cases, even though there are concerns that the necessary increases in capacity can be obtained without completely replacing the platforms. The close interaction between UPPMAX and SciLifeLab has proven successful and it is clear that closer communication between everyone involved, including SNIC facilities, software infrastructure platforms and users, will be needed in the future. Possibly all these facilities should be organized under a common higher structure, in order to ascertain that the facilities are able to meet the needs of the users.

A continuous and rapid increase in the capacity for both computing and storage is needed, both short and long term, based on the projections for increase of data output. At present, the technology for sequence data production is developing at a more rapid pace than computing power. Possible solutions include improved computer technology, providing both superior computational power and larger storage, and there are indications that such developments will come. Another possibility is the use of cloud computing, where the cost is currently prohibitive. This will most likely change and it is a very promising area which could completely change the hardware needs in the future.

8.2 Systems Biology

There are many definitions of systems biology, but most of these contain elements such as mathematical modelling, global analysis (or -omics analysis), mapping of interactions between cellular components, and quantification of dynamic responses in living cells. In most cases the objective of systems biology is to obtain a quantitative description of the biological system under study, and this quantitative description is typically in the form of a mathematical model. In some cases, the model may be the final result of the study, with the model capturing key features of the biological system,allowing the model to be used to predict the behaviour of the system under conditions different from those used to derive the model. In other cases, mathematical modelling rather serves as a tool to extract information from the biological system, enriching the information content in the data. There is generally a distinction between top-down systems biology and bottom-up systems biology. Top-down systems biology is basically a data-driven process, where new biological information is extracted from large, often extremely large, datasets. Bottom-up systems biology is, on the other hand, based on presence of very detailed knowledge, and this knowledge is then translated into a mathematical formulation that is then used to simulate the behaviour of the system and for designing experiments.

The systems biology community in Sweden is growing rapidly with key nodes in Gothenburg, Uppsala and Stockholm. The Swedish Research Council has recently provided a starting grant for establishing Systems biology Infrastructure for Life Sciences (SILS), and building systems biology infrastructures in the form of computational platforms is also taking place within the frame of SciLifeLab. Gothenburg University and Chalmers University of Technology are also partners in ISBE (Infrastructure Systems Biology Europe), which is an EU-sponsored infrastructure initiative.

Both approaches to systems biology involve high-throughput computing, but the demand is different depending on the approach. Thus, bottom-up systems biology is similar in computational requirements to traditional simulations performed in physics, meteorology or engineering, and it can be performed using parallel processing. Top-down systems biology on the other hand typically involves processing of very large datasets, and it is often difficult to distribute the computing into parallel processes. Furthermore, the requirement for analysing the large dataset at the same time imposes requirements for very large memory. The technical advances in microscopy and imaging has resulted in increasing awareness that we need to consider spatial as well as temporal aspects of intracellular processes. For bottom-up approaches this implies that models need to deal with geometries and spatially resolved concentrations, which are orders of magnitude more computationally demanding than conventional models based on ordinary differential equations. For the top-down approaches an emerging need for significant computational and storage resources results from analyzing and handling data from time-lapse microscopy, where the rapid increase in temporal and spatial resolution generates overwhelming amounts of quantitative data on intracellular dynamics. Furthermore, within SciLifeLab large datasets on genome sequencing, RNA sequencing, proteomics and metabolomics, are continuously being generated, and it will be a challenge for the above-mentioned systems biology infrastructures to ensure efficient analysis and visualization of these data. If these infrastructures are to handle the increasing amount of data then they will, to an increasing extent, have to involve web-based analysis, computation and visualization with user-interfaces that are easy to use. It is therefore predicted that there will be a large increase in computational demand for systems biology through SNIC in the coming years. In this context it is, perhaps, important to mention that compared with other big data generating scientific communities, such as nuclear physics, meteorology and astronomy, the life science scientific community is very large and covers a wide range of competences, and there will therefore be an increasing demand for service functions that enable efficient use of generated big data sets.

8.2.1 Potential breakthroughs

Potential breakthroughs in systems biology will be the use of mathematical models for drug discovery, drug evaluation and identification of modes of action of new drug candidates. Furthermore, systems biology is likely to contribute to improved stratification of patients in different diseases, hence enabling development of personalized medicine regimes. Finally systems biology will enable enhancement of cell factory design, which will form the basis for sustainable production of fuels and chemicals.

8.2.2 e-Science Requirements

In order to ensure efficient use of leading competence centres for systems biological studies, there is a move towards systems biology infrastructures that consist of databases of models combined with the ability to perform interactive web-based simulations and large-scale data analysis, often with input of the user´s own experimental data. Combining such simulations with advanced visualization will require significant computational power.

Calculation and analysis of high dimensional data needs to be performed in high-performance computer clusters (HPC). There is therefore a requirement for efficient media systems for communications between users, who can input their data through a web interface and perform calculations on an appropriate HPC system. The media system needs to have the capability to handle security issues of the data. The use of cloud computing is another alternative choice to operate interactive web-based analysis and it is important that SNIC takes a leading role in order to prevent dispersal of activities through the appearing of private cloud systems. Typical large-scale simulations and data analysis need large amounts of memory (30-300 GB of RAM) on computing nodes, which is a major limitation in the currently provided Swedish HPC resources. Moreover storage of data before and after computing is another limiting factor. A typical study involving NGS data, such as might be produced in a metagenome analysis of the gut microbiome,requires a large amount of storage (500 GB - 3 TB) to handle the results from the analysis workflow. As the amount of generated data is currently growing faster than the computing power it is essential to develop workflows that can parallelize the computation of large data sets.

In Sweden the computation demand for systems biology is facilitated through BILS, which is a distributed national research infrastructure supported by the Swedish Research Council. BILS provided bioinformatics support to life science researchers in Sweden and is also the Swedish contact point for the European infrastructure for biological information, ELIXIR. In addition to assisting Swedish scientists on a consultancy basis, BILS organizes training and provides infrastructure, including bioinformatics tools and data, and access to computing (in close collaboration with SNIC). Recently the Swedish Research Council funded the start-up of SILS (Systems biology Infrastructure for Life Sciences), another national infrastructure, and SILS is planning to establish several web-based simulation tools combined with advanced visualization. Through BILS and SILS there will be an increasing demand for SNIC resources for life science application, which will create a need for high speed I/O nodes, but these national infrastructures also plan to expand the use of cloud computing. Such cloud-based SNIC resources will be used for:

- deploying course-specific virtual machines for teaching. These would then be typically used intensively for short times (the duration of the course).
- deploying bioinformatics services, such as databases and web-based tools. The services would run continuously. A cloud based PaaS would allow for efficient usage of hardware resources, with the possibility of scaling up services when needed.

BILS and SILS may interact closely with SNIC for construction and implementation of webbased interfaces and command line software, and thereby ensure development of an efficient computational infrastructure for life science. There is an urgent need for expansion of computational nodes with large memory that can link with web-services with sufficient security.

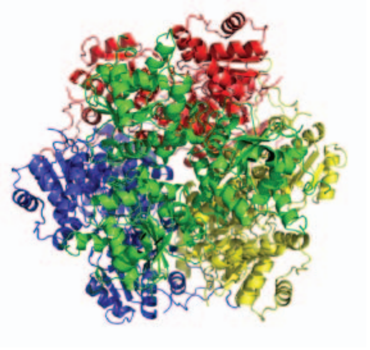


Figure 8.2: Structure of a putative aromatic acid decarboxylase from the ubiquinone biosynthetic pathway of the pathogen Pseudomonas aeruginosa. Image: Gunter Schneider.

8.3 Structural Biology

Swedish Structural Biology has had a very strong international standing in the scientific community for decades, in particular in the area of protein crystallography. In recent years strong research groups in protein NMR spectroscopy, cryo-EM and computational biology have broadened the portfolio of methods and approaches within the Swedish Structural Biology community. There are also several national infrastructures within this research area, the Macromolecular Crystallography beam lines at the MAX II storage ring in Lund, the Swedish NMR Center in Göteborg, the Protein Production Facility (PSF) at KI and SWEDSTRUCT. Research in structural biology requires considerable access to computing, short-term data storage and long-term data archiving. The needs for computing and storage differ significantly between the different research areas.

Computing: Determination of the structures of large macromolecules such as proteins, DNA/RNA and their complexes by X-ray crystallography, NMR or cryo-EM requires considerable computing resources. However the tremendous advances in computer power over the last two decades and the joint efforts by the international scientific community concerning program development has led to the fortunate situation that the necessary computing in most cases can be carried out using resources available to individual research groups. The cryo-EM group, however, uses SNIC facilities for part of the calculations. For the more specialist theoretical computational biology groups, access to national facilities through SNIC is crucial. Computational structural biology groups routinely employ hundreds or thousands of processor cores, for instance on the Cray XE-6 in Stockholm. In addition there are high-throughput applications that require access to high performance computer clusters.

Short-term data storage: Data collection using X-ray crystallography, NMR and cryo-EM generates very large amounts of raw data. For instance, depending on the type of detector, a typical data set in protein crystallography is of the order of 5-10 GB, collected in a few minutes at the more brilliant synchrotron light sources such as the ESRF. Storage and processing of these data sets are carried out in the individual research laboratories, although at several synchrotron facilities automatic processing routines are already in place, which reduce the file size for processed diffraction data to around 10 MB. In cryoEM the raw image data set for a 3D reconstruction is typically of a size of 5-30 GB. For pure computational biology applications, short-term storage at SNIC centres should be proportional to the amount of processing capacity. Currently most of the SNIC centres have too little storage to cater to all users.

Data archiving: A medium sized research group in protein crystallography presently collects about 5-10 TB /year of diffraction images. The long-term storage/archiving of these data are handled differently by the research groups, and are sometimes confined to the processed data, which are of significantly smaller file sizes. It should be noted that data processing programs/routines are well developed in protein crystallography and follow established protocols in the field. Structural biologists have a long tradition of depositing data (diffraction data or NMR data and derived atomic coordinates or cryo-EM maps) in a public data base, the Protein Data Bank (PDB) and the EMDataBank, upon publication of the results. Depositing of structural data in the PDB is now required by all scientific journals and these data are usually publicly available upon publication (some journals still allow a one-year hold). Computational biology groups typically cater for their own long-term storage, although there have been some efforts to generate public repositories as well.

8.3.1 Potential breakthroughs

There will be a shift from structural determination of single molecules to determination of large macromolecular complexes/machineries using X-ray crystallography, electron microscopy and NMR.

8.3.2 e-Science Requirements

Over the next decade the amount of data generated by structural biology groups will grow considerably. A major source of new data will be the Macromolecular Crystallography beam lines at the MAX IV synchrotron source, expected to see first light in 2015 and with the first beam line in operation in 2016. The MAX IV laboratory will be a world leading synchrotron facility and is expected to run at least two MX beam lines by 2020. The size of the computational problem and the necessity of a fast response time to inform users at these facilities about on-going experiments requires access to large scale computing/processing capacity, potentially through centralized computing resources. For pure computational structural biology the demands will also grow with time, and the SNIC facilities need to be updated regularly. Finally, single molecule imaging using X-ray free electron lasers, while in its infancy, will in the future require large amounts of dedicated computer time. A test bed containing GPU resources is being procured at UPPMAX right now.

Data storage/archiving. Estimates of the amount of raw data (diffraction images) generated by these beam lines varies dependent on the type of detector available at the time, but are in the area of 10 TB/day, which, with 200 days of user beam time would add up to 2000 TB/year (2 PB). Increasing data amounts are also expected from the protein NMR field, with the Swedish NMR Center in Göteborg and the NMR groups in Umeå providing national infrastructures in Protein NMR and Metabolomics. The cryoEM facility in Huddinge will produce increasing data volumes, in particular after the introduction of automatic data collection routines and larger detectors. The long-term storage requirements for computational biology may likewise grow beyond the capacity of a single research group. For single molecule imaging demands will most likely become even larger.

These amounts of data would require different approaches to archiving raw data, if such a policy would be adopted, as the costs are far beyond the resources of any individual research group. Discussions are on-going within MAX IV concerning storage of all data collected at this facility, including the fields of physics, material science and chemistry. This also involves questions of data transfer to potential sites for short-term storage and long-term archiving. The SNIC Swedish storage initiative (Swestore) could play a central role here.

