

e-IRG Workshop 11-12 June, Copenhagen



Around 90 participants attended the e-IRG workshop organized in Copenhagen on 11-12 June 2012, under the auspices of the Danish EU Presidency. The renewed e-IRG strategy was the main focus of the workshop; e-IRG aims at acting as the main advisory body in e-Infrastructures worldwide, as well as facilitating the coordination of all e-Infrastructure components. Topics included e-Infrastructure best practices, outsourcing e-Infrastructure to commercial providers, assessing the cost of e-Infrastructures, while ample attention was given to the user perspectives in several areas including networking, computing and data infrastructures.

Welcome and Introduction, Rene Belsø, (Danish delegate, workshop co-host)

Rene Belsø, as the host of the event, welcomed the workshop participants and opened the workshop. In his introduction he stressed that the focus of the event is on e-Infrastructures and the e-IRG strategy. He mentioned that diverging opinions are welcome to stimulate discussions, even if sometimes consensus is challenging.

Opening - e-IRG and the quest for Sound e-Infrastructure Strategy Development, Gudmund Høst (e-IRG Chair)

Gudmund Høst, chair of e-IRG, in his turn, warmly welcomed the attendees of the in Copenhagen to the e-IRG workshop. In his opening talk he introduced e-IRG, draw the audience' attention to the last e-IRG publications and announces the upcoming publications, namely the e-IRG Task Force Reports and the Roadmap 2012. Gudmund Høst finally provided some outlook about the new strategy of the e-IRG, which aims at strengthening the e-IRG as an advisory body, but also to transform e-IRG into a coordination platform for all e-Infrastructure components. He proposed that e-IRG should be more action-oriented and that the relations to ESFRI should be more cultivated. e-IRG as a coordination platform for all European e-Infrastructures should be open and transparent to stimulate the discussions, but also to facilitate the coordination and integration aspects. Finally, he pointed out that more attention is now needed to the use and management of scientific data and related expertise is valuable.

Welcome to the Danish e-Infrastructure Landscape, Børge Obel (Chairman of the Board, Danish e-Infrastructure Cooperation)

Børge Obel, the chairman of the board of the newly created Danish e-Infrastructure Cooperation (DeiC) presented DeiC, which is merging the activities of the Danish Center for Scientific Computing (DCSC) and Forskningsnettet, the Danish NREN. He said that in Denmark high performance computing is recognised as the third pillar in science besides theory and empery. It is the conviction of the Danish government that scientific computing and e-infrastructures are

crucial for the future and that “Big data” will appear everywhere. The aim of joining the activities is to find an appropriate sharing of responsibilities between national and local activities and the efficient resource usage.

Indicators - Mapping the Impact of European e-Infrastructures, Jorge-A. Sanchez-P. (European e-Infrastructures Observatory (e.nventory))

Jorge Sanchez started by stating that he has been working for the development and sustainability of e-Infrastructures in South Europe for many years. He was also one of the initiators of e-IRG during the Greek presidency in 2003 and it is good to be back in this environment. He started his presentation by introducing the need for an evaluation and an impact assessment framework with proper indicators and scoreboards, and reviewed related efforts and findings. Jorge Sanchez then introduced the e.nventory vision, being the establishment of the European e-Infrastructures Observatory, an on-line, single entry point platform offering visualisation tools, benchmarking indicators and a European-wide footprint. He then announced the opening up of