Computing and data-storage facilities can be used effectively only if there are accessible tools to use those facilities. Some projects are under way for this; for instance SNIC and BILS contribute to the iRods project in order to facilitate use of Swestore. It is crucial however to treat software – from the systems to the application level - as an integral part of any e-Infrastructure project. A significant amount of software tuning, maintenance and development will be needed.

There are a number of policy and legal questions that need to be resolved before deciding/implementing a policy for data storage/archiving. Most of these issues have been intensively discussed in the CCP4 board (a discussion forum within the international protein crystallography community). These issues can be summarized as follows:

1. Who owns the data? If the data are stored at a national/international infrastructure, who controls access? After all, the data are the result of the intellectual input and experimental efforts of individual research groups. Who is scientifically and legally responsible for the data?
2. Should these data be publicly available, and if yes, when? This is one of the most hotly debated points in the community because it is feared that an open access policy can potentially harm the competitiveness of research projects/groups and in particular harm the career of young scientists. What about data generated by industrial users?
3. Cost/benefit analysis - i.e. is it worth it - Archiving all data is not cheap and will drain funding from other research programs for, perhaps, little benefit. Much of the data collected is not useful and is part of initial sample screening. Finally the point is made that for the very few cases when anybody wants to have a look at raw data it is cheaper to re-collect these.

The International Union of Crystallography has appointed a working group that is addressing the issue of central archiving of raw data and may define recommendations on this issue, provided a common view can be found.

8.4 Imaging

Imaging has become an extremely rapidly advancing field thanks to a number of new revolutionizing technologies like Computer Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), resulting in increased volumes of imaging data. Several national initiatives like the 7T MR scanner (in Lund), the integrated PET/MRI (in Uppsala), the Magnetoencephalography (MEG) (in Stockholm), and, potentially, one more ultra-high field MRI in the future will increase the volume of generated data and necessitate collaboration of national infrastructures. Just as in clinical medical imaging the advances in light microscopy and biological imaging with increased resolution and increased dimensionality of

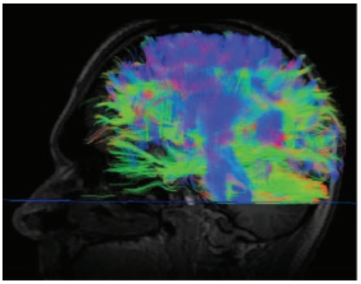


Figure 8.3: Diffusion Tensor Imaging (DTI) of the fibers in the brain. Image: Pia Maly Sundgren.

the images have led to a significant increase in data volumes to be analyzed and stored. The Swedish Bioimaging network (www.bioimaging.se), formed in 2009 and funded by The Swedish Research Council, is a national collaboration of more than 100 scientific members with the present aim to coordinate the use of equipment and training, and with an additional purpose of developing national data servers for storage, archiving, post-processing and analysis of imaging data to serve several national initiatives. The Swedish Biomaging network is collaborating with the European scale Euro-Bioimaging and other national networks.

Currently there are 7 national initiatives that will produce imaging data with an increasing need of national infrastructures for data handling, storage, post-processing and analysis. Large local and national multi-centre studies including both medical imaging data and patient information are conducted; with SCAPIS (www.scapis.se), Big 3 (Region Skåne), and U-CAN (www.u-can.uu.se) among others.

Storage of medical images: standardization of storage in a Picture Archiving and Communication System (PACS) has been successfully achieved, with all vendors of imaging equipment adhering to the strictly-defined DICOM standards.

Storage of non-DICOM data or non-reconstructed (projection) data: these data from CT, MRI, PET and SPECT as well as raw and post-processed data in, for example, MR spectroscopy, functional MRI, cardiac MR, all produce large volumes of data and tend to require large storage space (estimated to be 80 TB for 2014). For these data a special solution will be required.

Post-processing and analysis: Several methods, both provided by the vendors/manufactures or programs provided by third part companies, as well as home-made developments at different institutions, do exist for analysis of advanced clinical and research medical imaging data. Presently no standardization in post-processing of imaging data exists.

Archive: Archiving needs differ for light microscopy and in-vivo imaging fields and these differences need to be addressed properly in the future planning.

Data volumes and data sharing: The data volumes produced by different bioimaging modalities make conventional data transfer inadequate and there is a national need for an infrastructure for data storage, data handling, and sharing of image data.

8.4.1 Potential breakthroughs

Potential breakthroughs in imaging are the use of different advanced imaging techniques for deeper understanding of the pathophysiology underlying different diseases. In addition advanced imaging techniques allow for earlier detection of diseases and are likely to be increasingly used as a basis in treatment prediction models. This will improve stratification of patients and hence enable the development of personalized medicine in which tailored medical algorithms will be used.

8.4.2 e-Science Requirements

With the establishment of nationally-funded imaging equipment like 7T MRI, MEG and integrated PET/MRI, which will be used by researchers nationwide, the need for data handling, storage, post-processing and analysis will significantly increase. Therefore, there will be a demand for solutions for integration and transfer of image data between sites, standard postprocessing methods and data analysis, storage of raw and post-processed data, cost calculations, jurisdiction and rules for storage versus archiving. As the overall need can only be estimated, dynamic secure storage capacity which is easily accessible is needed. From a researcher´s perspective prerequisites for well-functioning national facilities are fast on-time access, qualified post-processing and data analysis at national servers with rapid feed-back functions to the researcher at a local level. Here, one possibility would be to use, the SweStore server hardware of SNIC and dedicated user interfaces (web portals), suitable for data with a wide range of open and proprietary image formats. A few examples of the development are:

- The National 7T MRI in Lund will begin to produce significant volumes of imaging data as of 2015 estimated to rise to 650 TB by 2020.
- Whole-body integrated PET/MRI in Uppsala will produce images that will primarily be stored at the local research PACS server but national data storage space will be needed estimated to reach 250 TB by 2020.

- The National MEG facility in Stockholm will produce 1 TB per research study and therefore, depending on the number of research studies/year, be estimated to require at least 50TB by 2020.
- Super resolution and 4D microscopy center (UMU) will need at least 50 TB/year by 2020 at a national level, in addition to the 16 TB required at the local level.
- U-CAN project, a joint cancer project between Uppsala and Umeå Universities that will need storage space for data in DICOM as well as non -DICOM formats on national services of 10 TB in 2014 with an increase to 25 TB by 2020.
- Large national study (SCAPIS) funded by the Heart-Lung Foundation will need archive and storage space for large volumes of imaging data, in overall estimated to at least 100 TB/year.
- SciLifeLab in Stockholm, with 2 new STED microscopes, is expected to need 5 TB/year so in total 40 TB to 2020.
- CARS and super resolution fluorescence microscopy in Gothenburg is expected to need 6 TB/year, totalling 48 TB to 2020.
- Overall estimations for need of storage space for microscopy data is at least 180 TB by 2020 (starting 2014 with 85 TB and estimated increase with 15 TB/year). The need for storage space for medical imaging (CT, PET, MRI, MEG etc.) will be at least 1000 TB by 2020 (starting with 300 TB in 2014 and estimated increase of 80-100 TB/year).

A limited number of standardized post-processing methods for analyses of medical images need to be provided, enabling the submission of imaging data by individual researchers or a group of researchers. These data, after post-processing and/or analysis, should be sent back to the researcher/group. There is a need for national consensus in the research community for which standardized post-processing methods should be provided. The possibility to "hook-on" individual post-processing software for analysis of data will be required. Solutions for handling patient/subject information need to be provided.

The access to and transport of imaging data needs to be secure, fast, and constantly maintained. There will be a need for service in the form of software developers, developers of medical imaging tools and post-processing programs and for implementation, security management and administration both of the SweStore hardware servers and the different hubs. The possibility to add national or individual hubs and web portals and increase the data storage space has to be a dynamic process. Currently, a group within Lunarc, at Lund University, is doing pilot studies of different possible scenarios for such implementation for the national 7T projects/ LBIC (Lund Bioimaging Centre). Finally, the costs for the individual researchers need to be kept low to encourage high demand and use.

There are currently several limitations that need to be addressed on a national level, legal as well as ethical and issues related to maintenance of data servers. For example (i) the proposed changes in laws regarding the use of social security numbers to identify subjects and regarding data base registers, (ii) the need for national regulations regarding the right to access data through local firewalls, (iii) the ownership of acquired data, and (iv) the increasing demand for open access. Other issues are (v) access time windows for short term (storage) and long term (archiving) data access as well as archiving responsibilities.

In the near future, there will be an increasing need for an infrastructure related to tracers for PET and microscopy. There are also new imaging modalities coming on the market that might, in the reasonably near future, be of national interest and a need for national imaging modalities, such as phase contrast CT, Electron Paramagnetic Resonance Imaging (ERPI), Magnetic Particle Imaging (MPI), micro-MRT might require additional national infrastructure support.

8.5 Molecular Medicine

Molecular medicine is a broad field, where physical, chemical, biological and medical techniques are used to identify fundamental molecular and genetic errors leading to disease, and to develop molecular interventions to correct them. To merge molecular patient data with clinical information is believed to be one of the keys to successful medical research in Sweden. Our advantages include our unique personal identification numbers, a long history of keeping patient records, national quality registers and our biobanks. In Sweden it is possible to perform prospective longitudinal studies which are of great value to research. Internationally and nationally there are efforts to include next generation sequencing and other large scale analysis techniques into healthcare and diagnostics. The results should form the basis for new treatment strategies. A common problem is that the local health care system fails to offer sufficiently large cohorts of patients with unusual diseases to generate statistical relevance in clinical studies. There is then a need for a well-functioning system for collaboration at the national or sometimes at the international level. A good model is haematologic malignancies, which are heterogenic and each disease variant is rare. In total there are about 2,000 new cases every year. Samples from the different sub-types are biobanked nationally and form the basis for reaching clinical relevance. The Swedish government has made a large effort to improve the national quality registers by increasing the financial support step-wise during coming years. There has also been a commitment to build a national biobank infrastructure financed by the government through VR. Also many of Swedish county councils are committed to develop the biobanking infrastructure locally especially at the university hospitals with national collaboration through the national biobank council.

The clinical users are often in need of bioinformatics competence and computing skills. It is therefore important that there exist wellfunctioning service centres that can provide support to the clinical scientists.

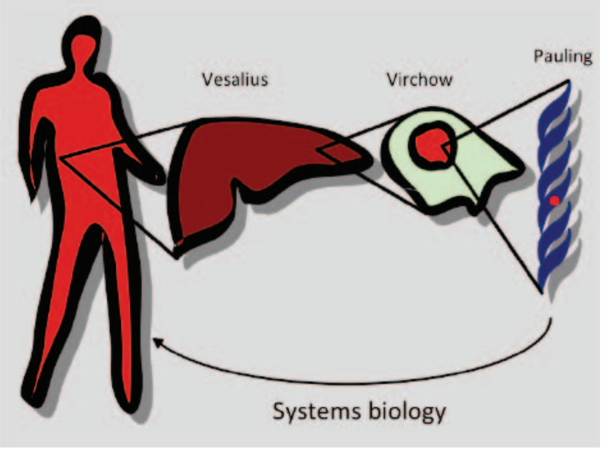


Figure 8.4: Medical sciences have moved from organs (Vesalius) to tissues (Virschow), protein (Pauling) to DNA. Image: Ulf Landegren.

8.5.1 Potential breakthroughs

- The aetiology of common diseases like hypertension, diabetes, and schizophrenia will be understood at the DNA-level allowing new tools for diagnosis and treatment to be developed.
- The era of personalized medicine will be here.

8.5.2 e-Science Requirements

It is difficult to collaborate on a national basis due to a lack of standards within health care, and autonomy for universities and county councils. There is a lack of driving forces to make national agreements at several levels. Medical research personal data is controlled by legislation in Sweden, such as by the personal data act. There are strict demands for informed consent from the donors as well as an approval by a regional ethical committee. How the personal data is treated must be described in the information to the donor. Samples and genetic data are also regarded as personal data and must be handled according to legislation and to what has been included in the informed consent. It must thus be clear for the e-Science infrastructures how this type of data can be transferred, used and stored. In the discussions about future development and increase of capacity the use of “the cloud” is discussed. This poses problems since personal data cannot be sent abroad without a signed agreement between the owner of the data and the recipient. There is a strict requirement to know where the data are stored under all circumstances.

A national e-Science infrastructure must support secure analysis, storage and access to data with clear rules and routines for access making sure that all activities follow current legislation. There seems to be a need for expert centres for external users to access data in a coded form with the ability to go back for follow-up data, but at the same time the databases should be built with a high level of e-security, for example using technology developed by the banking sector. It should be discussed at a higher level how this should be set up. It might be that the e-Infrastructures should be included in expert centres with honest broker functions together with analysis platforms such as SciLifeLab, local biobanks, national quality register holders etc. Service centres should not only include competence in biostatistics, information security and data management but also in ethical and legal matters associated with the use of personal data or collaborate closely with units that hold such competence. The centres need to be located close to the researchers ,in a distributed manner, and with a national coordination. Quality registers fall under separate legislation which must be taken into consideration.

The e-infrastructures need to be able to guide and help researchers to manage the legal aspects around management of personal data and offer secure solutions for encrypted transfers of data when needed. Most data collected within health care remain the health care principal´s responsibility. All county councils and university hospitals have biobank organizations or coordinators today, that manage the legal and ethical administration needed for use of biobank samples. The National biobank council has produced a material transfer agreement for sending samples and data to another principal. They have also produced national templates for applying for access to samples and data for research, that are being used all over Sweden. The local biobanks have routines to manage this and can be an important part for future collaboration between health care and research using national e-Infrastructures.

8.6 Summary

Life science researchers are today heavy users of e-Science tools. The structural biologists were the first in the life science sector to require advanced computer resources. The birth of genomics and systems biology has profoundly increased the requirements for advanced e-Infrastructure platforms and networks in the field. It is clear that the needs will grow at an unprecedented pace due several reasons. Firstly there is a monumental increase in the speed of sequencing at the same time as the cost is decreasing. Secondly there will be a massive increase in data due to the Macromolecular Crystallography beam lines at the MAX IV synchrotron source, expected to be in operation in 2016. Moreover, with the development of new imaging methods and imaging research at the national level, there is a need for advanced e-Science platforms for post-processing and sharing of imaging data between sites. The current needs have, across the entire field, reached the ceiling of the current national e-Science infrastructures in Sweden. The life sciences have special needs in that the e-Science infrastructure must take the integrity issues into special account. Another character of the field is that user support is particularly important as computational skills are less advanced in some of the areas.

Panel members

- Professor Ulf Pettersson, (Chair), Department of Genetics and Pathology, Uppsala University. Expertise: Medical genetics and virology.
- Professor Björn Andersson, Department of Cell- and Molecular Biology, Karolinska Institute. Expertise: Genomics and bioinformatics.
- Dr. Anna Beskow, Head of the Uppsala Biobank, Uppsala Clinical Research Center. Expertise: Molecular epidemiology.

- Professor Jens Nielsen, Systems and Synthetic Biology, Chalmers University of Technology. Expertise: Systems biology of metabolism.
- Professor Gunter Schneider, Department of Medical Biochemistry and Biophysics, Karolinska Institute. Expertise: Structural biology.
- Professor Pia Sundgren, Department of Diagnostic Radiology, Clinical Sciences Lund, Lund University. Expertise: Neuroradiology.

9. ENGINEERING SCIENCE

- *40% of SNAC resources are used for solving problems involving the Navier-Stokes equations*
- *e-Science in engineering is, to a large extent, HPC and code development for HPC*
- *Application experts and research engineers are important for successful application of e-Science in engineering*
- *Knowledge transfer from academia to industry is important in order to achieve real impact*

Engineering science spans the gap between fundamental basic science and industrial application. The area of engineering science includes many subdisciplines and they, in turn, can be divided into more specific subjects. However, all of them aim at covering the area all the way from understanding and explaining to predicting and applying. The practical applications and usefulness in society resulting from progress within engineering science cannot be overstated. Energy production, transportation and communication are just a few examples.

During the last 30 years, the field of engineering science has experienced a radical change due to the introduction of the computer. Digital tools are used not only in academic and corporate research, but also in the complete design process in industry through the use of finite-element calculations in solid mechanics to determine the structural loads and stresses, optimization of production processes, computerized design, etc. In particular in the area of computational fluid dynamics (CFD), numerical simulations of initially laminar, but now also turbulent flows have dramatically increased our knowledge and design abilities. Turbulence models of varying complexity and accuracy have been developed over the years, with profound impact on the daily engineering workflow. The simulation of fluid flows usually requires a large number of degrees of freedom. For example, when considering turbulent flow, the range of excited scales is very wide, ranging from integral scales (of the size of the considered domain, such as an aeroplane wing or even a typical atmospheric eddy) down to small viscous scales on the order of a few micrometres. Thus, CFD applications have always been among the best customers of computer centres. Nowadays, DNS using up to the order of 10 billion grid points are feasible, and a few million grid points are considered to be small simulations. Large-scale simulations were typi-

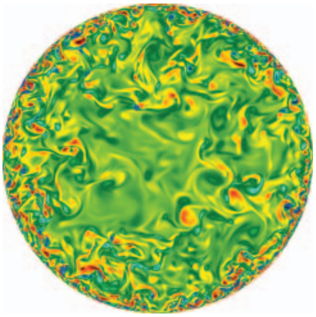


Figure 9.1: TCross-stream cut of turbulent pipe flow at high Reynolds number, showing colours of axial vorticity. The data was obtained from a high-resolution direct numerical simulation (DNS) using about 2.5 billion grid points. The fine scales close to the walls and the sharp gradients in the pipe centre are clearly visible which highlights the importance of adequate numerical methods and efficient codes. Image: Philipp Schlatter and George El Khoury.

cally first performed within an academic framework, but propagate into the normal engineering workflow within a few years.

The area of engineering science has an immediate industrial importance and hence there is need for basic as well as applied engineering science research. Furthermore, there is a strong need for moving the basic engineering science research knowledge into the area of applied engineering science and further into practical application. This is in particular important for e-Science aspects such as codes, data management, best practice as well as postprocessing and visualization.

Engineering science, however, is a quite diverse area that covers many (sub)disciplines such as aeronautical and aerospace engineering, automotive engineering, civil engineering, oil and gas engineering, chemical engineering, nuclear engineering as well as biological and medical engineering. The research area includes turbulence, multiphase flow, combustion, fluid-structure interaction, structural mechanics, electromagnetics, and the general area of multi-physics simulations. The term multi-physics simulations incorporates the combined effort to modelling and subsequent simulation using different physics, which is now becoming strong both in academia and industrial applications, one example being combustion. Note that this should not be confused with multiscale simulations, which is rather routine in any scale-resolving simulation, for example in general turbulence.

In Sweden, e-Science research in the area of engineering science is today clearly dominated by theoretical, computational and applied research on fluid-dynamics related problems, mostly based on solutions to the Navier-Stokes equations. The usage of SNIC machines in Sweden supports this view, with about 40% of CPU time for Navier-Stokes-like simulations (including climate, astrophysics, etc.), and about 25% more direct engineering Navier-Stokes simulations (turbulence, combustion, multiphase flows, etc.). Recent numbers from the INCITE project (USA) show very similar numbers, also with about 40% used for Navier-Stokes-like runs, of which half is used for engineering applications.

Apart from the engineering knowledge, the area is highly dependent on development in computer architectures and computer science, numerical methods as well as other applied areas of mathematics such as statistics and optimization theory. Increasing usage of HPC both in academia and industry constantly expands the possibilities. There are several trends at work simultaneously. In the case of CFD, for example, on the one hand there is a lot of interest in modelling and simulating increasingly more complex flow situations at increasingly higher and higher Reynolds number, whereas on the other hand simpler turbulence models have gained such confidence that there is now a utilization of optimization techniques in many industrial applications.

In the following, a few of the key aspects relevant to e-Science in engineering are discussed:

Software: The development in computer hardware and HPC has been phenomenal over the last decades. New and more powerful computers, however, require accurate and efficient simulation codes. Despite the effort of writing modular parallel codes, scaling to very large numbers of cores (on the order of millions) for certain applications remains a challenge. The need for scaling to very large numbers of cores is only one part of the problem; for many engineering applications there is a need for quite general codes to be developed which allow for a combination of different physical modelling elements (multi-physics) which might happen on different length and timescales. It can be very difficult to develop efficient methods which scale well in such situations. Furthermore, emerging new computer architectures may require the development and adoption of entirely new algorithms for the solution of the governing equations. Such a development effort, which goes beyond just adapting existing codes, turns out to be very expensive. The community in engineering science is less well organized than some other scientific communities. Whereas other areas use a few well-established computational platforms, instead there exist several different initiatives within engineering, ranging from in-house codes, through GPL “community codes”, to commercially available codes. Many institutions and companies make use of various codes with different licenses. There are also many projects that are performed collaboratively between academia and industry.

Post-processing: Data analysis of output presents another great challenge, for example in storage and visualization. The storage needs to increase dramatically as more research effort is put into scale-resolved simulations. This trend is true both for academic and corporate applications. Storage, either short-term or long-term, is a challenge of the same magnitude as the computations themselves, but receives less attention. Connected to that is the ability to post-process and eventually visualize data sets. This is made difficult by the mentioned variety of simulation codes (data formats), but also the wide range of applications and thus the needs for visualization (for example what quantity to display). Visualization of large data sets requires (automated) methods for data reduction.

Multi-physics vs. multi-scale: There is significant interest in multi-physics simulations, where a significant effort is invested to couple various types of continuum-based software such as structural mechanics, acoustics, fluid dynamics, chemical kinetics, and heat transfer in a combined manner. This is an important field with a great deal of interest from industry, where such methods are being routinely developed and utilized. There is still a significant need for research in this area both on the actual modelling of different types of physics and their interaction, as well as very practical aspects of how to couple the numerics. On the other hand, multi-scale models are very complex in their nature, requiring bridging between different length and time scales that may span many orders of magnitude.

Uncertainty quantification: The assessment of the accuracy of simulation results is becoming increasingly more important in the field of computational engineering; this is particularly true in solid mechanics. This new and exciting area will require substantial research in the coming years as uncertainty quantification (UQ) is likely to become an important industrial reality. UQ can easily be standardized, hence requiring best practice type simulations.

Other fields: In addition to the mentioned fields in which e-Science and, in particular, large-scale simulation has a long tradition, there are also quite a number of emerging fields. These disciplines include control theory, multi-body simulations for robotics, very large scale production system and traffic simulations. However, merely involving more and more parts in a simulation, and so making it large and complex, does not necessarily make it e-Science. Hence, in the following description we restrict the appended examples to areas that are truly e-Science, with connection to HPC.

The five main areas of e-Science within the field of engineering science in Sweden today, turn out to be four areas that involve fluid dynamics (turbulence, combustion, aero-acoustics as well as complex/biological flows) and one area involving mechanical and electrical engineering in general. As mentioned above, engineering and, in particular, computational fluid dynamics, is traditionally a large user of computer time, consuming about 40% of all available CPU time. Therefore, research groups in this area are large and experienced HPC users, and a number of groups in Sweden are among the best and most active groups within the area worldwide. A number of local and national centres have been formed within the area, with researchers as an integral part of activities, including the Linné FLOW Centre (at KTH Stockholm), Vindforsk, CeCOST (national Centre for Combustion Science and Technologies), CCGEX (combustion research, at KTH), and a number of EU projects (such as RECEPT). Engineering sciences also constitute an important application area within the Swedish e-Science Research Centre (SeRC) and e-SSENCE. Furthermore, active groups and individual researchers are well represented in the various HPC activities within Europe, such as DECI and PRACE.

e-Science challenges

The area of engineering science has seen a tremendous development over recent years, both within academic and corporate research. A number of main challenges, listed below, can be identified and these should be addressed in Sweden in the coming years.

1. Turbulence: Better understanding of turbulence and associated modelling of turbulence for industrial use, which will enable optimized design procedures, for example in aeroplane wing or wind-turbine design. Inclusion of more complex aspects such as aero-acoustics (noise reduction) or flow control (drag reduction). E-Science aspects are the development of simulation methods, computer implementation and advancement of physical and model understanding.
2. Development of multi-physics and multi-scale methods within engineering fields of importance. This effectively means coupling of different codes in an efficient manner, together with improved physical and mathematical modelling.
3. Reduction of cost of experimental set-ups by performing certain experiments in virtual facilities, such as virtual wind tunnels or virtual crash tests (areas such as automotive and aeronautical research) as well as virtual material modelling.
4. Algorithms and codes that scale to tens of thousands and hundreds of thousands of cores. At the same time, exploiting new computer architectures (for example accelerators) in an efficient fashion.

5. Reduction of turnaround time in simulations (approximately 12 h) for industry usage. This is a key requirement for practical usage of advanced computing in industry.

The panel has chosen to exemplify the area of engineering science using five areas: turbulence, combustion, aero-acoustics, complex flows and mechanical engineering. All but the last area have a long-standing tradition of HPC utilization whereas the fifth area is considered emerging and to have significant potential. The general conclusions of the panel are summarized in section 9.6.

9.1 Turbulence

Even centuries after the formulation of the governing equations for fluid flows, the Navier-Stokes equations, one of the most intriguing consequences of their inherent non-linearity, the appearance of turbulence, remains an open issue. Turbulence is thus named as one of the most important unsolved problems in classical mechanics. Its relevance comes from the fact that most flows appearing in nature, biology and industry are, in fact, turbulent. Atmospheric flows such as hurricanes, wave dynamics on oceans, mixing in combustion engines and coffee cups, pressure drop in pipelines all exhibit turbulence effects, as do blood flows in arteries or air flows in human airways. In the latter systems, turbulence might be the cause of medical conditions and important as such, whereas in technical applications both its favourable properties (increase of mixing) as well as its negative consequences (increase of drag) need to be considered and understood.

As a special, but very important subclass of turbulence, we will focus here on so-called wall turbulence. This turbulence emerges, for example, along the surface of moving ships and aeroplanes or in pipelines, and is the main cause of the skin friction and drag exerted on those bodies. A rough estimate says that about 10% of the total energy in the world is used to overcome turbulent friction in one way or another. Understanding (and potentially affecting) turbulence has thus a direct impact on societal issues such as reduction of CO_2 through lower fuel consumption, and optimization of wind turbines. Even a drag reduction on the order 10% on the surface of ship´s hulls would bring an overall saving of about 5 billion USD per year.

Traditionally, research on turbulence has been mainly performed using experimental methods in specifically constructed wind-tunnel facilities. Direct numerical simulation (DNS) of turbulence (the term "direct", here, refers to "directly from Navier-Stokes" as suggested by Stephen Orszag in 1970) has become an important tool during recent decades due to the availability of supercomputing. Turbulence features a large number of degrees of freedom, which increases rapidly as the so-called Reynolds number becomes large. This non-dimensional number essentially characterizes the speed of the flow, and is thus large for problems of direct technical and engineering relevance. This means that real industrial flow still needs to rely on quite advanced turbulence models, and will do so for some time. Fundamental understanding at reasonably high Reynolds numbers, however, needs to rely on DNS, which is computationally very costly. DNS is, in principle, a numerical experiment, with increased control of boundary conditions (such as inflow and outflow conditions) at the expense of reduced control of aspects such as grid spacing and accuracy of the numerical method employed. This calls for a more serious scrutiny of computer-provided data; even though measurement noise, as inherent in experiments, is not a problem, simulated data is not always correct either, and might be seriously affected by errors made in the problem set-up, or the solution method (numerics). Therefore, uncertainty quantification (UQ) has recently become a very active area of research for turbulence problems. A few months ago, Prof. Moin (Stanford University) presented his first DNS on 1 million cores, and concluded: "We have asked experimentalists for a long time to give us uncertainty bars on their measurements. But I think the time is right to ask computationalists to do the same." One reason computational uncertainty quantification is a relatively new science is that, until recently, the necessary computer resources simply didn't exist?

Present, high-quality direct numerical simulation in turbulence routinely uses $10^{1}0$ grid points, consuming millions of core hours on thousands of parallel processors. These simulations are primarily cases of fundamental interest, whereas particular DNS simulations of complex applications are coming on strongly in both academia and industry. Even though the largest simulations are still performed in the US and Japan, Europe and in particular Sweden has also been an active player in pushing the limits. Whereas flow cases were simple geometries just a few years ago (channel flow for instance), the geometrical and physical complexity of problems undertaken has been increased to more complicated cases such as spatially developing flows (boundary layers, diffuser flows etc.) and this trend is likely to continue and we expect even more realistic geometries to be considered, such as aeroplane wings, surface roughness, combustion chambers etc.

As an example of a future DNS to be performed in Sweden within a 5 year horizon, it is interesting to estimate the necessary computational cost for a resolved simulation around an aeroplane wing section. If we assume a typical wind-tunnel experiment, the Reynolds number based on the cord length reaches up to about one million for transition experiments, and to about 10 million if turbulence is studied. Within an order of magnitude, we can estimate that about 100 billion grid points will be necessary. Given a modern high-order numerical method, about 10 billion core hours would be needed to simulate such a flow. With supercomputers of petascale performance expected to be available within the next 5 years, such simulations are feasible and will be performed.

It is however necessary to point out that for turbulence simulations such Petascale capability is not the end, but rather an intermediate step. Reynolds numbers in nature and industry are still much larger, clearly indicating that for such simulations larger, and equally relevant, problems can be formulated. DNS will always challenge (and always exceed) the available computer resources and, as such, is a prime example of a discipline truly requiring supercomputing.

9.1.1 Potential breakthroughs

- Large-scale simulations of the fully resolved flow around a real aircraft wing section at realistic Reynolds number. Such a simulation will guide the way towards establishing a virtual wind tunnel, cutting down costs of real experiments at greatly increased accuracy.
- Development of accurate and efficient methods for the simulation of turbulence in complex geometries that scale up to millions of processors (petascale simulations).
- Improvement of turbulence modelling based on large-scale simulations of turbulence in canonical geometries.

9.2 Combustion

Combustion has played an important role in the history of human evolution and today it is still very important for our society, as it is the major approach of converting chemical energy to heat and power, making up of more than 80% of the total energy utilization worldwide. In Sweden, about 60% of the total energy supply in 2010 was from combustion of various fuels, including traditional fossil fuels typically used in the transport sector, and biofuels, peat and waste typically used for heat and electricity production. Combustion of hydrocarbon fuels contributes to air pollution and greenhouse emission which affects the earth climate and our daily life. Fire accident and explosion are other undesirable combustion processes. The advancement of combustion science is essential for the control of combustion accident and for the design of efficient combustion devices that have low pollutant emissions.

Numerical simulation of combustion processes is, today, one of the three major approaches, together with theory and experimentation, used by scientists and engineers to develop quantitative description of the various details involved in combustion, such as flame structures and propagation, intrinsic flame instability, ignition and quenching, pollutant emissions, and turbulence/chemical reaction interaction. In Sweden there is a strong industry in combustion engines through companies such as Volvo and Scania, who are world leaders in combustion engine development and manufacturing. High fidelity numerical simulations are used by engineers in these companies for the design and optimization of their combustion devices. It is expected that high performance numerical simulations will play a more important role in the future R&D work in the Swedish industry involved in energy and power systems since it will lead to reduced cost by avoiding extensive and expansive experimental campaigns. This will improve the competitiveness of their products on the worldwide market.

Combustion is a multiple disciplinary science involving physics (fluid flow and heat transfer), chemistry (chemical kinetics) and mathematics (partial differential equations). The flow velocity involved in combustion of engineering interest is typically high enough that the flow is turbulent. Turbulence eddies with a wide spectrum of scales interact with chemical reactions in a highly nonlinear way, leading to various combustion phenomena such as flame stabilization and flashback, thermo-acoustics induced oscillations in gas turbines, and lift-off/self-ignition, knocking, and cyclic variations in piston engines. For gas turbines and piston engines, liquid fuel injection, spray evolution, vaporization, and mixing of vapour fuel with air need to be considered. Combustion of biofuel that makes up of one third of the Swedish energy supply today involves pyrolysis and inter-particle transport processes. Numerical simulation of these combustion processes is a challenging research area of crucial importance.

Currently there are three different approaches used in combustion simulations. The first is the model-free direct numerical simulation (DNS), which provides detailed spatial and temporal distribution of all important flow and thermo-chemical variables. It is typically used to gain insightful information into a specific combustion process such as the onset of ignition in homogeneous charge compression ignition conditions in a $5^3 mm^3$

domain. DNS of this type for simple fuels, such as methane, requires several millions of CPU core-hours. It will therefore remain not feasible for combustion simulations in practical combustion devices in the near future. The most widely used simulation approach in industry R&D groups is based on the Reynolds averaged Navier-Stokes (RANS) equations and transport equations for species and energy. The smallscales of turbulence and reaction layers are not resolved in this approach, which significantly reduces the resolution of grid and the number of time steps required to achieve a statistically averaged field. The third approach is the large eddy simulation (LES) approach, which resolves the energy containing eddies of turbulence, and hence it is more suitable to simulate processes involving large-scale coherent structure such as vortex-shedding in jet burners and bluff-body stabilized flames, precessing vortex core (PVC) in swirl-stabilized flames in modern gas turbines and cyclic variations in piston engines. LES requires a much finer grid and hence it is more computationally expensive than RANS; however, due to the advantage in capturing unsteady phenomena, LES is expected to be more frequently used in future industrial R&D work.

As an example we consider here a combustion LES project carried out at Lund University. The problems examines the use of LES to study liquid n-heptane combustion in a diesel engine with two different injector geometries. LES was shown to have a superior capability in capturing the process of spray evaporation, mixing of the vapour fuel with the ambient hot air and ignition of the fuel/air mixture, as well as the stabilization of the diesel flames. The influence of flow swirl is clearly identified in the LES, with the upwind side of the flame being stabilized further downstream of the jet than the downwind side of the flame. The inter-jet angle is shown to affect the flame structures significantly. For this simulation 3.5 million mesh cells were used; with 512 cores the run took 3 weeks for simulation of the fuel injection and combustion process.

Since both LES and RANS do not resolve the reaction zone structures, a model is needed to take into account the interaction between turbulence eddies and the chemical reactions. For simple problems, such as low Karlovitz number flamelet combustion, there are different models developed and tested for engineering combustion simulations. For more complex problems, there is a need to do further model development and validation. This is currently being done in university research groups where DNS is used to assist the development of combustion models. This trend is expected to continue in the future.

A typical three-dimensional DNS combustion simulation of academic interest will require at least several millions of CPU core hours and a few billions of grid cells. This requires computers with 10^4 cores in order to finish within a few weeks. For RANS and LES, however, the requirement on computers are quite different. Instead of large number of cores needed for one DNS job run, there will be more jobs run for RANS and LES in order to study a wide range of combustor geometry and/or operating conditions for the design and optimization of the combustors. Each job employs far fewer grid cells, typically 10^5 for RANS to 10^6 for LES. The number of cores required for each job in RANS and LES then ranges from $10–10^2$. Currently the SNIC Tier-1 computers are suitable only for RANS and LES. To perform combustion DNS the SNIC computers should be upgraded to Tier-0 machine level.

9.2.1 Potential breakthroughs

- Direct numerical simulation of full combustion processes, both in free space and close to solid walls, for realistic Reynolds numbers and temperatures.
- With the help of DNS data and experimental data, development of high fidelity models for the interaction of turbulence and chemical reactions, and application of LES and RANS simulations to study large-scale engineering combustion processes.
- Improvement of engineering design processes in the area of automotive engines, gas turbine engines, and industrial furnaces.

9.3 Aeroacoustics, flow control, and complex geometries

In addition to the research detailed in the above section relating to canonical turbulence simulations (such as boundary layers developing on aeroplane wings), additional physical effects and inputs might be added and considered. We will briefly discuss a few of these aspects in an effort to highlight the physical (and computational) complexity that can be added on top of an already complex turbulent flow. Note however that this list is certainly not complete.

During the development of new or improved technical devices, such as aircraft turbines, combustion engines, wind turbines, cars and trains, the prediction of the noise fields due to the flow fields bears large societal importance. For instance, aircraft noise in the areas surrounding airports is crucial, leading to stricter regulations. Any better understanding of the interaction of noise and turbulence, and thus the potential of further noise reduction, will lead to direct positive feedback. The methods currently used in industry, however, are mainly based on experience and existing data bases, and cannot be very accurate. Therefore coupled simulations of both flow and acoustic field will be necessary. This is complicated by the need for different adapted meshes for each of the two parts, the potential of including not only turbulence but also chemical reaction (combustion), and the very complex geometries which are usually involved. The first, very promising, steps have already been undertaken and the potential of aeroacoustic simulations can be considered very large, and definitely require large computational resources.

Similarly, one of the goals of simulating and understanding turbulence is the possibility to also affect turbulence in a favourable way, for instance with the aim of reducing drag, reducing noise, or increasing mixing. This discipline of flow control sits at the intersection of control theory and fluid dynamics. The development, optimization and validation of appropriate control methodologies, either based on active or passive control, is again crucially dependent on accurate flow simulation which captures the relevant details of the physics. In addition aspects such as the accurate modelling of actuators (for example of blowing/suction through small holes in the surface, or plasma actuators), will add to the complexity of the simulation set-up. Sweden has very active groups dealing with adapting control methods to real situations, and will thus require large computational resources.

Another emerging area, mostly driven by the increased capability of modern computers and the societal need, is the accurate simulation of technical devices such as wind turbines, trains and cars. As discussed above, the Reynolds numbers involved are too large for fully resolved DNS, but high-quality large-eddy simulations are possible. In particular, the simulation of wind-turbine wakes, leading to a better understanding of the mutual interaction of turbines in so-called wind farms, is an active area in Sweden with great potential. The improved accuracy of such simulations, based for instance on the actuator-line method in which each rotating blade is included in the simulation independently, plus inflow conditions based on realistic atmospheric turbulence, will allow for optimization of such wind farms, and thus improved energy efficiency. Whereas the state of the art is simulations with up to 50 million computational cells, within a 5-year horizon at least a hundred-fold increase can be anticipated, which is necessary for predictive capability.

9.3.1 Potential breakthroughs

- Investigate methods to predict sound that is generated and scattered by fluid-structure interaction via advanced computer simulations.
- Numerically simulate the tip-vortex breakdown generated by wind-turbine rotor blades and the mutual interaction of wind turbines. This will lead to a better design of wind farms with increased energy output for a given area.
- Feedback control of boundary layers leading to practical implementation of controller strategies. Such devices, consisting of actuators and controller unit, could be implemented in aeroplane wings, leading to transition delay.

9.4 Complex and Biological Flows

Complex fluids abound in nature, with examples such as polymeric solutions, milk, foams and emulsions (mixtures of two immiscible liquid substances). Simple Newtonian liquids with immersed elastic or solid particles form suspensions that can display very complex and non-Newtonian dynamics. In general, flows involving such complex fluids exhibit intriguing two-way couplings: the micro-scale dynamics of the particles affects the macro-dynamics of the flow and vice versa. One interesting and particularly challenging class of complex fluids are fibre suspensions, suspensions of elongated particles in a liquid. Examples of applications where such liquids appear are pharmaceuticals, food, and pulp and paper processes.

In the area of biomedical flows there is a trend towards image-based CFD where it is possible to create subject- or patient-specific models of different human organs, for example the human heart and the larger blood vessels to understand stenotic and/or aneurysmal flow, or the nasal cavity to simulate the dispersion of aerosols. In any such case the trend is strongly directed towards the application of increasingly more sophisticated computational tools. The simulations in the field of biomedical flows have recently migrated from classical, industrially available RANS-type methods to LES modelling. This step has only been enabled by the introduction of supercomputer performance and is of the utmost importance: the RANS models are simply inadequate to describe the complex turbulent flow field that occurs in a stenosis (usually followed by a dilatation). In such areas advanced methods must be considered that can correctly capture time-dependent flow structures such as vortices and recirculation, as well as separated flow. Accounting for wall motion or taking the actual behaviour of the vessel wall into account calls for a fluid-structure interaction (FSI) approach. Hence it is necessary to couple a fluid solver with a solid mechanics solver. Such problems can either be solved by developing a monolithic code, where all the physics is solved at once, or by utilizing a multi-physics approach and coupling the different independent solvers. The computational burden will, in such cases, increase by at least ten times.

There is a trend towards simulation driven optimization of interventions where it is now possible to simulate, for example, the balloon dilatation of a stenosis. Such methods require a substantial amount of computational power since several scale-resolved fluid flow simulations are run to optimize the intervention during intervention planning and hence, on an individual basis, predict the outcome of a particular intervention. Typically, a minimum of 50 million computational cells is required for a model of the human aorta.

9.4.1 Potential breakthroughs

- Paper process simulation for design of material properties.
- Methods for patient-specific intervention planning where, for example, the outcome from a stenting or a balloon-dilatation procedure can be simulated and evaluated.

9.5 Mechanical Engineering and Engineering Materials

The area of solid mechanics has been at the forefront of computational mechanics for several decades, ever since the development and popularization of the finite element method in the late sixties/early seventies. The ability to design structures with different types of complex materials, metal as well as composites, has revolutionized its industrial utilization. Furthermore, such analysis is frequently performed at either very low or very high temperatures. The need to handle heat transfer problems cannot be overestimated. Cooling of hot critical components remains one very important aspect and while even heat conduction modelling may be difficult due to a lack of reliable data on material properties, convection and radiation are both even greater challenges. Convection is strongly coupled to the turbulent behaviour of the flow whereas, in the case of radiation, both actual geometry and the complex interplay of transmission at different wavelengths must be handled. The development during recent decades has also been phenomenal when it come to applications including transient analyses of very complex geometrical structures, vibration and fatigue problems. The latest applications include reliability analysis of, for example, cars undergoing numerical crash tests subjected to realistic loads.

Structural optimization as well as emerging areas such as topological optimization are currently increasing in popularity. The industrial relevance for these areas cannot be overstated, but the structural codes of today do not scale well to thousands of cores. In certain optimization applications, however, this can be counteracted by the fact that several cases are run in parallel, exploiting the "embarrassing parallelism" such processes exhibit. One future goal in the field of solid mechanics is a more complete merger with the area of engineering materials to address the need for thorough analyses of fatigue and/or crack propagation in complex and industrially relevant applications such as gas turbines and jet engines. There is a significant need for developing very accurate models of engineering materials that can be included in expected life span simulations that are used for prediction of stress limits as well as service intervals.

Another area of increasing importance is industrial applications of ab-initio simulations for the development of new materials. Not only can completely new materials be modelled, but also modifications of existing materials as well as deformation mechanisms and processing pathways.

From having relied on mainly thermodynamic modelling for many years, the materials industry is now employing ab-initio techniques as an established development tool.

Academia and industry is working side-by-side in order to tackle very complex problems within an industrial setting. The fundamental work including the description of future computational capacity requirements within this field has been covered in the Chapter on Material Physics but the importance of this area is further stressed here.

Advanced modelling shortens industrial development cycle times and gives a competitive edge in a global environment with increasingly tougher competition. In chapter 7 one example of a major effort outside the EU and Sweden is described, the US Materials Genome program. The US materials industry is certain to benefit from this extensive initiative.

One specific case where advanced materials modelling will be of paramount importance is the Raw Material initiative from the European Commission as one of several major societal challenges. The initiative covers subjects such as substitution of scarce metals (or "critical minerals"), new methods for recycling, enabling the usage of waste deposits ("urban mining") and understanding factors influencing rock excavation for creating more environmental friendly mining methods. The success of developing all these areas will be highly dependent on advanced and efficient modelling resources for materials. Swedish industries and researchers are taking a prominent role in preparing for the agendas within the initiative.

Another case is the increased use of light construction materials, such as fibre reinforced plastics (CFRP) and high strength special steels. CFRP is expected to revolutionize, above all, the transport industry leading to vehicles with lower energy consumption and less carbon dioxide release, again addressing important societal challenges. Understanding the properties and manufacturing of lightweight materials as well as deformation mechanisms related to their use will need modelling tools.

One example of already applied and successful use of ab initio methods within an industrial context is the design of tooling concepts for metal cutting of modern lightweight materials. Rising costs of energy and raw materials put high demands on the aerospace and automotive industry, resulting in the replacement of traditional materials with modern lightweight alloys, such as aluminium, titanium and metal matrix composites. Most of these materials are expensive and difficult to machine. Chemical and diffusion wear are important destructive mechanisms, causing extensive material removal from the cutting tool and reduction of tool life. By applying ab initio calculations, industry researchers have been able to design cutting tools that minimize the chemical degradation of, for example, the thin films used as wear-protective layers on top of cemented carbide based tool materials, giving enhanced performance and end-user benefits.

Connected to the increased use of new construction materials are new demands on the machining tools for manufacturing parts and components out of them. Here we can see a trend towards unconventional machining tool components such as superhard materials like cubic boron nitride and diamond. These materials are manufactured using high pressure, high temperature processes and for understanding the thermodynamics and kinetics of these manufacturing processes it is necessary to use modelling, since it is impossible to make the analysis using conventional experimental methods. Again, there is also a need to understand the deformation behaviour of the super hard materials during their use in machining operations.

9.5.1 Potential breakthroughs

- Virtual prototype testing, for example in crash analysis, in automotive engineering.
- Development of materials with tailor-made properties such as surfaces with load carrying capacity for extreme conditions (high temperatures, harsh chemical environments, high mechanical loads).

- Replacement of scarce or hazardous metallic elements in the design of industrially relevant materials for the automotive, aerospace and machine tool industries.

9.6 e-Infrastructure requirements

The e-Infrastructure requirements for engineering science must include both capability and capacity resources to cover the broad spectrum of research from basic engineering research to direct industrial applications. The main requirements are capability/capacity machines, storage, advanced user support and code development.

Capability vs capacity: A variety of different computer architectures should be available to researchers, ranging from foundation-level systems to massively-parallel systems. Also, a few machines designed for specific tasks (for example having large fat nodes, visualization capabilities, or GPU clusters) should be available. It is important that Swedish researchers have access in Sweden to large-scale (at least Tier-1) systems to develop their own massively parallel simulation codes, and to gain experience in this area. Collaboration within Europe (e.g. PRACE and DECI projects) is instrumental for Swedish research to get access to the largest Tier-0 computers.

Storage: Reliable storage solutions that allow researchers to generate and store large-scale data, in both the medium and long-term (short-term storage is well organized at the computer centres already). A thorough strategy needs to be devised of how storage is being provided to individual groups and how collaboration (nationally and internationally) can be organized efficiently.

Advanced user support: Both application experts (sitting at computer centres) and research engineers (sitting in research groups) are required to allow for efficient e-Science application development. These experts should bridge the gap between the domain/application scientist, computer centres and pure method developers, possibly also taking into account industry users and their usage patterns. A long-term funding strategy needs to be developed to support this.

Code development: In order to cope with the rapid development of computer architectures, special emphasis should be put on the development of new simulation methods and codes, not only on a theoretical (method development) level but also in actual implementation on current and future supercomputers. Such programming activity, traditionally difficult to fund, is however crucial for the future ability of scientists to perform large-scale simulations. Note that code development is fundamentally different from user support.

Panel Members

- Professor Matts Karlsson, (Chair) Department of Management and Engineering, Linköping University. Expertise: Applied Thermodynamics and Fluid Mechanics.
- Professor Xue-Song Bai, Division of Fluid Mechanics, Lund University. Expertise: modeling of turbulent reacting flows.
- Dr. Ingrid Reineck, Sandvik Machining Solutions. Expertise: Material science.
- Associate Professor Philipp Schlatter, Mechanics, Royal Institute of Technology, Stockholm. Expertise: Turbulent flow.

Further input has been obtained via contacts with researchers associated with large-scale projects at SNIC-funded centres.

10. SOCIAL SCIENCES, HUMANITIES, EDUCATIONAL SCIENCES AND EPIDEMIOLOGY

- *Dramatic expansions of digitized data demand improved systems of curation (from collection to dissemination) and greater cooperation across disciplines and national boundaries to ensure common standards and data sharing.*
- *New methods for data security – including anonymization and copyright protections – are required to enable increased access to personal information and proprietary data for research.*
- *Investments in human capital will be extensive in order to bring a diverse community of users to levels of expertise and creativity required to fully exploit e-Science possibilities.*
- *Possible breakthroughs include new understandings of:*
  - *Cultural heritage & contemporary production*
  - *Pre-historic social & economic organization*
  - *Genetic susceptibility to illness & disease*
  - *Contextual influences on cognitive development and academic performance*
  - *Segregation across time and space*
  - *Political discourses and their transmission across time and place*
  - *Intergenerational transmission of inequality*

The social sciences, humanities, educational sciences and epidemiology study humans and their social and material products. They investigate such diverse phenomena as mental and physical processes within individuals, interactions between individuals, organization of economic, social, cultural and political life, and the creation of and meaning of artefacts. They analyze data ranging from the millions of biological elements that comprise a particular human being to physical and social structures created by collections of humans over centuries and across the globe. Methods of data collection and analysis are often highly specialized and labour intensive. Research has for too long relied on small-scale studies, underestimating by orders of magnitude the true costs of understanding the human condition. Considerable diversity exists in the extent to which scientists in these fields use, or are even aware of, potential e-science solutions to the questions they attempt

Figure 10.1: Traces of Human Activity: Footprints 3.6 million years ago, Laetoli, Tanzania Reference: replica exhibit in the National Museum of Nature and Science, Tokyo, Japan. Image from http: // en. wikipedia. org/ wiki/ File: Laetoli_ footprints_ replica. jpg , downloaded 2013-09-26

to answer; on average, the level of knowledge and use is low in comparison with the natural sciences.

That situation is changing very fast. New methods of data collection have generated an explosion in the amount and variety of research data on human endeavours. Data are increasingly "born digital"; survey responses, texts, speech, images and administrative records are directly recorded in digital format. In addition, huge amounts of data recorded or preserved in other forms have been or are in process of being digitized. e-Science has also revolutionized analytic capabilities in these fields, as in others, - iterative model-fitting, spatial analysis, microsimulation, high-throughput "-omics", such as genome-wide SNP genotyping or metabolomic data, and deep linguistic and content processing. The potential for "big science" in the social and educational sciences, humanities and epidemiology is here, but new and improved e-Infrastructures are required to make it possible.

The massive increase in volume and variety of data will demand greater coordination to cross disciplinary as well as national boundaries. For example, if every data point on every individual and every artefact were linked by coordinates of time and place, humanists, social scientists and epidemiologists could generate new questions and answers by working together on the same set of data, organized historically and geographically. Possibilities also exist for linking such data with data from the natural world on climate, and geographical or geological features. Search capabilities are becoming of much greater importance as the volume and variety of data increase. Furthermore, because many of the data pertain to living persons, issues of privacy arise in the combination of data from different sources.

e-Science challenges

1. Faster and more accurate methods for digitization of data recorded in other formats, including translation into structured formats suitable for analysis.

2. Increases of orders of magnitude in the size and complexity of data, demanding levels of computing power and storage much closer to those of the natural sciences.
3. Improved data curation – selection, digital transformation, organization, documentation, storage, protection and distribution – including the development of common standards across varieties of data types.
4. e-Science solutions for data security, including anonymization and copyright protection.
5. Extensive investments in human capital to enable scholars in these fields not only to accelerate and improve the precision of their research, but also to be able to imagine the questions that e-Science can help them answer. As a result, greater demand for computing capacity.

The following science cases are presented to exemplify these needs. The cases focus on different types of data used in the social and educational sciences, humanities and epidemiology. They do not, by any means, completely represent the full complement of disciplines and methods in these fields; the panel is aware of many other research areas that would be enhanced by improvements in e-Infrastructure but could not be fitted into a single panel's report. In combination, however, they identify the overlapping needs for e-Infrastructure that will also apply to other research areas.

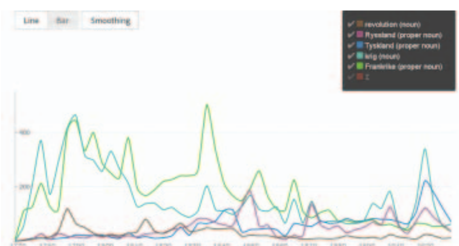


Figure 10.2: rTime Series, Word Frequency in Digidaily (krig=war). Image: Språkbanken, Gothenburg. University.

10.1 Language

Language-based research deals with all of recorded history and that which continues to be created from moment to moment. The amounts of available text and speech have grown far beyond the capacity of even the fastest and most multilingual reader. As noted above, much of today´s text and speech is "born digital", as in the billions of words of Swedish tweets and blogs, and older historical texts are being digitized. Litteraturbanken now offers close to one thousand books of mainly out-of-copyright classical works of Swedish literature in thoroughly proofread digital versions, and the Digidaily project (a collaboration between the Royal Library and the National Archives) has digitized some 300,000 pages of Swedish newspapers from the last 300 years. However such data are, for the most part, only in picture form. In order to make them accessible for searching and textual analysis, they must be transformed into "structured data" using high-quality, robust language technology in a basic infrastructure of standardized data and metadata formats, content models, and APIs (An Application Programming Interface (API) is a formally defined format for a computer program to communicate with other computer programs, enhancing modularity in large software systems).

Current e-science infrastructures for textual analysis include those being developed at Språkbanken, University of Gothenburg, where efforts are also under way to digitize the Swedish cultural heritage. The field is extremely dependent on international infrastructures, such as CLARIN ERIC and ISO TC37 [1]. In order to make breakthroughs in language-based research, e-Infrastructures must offer the capacity to access and formally manipulate the content of vast amounts of text or speech in many languages.

10.1.1 Potential Breakthroughs

- Comparisons of discourse in social media to that elicited in opinion polls, in order to identify adjustments that would be necessary to obtain accurate information about how public opinion evolves in real time in different kinds of polities; if data are identified by time and place, they can be used to study the diffusion as well as the disappearance of political ideas and actions.
- Understanding how such central human concepts as CULTURE(D) are perceived in different ways across time and space. By linking texts and speech to persons, it is possible to identify the carriers of culture, their social networks, and the institutional arrangements that may foster one or another understanding of culture(d).
- Testing the validity of the Aarne-Thompson folktale motif classification for a very large and varied folktale material in many language´s
- Through the use of avatars to represent video transcripts of student-teacher interaction, identify the behaviours and interactions, verbal and non-verbal, that produce effective teaching and learning.

1 Common Language Resources and Technology Infrastructure (CLARIN) is a European Research Infrastructure Consortium (ERIC) aiming to improve access to digital language data and provide tools for use in research; see www.clarin.eu. Technical Committee 37 (TC37) of the International Organization for Standardization (ISO) develops international standards for terminology and other language and content resources.

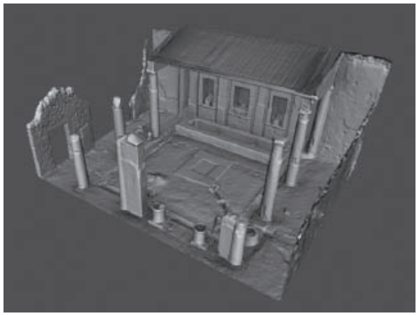


Figure 10.3: Orthomodel, House of the Bronze Bull, Pompeii. Image: The Swedish Pompeii Project, Lund University.

10.2 Images

Archaeology and the visual arts have been oriented towards a "spatial turn" for some time; the first models appearing 15-20 years ago. Large amounts of visual data, including video, are, however, disappearing into a black box because researchers have no common standards or archives through which they can share such data. In the past several years, digitization has become much more accurate and new software tools have been developed to make use of the full complement of digital information about a given image. The research has now arrived at a point where standardization and archiving have become central to further progress.

A pioneering use of e-Science infrastructure is the pilot project to visualize standing structures in Pompeii. The fieldwork is recorded by means of digital photography montages (2D documentation) and scanning (3D documentation) in a digital, open access platform using a syntax based on images (www.pompejiprojektet.se/insula.php). The platform provides a comprehensive overview of one city-block through a systematically ordered presentation of its basic spatial constituent, the room (each represented by images of at least its four boundary walls). The digital visualization permits passing walls, thus enabling immediate assessments of wider contexts, incomparably more rapid than real time investigation within the ruins. Further, scanning with 3D modelling furnishes a new and stunningly rapid means of acquiring fully covering basic documentation.

---

2 ARIADNE (http://www.ariadne-infrastructure.eu) brings together and integrates existing archaeological research data infrastructures so that researchers can use the various distributed datasets and new and powerful technologies as an integral component of the archaeological research methodology.

3 DARIAH completed a preparatory phase and has been working on achieving ERIC status in order to begin construction. The Swedish National Data Service (SND) has participated as an associated partner in DARIAH.

The Department of Excavation, Swedish National Heritage Board, responsible for approximately 30-40% of archaeological fieldwork conducted in Sweden, owns some 3800 individual databases, all based on a similar spatial, geographically-based syntax (GIS). If these databases were structured to communicate with each other then it would be possible to search for common as well as distinct features across space and time. A common data structure would also enable linkages between museum objects and their archaeological history. Current work on such linkages is being conducted in Sweden but e-Infrastructures are needed to coordinate this effort with the syntax for data registration advanced by the European ARIADNE project 2. Sweden currently participates in DARIAH (Digital Research Infrastructure for the Arts and Humanities), a European-coordinated technical infrastructure to improve and reinforce digitally based research in the humanities 3, and DC-NET (Digital Cultural Heritage Network), an initiative to digitize Sweden´s cultural heritage 4.

10.2.1 Potential Breakthroughs

- E-infrastructures to navigate the 3D model and, on demand, to switch directly to databases such as the digital platform, could revolutionize archaeology, providing new ways to relate to ancient life. The technology can easily be adapted to other sites, allowing for comparisons not only of ancient life at different times in the same place or different places at the same time, but also the development and practice of archaeology.
- Works of multimedia art created with digital equipment are fixed in terms of the technology used at the moment of creation; documentation of the original set of digital media/instruments used and how these are made to interact, the characteristics of the place of creation, and a chart on the ideas creating the work of art enables the study of how available technology influences artistic expression and the possibility to identify other dimensions of artistic change independent of technology.
- Complete digitization and documentation of museum collections under way at EUROPEANA [5] would increase possibilities for identifying migration patterns and cultural transmission in the past.

4 Sweden´s participation is coordinated by the Swedish National Archives.

5 EUROPEANA (www.europeana.eu) is a single access point to books, painting, films, museum objects and archival records that have been digitized throughout Europe. It also supports knowledge exchange between librarians, curators, archivists and the creative industries.

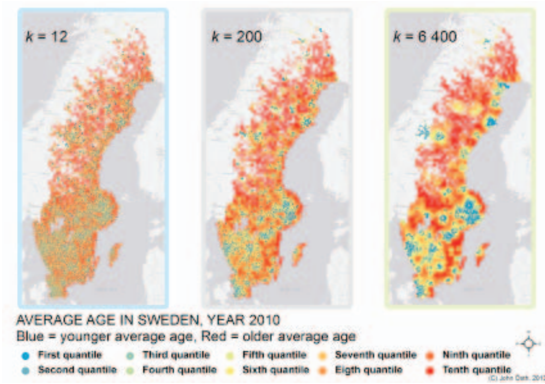


Figure 10.4: Average age from geo-coded register data for neighbourhoods defined by 12, 200 or 6400 nearest neighbours. Image: John Östh.

10.3 Place and space

Over the last decades, geography and other spatially-oriented sciences have been paying more and more attention to digitized and geographically-coded resources and to computer-assisted technologies for the analysis of spatial relationships. The technologies are usually referred to as GIS or GIT (Geographical Information Systems/Technologies) and today encompass a wide range of methods facilitating the analysis of satellite imagery, planning, complex networks, transport flows and much more. Geographically coded data are also increasingly used to provide contexts for individual or institutional analyses.

The possibilities for research using geo-coded data on human populations have, however, hardly been tapped. Focusing on a particular place at a particular time, one may gather text, images, administrative records, survey responses, climate data, and more, to understand the entirety of the human experience in context. Focusing on a particular human behaviour, expression or artefact, one may link data across places and time to understand social networks, communication channels, or the spread of disease.

An example of such a novel technology is embedded in the EquiPop software which is currently under developed at the Department of Social and Economic Geography at Uppsala University and in early use by geographers and sociologists at Uppsala and Stockholm universities. EquiPop analyses individual-level geo-coded data, and identifies the K nearest neighbours for each individual, or all neighbours within a fixed distance. It is designed to find KNN in less organized geographies containing large unpopulated areas and "clusters" of densely populated space. Each set of neighbours can be described in terms of aggregate demographic and socio-economic characteristics, with distance decay algorithms that take into account spatial dispersion within the neighbourhood. Individually-determined neighbourhoods generate more precise and valid indicators of the individual´s economic, social, cultural and political context than neighbourhoods based on administrative boundaries.

10.3.1 Potential Breakthroughs

- Individualized neighbourhoods will dramatically improve the measurement and understanding of residential segregation along various dimensions, as well as estimates of neighbourhood effects or moderation of individual-level effects on such diverse outcomes as children´s school performance, fathers´ uptake of parental leave, unemployment and hospitalization.
- Geo- and time-coded data on vaccination centres and influenza events would enable more precise identification of mechanisms of contagion and the most effective distribution of influenza vaccine.
- Multi-level statistical models together with geo-coded data will enable observed variation in, for example, student performance, to be connected to the various spatial or organizational levels where the student spends their time, with implications for educational and residential policies.

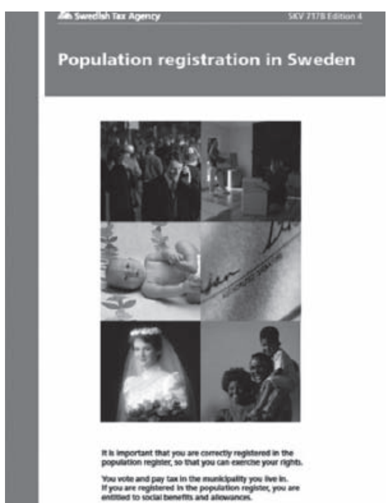


Figure 10.5: Swedish Tax Agency Brochure on Population Registers.

10.4 Administrative registers

In all of the Nordic countries and a few others, individual-level data on demographic events and socio-economic characteristics are maintained in population registers that can be linked to generate data over time on individuals and families. The Swedish Research Council has invested heavily on several fronts to increase the use of population register data for research, most notably in the Swedish Initiative for Microdata in the Social and Medical Sciences (SIMSAM) and in the Microdata Online Access (MONA) system developed and managed by Statistics Sweden.

Several problems remain for the most efficient exploitation of these data. The first problem is poor documentation from the statistical agencies. The SIMSAM-Infra project, under the umbrella of the SND, is working to generate complete and uniform documentation. Second, access to combinations of administrative data required to answer a given research question is difficult. The MONA system provides the technical capacity, but each collection must be associated with a particular project. As about 80 percent of the variables required are the same across projects, the duplication of work and the additional costs produce an extremely ineffective use of research resources. MONA has also led to a reduction in the types and amounts of register data that were, in the past, distributed directly to researchers, such as the random sample database LINDA. Third, interpretations of data protection and privacy laws have made it more difficult to share data with the international research community. SIMSAM-Infra plans also to review technical security systems for distributed data solutions, describe the ethical background to the legal system governing research on register-based personal data, and analyse bioethical issues.

10.4.1 Potential Breakthroughs

- Greater access to combinations of Swedish data registers will enable up-to-date estimates and more complete understandings of changes in living conditions and demographic behaviour. Data from the new (since 2012) Swedish dwelling register will be of particular importance for studying non-marital cohabitation and its implications for child and family well-being.
- Persons found in historical registers for Skåne and Norrbotten have been identified in modern registers, enabling up to 15 generations to be studied over time. If these data are made more widely accessible and linked to data from various health registers, it will be possible for the first time to study how the economic and social conditions of great-greatgreat (and more) grandparents influences the socio-economic status, family behaviour and health of their descendants.

- A common platform for Nordic administrative registers would provide unlimited opportunities to identify effects of differences in welfare state policies across time on economic, social and family behaviours and outcomes. Further advances in migration research would be made possible by following individuals across Nordic borders.

Figure 10.6: Computer-Assisted Personal Interview pad. Image: Tribe Research.

10.5 Surveys and Assessments

Subjective phenomena – attitudes, values, relationship quality – require direct reports from representative samples that are comparable across time and place. The same is true for comparative assessments of cognitive function or academic performance. Comparable cross-national data are particularly important for studying the consequences of economic, social and political change.

Sweden participates in several cross-national surveys and assessments: The European Social Survey and the Survey of Health, Ageing and Retirement in Europe have been designated as European Research Infrastructure Consortia by the European Commission. Sweden also participates in the Generations and Gender Programme that has applied for ERIC status, the International Social Survey Programme, the World Values Survey, the International Civic and Citizenship Education Study, the Progress in International Reading Literacy Study, the Programme for International Student Assessment, and the Trends in Mathematics and Science Study. In each of these programs, data are distributed to the international research community through a simple process of application, review and electronic download, and the quality of documentation is generally high. Documentation is also available for Swedish users through the portal of the Swedish National Data Service. Using SND's search engine, researchers are also able to find Swedish or other country-specific studies that may have common variables from which one can generate cross-national comparisons.

Long term needs for survey-based research include better access to survey data, including improved legal and practical possibilities to merge such data with register data across countries. Thanks to administrative registers, Nordic sample surveys can be based on random or stratified samples of the population, and survey time and costs can be dramatically reduced by linking survey responses to register data. These possibilities are, however, limited by requirements that surveys be conducted by the agencies themselves. Matching data demands both the development of practical standards for data as well as international cooperation of external organizations, including register data owners and research councils across Europe.

10.5.1 Potential Breakthroughs

- With greater standardization, improved documentation and more widespread access for existing and future surveys of the older population, it will be possible to specify more precisely how social, economic and cultural contexts influence the ageing process.
- If every sample survey could be linked to register data, both backward and forward in time, the cost of panel surveys for understanding change in attitudes and values, especially in response to expected or unexpected life events, would be dramatically reduced and the knowledge base increased by several orders of magnitude.
- Development of new methods to ensure that individuals cannot in any way be identified through combinations of characteristics included in a given data set would make possible greater international sharing of data, thus increasing the quantity and quality of research using a given dataset.
- Digitization, translation to structured formats, and analysis of handwritten material, drawings/sketches, etc. could generate a much richer understanding of individual development and performance in national assessments.

10.6 Clinical observations

Epidemiological studies are increasingly using a mixture of data collection schemes (clinical measures, questionnaires, registry data, and bio-specimens), all of which increase the need for e-Infrastructure for collection, storage and analysis. Mixed modalities are often used, and sample sizes are more often in the tens or hundreds of thousands than the hundreds or thousands typical of clinical studies or randomized controlled trials. Two new prospective cohorts are being developed in Sweden, EpiHealth and Life-

Gene, where over 500,000 individuals are responding to web-based questionnaires (with an average of 700 questions for the typical respondent), providing 40 ml of blood and urine that are processed and stored in 50 aliquots (requiring LIMS systems), physical measurements (which may entail up to 100 units of measurement and requiring LIMS for collection), and CRM to manage contacts with the participants. The participants will be followed longitudinally and with plans to link the data to national health registries. With the inclusion of "-omics" data, such as genome-wide SNP genotypes or sequencing and characterization of the proteome, there will be at least 2,000,000 data points on each individual participant at any one time-point. If imaging data are available, this number will double. Additional GIS information concerning place and time, frequently used for analysis of distribution of disease, adds yet another dimension.

Currently, epidemiological researchers are using predominantly local solutions for storage and analysis of data. Those working with highthroughput "-omics" data are using facilities such as E-max in Uppsala for analysis of data and sharing data for meta-analyses through EMBL-EBI. E-Infrastructure for biobanking is available through BBMRI.SE, but will surely need to be expanded. Many researchers use the Swedish National Data Service to make data available for others, while a few are archiving data through international repositories.

10.6.1 Potential Breakthroughs

- The role of infection/inflammation, gene-environment interactions or of epigenetic mechanisms in the exploding incidence of allergies and obesity could be identified by combining data from blood samples, dietary and other lifestyle information taken from mothers and fathers prior to pregnancy, from mothers during pregnancy, from cord blood and the placenta, and from the neonate (including blood and faeces) at regular intervals after birth, and linking to quality registries (for example Obstetrix), the prescription registry, and the national patient registry. Such data could be generated in a full scale roll-out of the LifeGene Born into Life Cohort.
- Combinations of data from population registers with the Patient registry and Prescription registry would make possible studies of the long-term effects of anaesthesia on child and adult outcomes (such as school grades, performance on national standard tests or military conscription tests, occupational attainment or dementia). Further links to genetically informative information, such as risk in family members (by linking to the Multi-Generation register) or by obtaining relevant genotype data from

a nested case-control material, would enable tests of differences in outcomes of operations for susceptible vs. non-susceptible individuals.

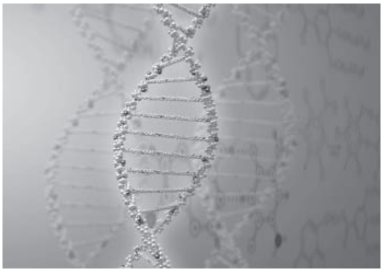


Figure 10.7: Double Helix. Image: Matton

10.7 e-Infrastructure Requirements

e-Science is essential for the development of the social and educational sciences, humanities and epidemiology. As noted above, research in these fields has not received the same kinds of investments as other areas of enquiry, and has therefore not yet attained the stature of "big science". The explosion of digital data is a stimulus for change in which increasing demands will be placed on e-Infrastructure.

Capability and capacity. Until recently, computing power has been more than sufficient for research in the social and educational sciences, humanities and epidemiology, not coming close to the demands of several disciplines in the natural sciences. Today, however, many researchers in these fields have access to billions of data points and require high-capability computing and associated storage to process the data. In language research, the content of vast amounts of text or speech in many languages must be simultaneously manipulated. The conversion from text to structured data requires high-performance computing capabilities, even more so in the case of speech and video data. Even fairly low-level linguistic processing of larger volumes of text may take days or weeks on a high-end server. The iterations required for such software as EquiPop to create a single neighbourhood variable for each individual in the population are enormous, requiring large amounts of RAM and digital storage, requirements that are also typical of other spatially based software. For biological data, current facilities such as E-max in Uppsala for computation, the EMBL-EBI for data-sharing, the BBMRI.SE (with its ERIC for biobanking), and the Swedish National Data Service will surely need to be expanded to accommodate the needs of epidemiologists and interdisciplinary research on social context and health. With the increase in data points and complexity, new bioinformatics tools for analysis such as those used in computational biology will become the standard. As research in these fields becomes more and more "digital", computing capacity will also face dramatic increases in demand.

Storage. Both language- and image-based research will require vast amounts of fast-access, flexible data storage solutions. The same can be said for geo-coded data and aggregates that can be constructed for each individual. In clinical research, e-Infrastructure needs for data storage and linkages are expected to increase exponentially as more investigator-initiated data collections are started, and ongoing large-scale cohorts accrue more data. The number of data points will soon be on the same order as in astronomy. In register-based research, variables created for the entire population take up large amounts of storage space, but could be more efficiently managed if a given variable were made accessible outside the research group that created it, to avoid duplication.

Databases, including documentation and distribution. The data explosion in the social and educational sciences, humanities and epidemiology is not yet over. e-Infrastructures are required to more quickly handle the vast quantities of historical documents, artefacts, and geographical and time coordinates that could be generated and/or added to existing databases. More immediately feasible would be improvements in the possibility to link researcher-generated data (sample surveys, clinical observations) data from Swedish registers; to link data from population registers in MONA with the registers held by Socialstyrelsen; and rapid evaluation and linking of the new dwelling register to other population registers. In the longer term, e-infrastructure must be developed to enable a common Nordic register platform. The work could begin with a pilot project using aggregated data. Such a project would facilitate inter-Nordic discussions of necessary legal change and multi-national agreements required for a common platform with individual-level data, including data from different countries about the same person.

The exponential increases in volume and variety of data make clear the need for extensive expansion of data curation. Although data is increasingly produced in digital form from the start, the rest of the curation process is quite slow, carried out in separate and sequential steps, highly localized and inefficient. E-infrastructures are needed to streamline the process of data transformation, organization, documentation, storage and distribution from the moment data are selected and acquired.

Curation also requires development of compatible syntaxes, common software and publication formats for all types of data. Images must be documented in terms of the mechanisms of their production and analyses must be recorded, documented and preserved (for example the processes of archaeological excavations, context and response to artistic creation and performance, and the mounting and dismantling of exhibitions). Methods are required to preserve a finished visualization as a set of syntactically organized data. Swedish participation in international infrastructures where standards are developed and data are archived – ARIADNE, CESSDA, CLARIN, DARIAH, DC-NET, and EUROPEANA – is essential for breakthrough research in these fields. Efforts underway in the SIMSAM Infra project will vastly improve documentation of register data in Sweden and across the Nordic countries and is therefore also an immediate need in data curation.

Software. Many software developments – such as EquiPop – are already on the horizon, but others have not yet been initiated. Specialized software is needed for image data to compare and tag diagnostic features within 2D- and 3D-models. New software is also required to identify individual instruments used in importing and post-processing data. Even more demanding of e-Infrastructure are reconstructions (as-was models of lost historical spaces) based on scanned data, whether created for scientific or museum use, or for the preservation of the original "creation context" of multimedia art. Geo-coded data and spatial statistics also add multiple dimensions of complexity to data management and analysis, requiring novel software solutions. Software is needed to link large-scale data of all sorts in a manner that allows for simultaneous search – a Google for science with greater precision and security. Software must be created or adapted for parallel computing, making better use of Sweden´s existing supercomputers for the increased requirements for data processing and analysis.

Much of the data in the social and educational sciences, humanities and epidemiology, requires a high level of security to protect individual privacy and/or copyrights. In the case of text and image data, software must be developed to enable search and analysis while protecting creators´ interests. Software and statistical solutions are needed to ensure the protection of individual identities while at the same time increasing access to data by researchers throughout the world. Geo-coded data on individuals greatly increase the risk of identification and therefore increase the demand for such solutions. The work of SIMSAM-Infra to review technical security systems for distributed data solutions describes the ethical background to the legal system governing research on register-based personal data, and analyses bioethical issues critical for the next several years.

User Support. The heterogeneity of experience with e-Science tools and their possibilities will continue to increase and require increased investments in human capital to make the most of the data that have been and can be generated. Geo-coded data and spatial statistics add multiple dimensions of complexity to data management and analysis, requiring a great deal of user support. Most large scale international survey infrastructures provide user support of varying degree, from courses to direct support of specific data issues. To increase national usability, however, user support is needed at the national level. The SIMSAM nodes and research school are critical for maintaining a cadre of researchers who understand and can master the possibilities of administrative registers. Ideally, these efforts could be expanded to the Nordic level, especially in relation to differences between registration processes, coding and management of data for research. Nevertheless, there will always be a need for continued support for users as they design, collect, record link, analyse and store data, particularly in provision of relevant meta-data.

Panel Members

- Professor Elizabeth Thomson, (Chair) Department of Sociology, Stockholm University, and Professor of Sociology Emerita, Department of Sociology, University of Wisconsin-Madison. Expertise: demographic data and methods, especially life histories from retrospective surveys and administrative registers; director of Linnaeus Center for Social Policy and Family Dynamics in Europe.
- Professor Lars Borin, Språkbanken, Department of Swedish, University of Gothenburg. Expertise: linguistically informed language technology, lexical resources for language processing, and language-technology based e-Science in the humanities and social sciences.
- Professor Mikael Hjerm, Umeå University. Expertise: survey research and anti-immigrant attitudes in comparative perspective; National Coordinator of European Social Survey in Sweden.
- Professor Anne-Marie Leander Touati, Department of Archeology and Ancient History, Lund University. Expertise: communication in and through images, particularly applied to archeological remains. Director of Pompeii Project, an open-access database for documentation and storage of archeological material through visualization and 3-D modeling.
- Professor Nancy Pedersen, Department of Medical Epidemiology and Biostatistics, Karolinksa Institutet, and Research Professor of Psychology, University of Southern California. Expertise: Analysis of genetically in-

formative populations (such as the twin registry) and large scale prospective cohort studies combining phenotypic, lifestyle and ”-omics” data.

- Doctor John Östh, Department of Social and Economic Geography, Uppsala University. Expertise: Geographic Information Systems and quantitative spatial analysis.

11. SUMMARY AND CONCLUSIONS

11.1 Ten major findings

Based on the texts provided by the panels, an overview of existing resources, international outlook, and technology foresight, a number of general noteworthy items have been identified. We list here the 10 most significant findings:

1. **Significantly enhanced resources and services will enable exciting breakthroughs in several disciplines and can be spearheaded by Swedish researchers.** This report provides evidence of potential breakthroughs that can be accomplished and what demands this will put on the future infrastructure. This evidence constitutes a solid documentation of possibilities for Swedish science to have large international impact through investments in infrastructure. It is, however, clear that many of these breakthroughs will require significant investments in the e-Infrastructure as well as the corresponding research agendas. It is also noted that the existing infrastructure and current level of investment is meeting current needs and enables internationally leading science to be done. Researchers do, however, have clear ideas about direct routes for more and better science through provision of more facilities, and are relying on continued investment and technology development to provide increased resources.
2. **Development of methods, tools and software within core disciplines is necessary to make breakthroughs**. To address some of the challenges outlined in the report, significant method development and implementation work needs to take place. Some of these efforts should be seen as a part of the infrastructure and it is important to broaden the view of e-Infrastructure to also contain software and software development.
3. **Advanced and long-term user support and human infrastructures are keys to e-Science adoption**. e-Science is a new approach and offers many possibilities but the breadth of knowledge of these techniques is currently small, and the difficulties associated with exploiting the approaches can appear very time-consuming, or even insurmountable, to an active researcher. It is thus important to invest in expanding the available expertise in the application of e-Science techniques, providing routes to educate and assist active researchers in making use of new methodologies. The range of increased efforts should span from standard support at centres to embedded research engineers in e-Science research groups.

4. **The simulation paradigm dominates the current Swedish needs for e-Infrastructure. A complementary and more data centric aspect of e-Science should be promoted.** While excellent research is being carried out in Sweden with existing e-Infrastructure, other countries are seemingly ahead in the introduction of support for data-driven e-Science. This means science that is based on exploration of large scale data coming from a wide range of application areas conducting simulations and experiments. Promotion of data-driven science is needed to maintain Sweden´s position as a leader in science, technology and business.
5. **International participation depends on access to national infrastructure compatible with international infrastructures.** e-Science, by its very nature, tends towards collaborative work with the sharing of applications, data and skills being facilitated by the distributed, networked tools and equipment involved. It is essential, however, that Swedish science be seen to be an active participant as well as an active contributor to international science and a national e-Infrastructure is thus necessary to create and maintain active international collaboration in e-Science, both within Sweden as well as on the international stage.
6. **User communities must be actively engaged in the prioritization, design, deployment and operation of e-Infrastructures.** Each user has a unique perspective on what e-Infrastructure best suits their needs and, of course, wants the system put in place to reflect that perspective. Consequently it is important to maintain a flexible and adaptable infrastructure which can meet the needs of each user. It is also important, however, to keep the e-Infrastructure developing in the light of new advances and offer services which go beyond what researchers are currently using, offering opportunities for new and potentially ground-breaking approaches to their problems.
7. **e-Social Science and e-Humanities are potentially very large users, but need active support like other communities new to e-Science.** There is increasing awareness of the potential of high performance data analysis methods in these areas and, coupled with the increased ease with which data can be gathered, social scientists and humanities researchers are becoming very interested in the use of these new methods to support research on much larger and more detailed data sets. Such approaches can be expected to provide insights that were not previously available and enable new and valuable research. Researchers in these areas, however, are also often less technically skilled, such skills having been less of a priority in past research work, and hence in their training. Consequently it can be anticipated that substantial support will be required during the start of this trend towards the use of e-Science methodologies in these

research areas, although this will reduce over time as the value of these new methods leads to their incorporation into the training of researchers in these fields.

8. **e-Science methods and tools are in increasing demand and will be instrumental in increasing interaction between tool makers and tool users.** Widespread use of e-Science demands new tools for visual analytics, data collection, data mining, and data curation. Support for development of these new techniques, the creation of platforms and tools to assist users to develop tools of their own, and direct user support and training should be considered an important priority. The development and tailoring of tools should be done in close collaboration between tool developers and users leading to new multidisciplinary projects.
9. **Effective secure and regulated access to, and distribution of, data, software and other resources must be enabled.** Research is to an increasing degree relying on networked resources and sharing of data. Access to these must be transparent and simple. It is thus important to further strengthen data related services, including not only hardware and software solutions but also institutional (legal, political) solutions and policy frameworks ensuring wide availability and distribution of resources. Included in the data network should be not only data generated by individual researchers but also open data collected by companies and other agencies.
10. **Improved co-ordination of the national e-Infrastructure and e-Science initiatives is needed.** As is clear from the brief account provided in the introduction, the current organization of e-Infrastructure in Sweden is complex and involves a lot of actors. Also the mandates and responsibilities of the actors and the interfaces between them are not always fully defined. To ensure that the most cost-efficient structure of the highest possible quality is provided, it ought to be investigated whether changes to the current set-up might be needed.

11.2 Recommendation for e-Infrastructure investments

This report provides evidence that e-Science methods and tools are gaining an ever growing importance across a wide range of scientific disciplines, in Sweden and globally. It is thus of the utmost importance that an internationally competitive and complete e-Infrastructure is provided to Swedish researchers. The national e-Infrastructure services must provide resources that enable Swedish scientists to continue to compete at the international forefront and participate as frontrunners in international research collaborations as well as making sure that the opportunities for ground-breaking research outlined in this report are realized and that the e-Science paradigm spreads to new disciplines.

11.2.1 Summary of needs

The panels have made an effort to gauge the current and future needs of e-Infrastructures. In the panel chapters these needs are described and examples of research projects motivating the needs are given. From the overall discussions among the panel chairs it is clear that even if we take into account the development of capacity, and given current level of investments in national e-Infrastructures, the needs will not be met. The items that are found to be most critical are:

- Capacity computing - HPC systems tailored for throughput of many independent jobs or large jobs with extreme scalability.
- Storage - Large scale storage solutions that are integrated with database and visualization services.
- Software - Efforts to develop new software to address new problems and new approaches.
- User support - The pool of human resources providing qualified assistance to users.
- Data access policies - Access and distribution of collected data is key to many fields, especially the emerging areas in social science and humanities.

Unless action is taken on these items the projected deliverables and breakthroughs described in the panel chapters will not be facilitated. In view of the demand-driven items presented here, it should be borne in mind, however, that the new possibilities afforded by better and more imaginative e-Infrastructures will enable new kinds of scientific research agendas that users are not currently foreseeing and so a purely demand driven e-Infrastructure will restrict what science can be done. The e-Infrastruture thus has to have a complementary technology-driven component to catch unexpected opportunities.

11.2.2 Development of an agenda for promotion of e-Science in Sweden

The responsibility for coordinating and supporting some areas in the e-Infrastructure landscape in Sweden is unclear today. This goes beyond the traditional hardware infrastructures. One example of this is found at the unfunded boundary between e-Infrastructure and e-Science research, where the adaptation of software to provide production-quality e-Infrastructure services must take place. Another area of concern is funding of and career paths for research engineers who will enable the use of e-Science. These engineers are often staff with research experience in an e-Science domain and with special competence in the use of the infrastructure. In other cases, they have research experience from one of the e-Science tool-making fields.

There is, as is shown in this report, an opportunity to accelerate the use of e-Science in Sweden and to contribute to, and indeed take the lead in, the on-going transformative e-Science evolution process. It is the conclusion of the work of the panels that an agenda for the promotion and acceleration of e-Science in Sweden should be developed and implemented.

The report is a result of the investigation initiated by the Swedish Research Council regarding the needs of e-infrastructure among Swedish scientists. The report provides an overview of existing e-infrastructures today, as well as seven subject field-specific chapters, describing the future needs within different scientific fields. The result show that the needs of adequate e-infrastructure continue to increase and e-infrastructures will soon be a necessity within almost all scientific fields.

Västra Järnvägsgatan 3 | Box 1035 | 101 38 Stockholm | Tel 08-546 44 000 | vetenskapsradet@vr.se | www.vr.se

Vetenskapsrådet har en ledande roll för att utveckla svensk forskning av högsta vetenskapliga kvalitet och bidrar därmed till samhällets utveckling. Utöver finansiering av forskning är myndigheten rådgivare till regeringen i forskningsrelaterade frågor och deltar aktivt i debatten för att skapa förståelse för den långsiktiga nyttan av forskningen.

ISSN 1651-7350
ISBN 978-91-7307-240-3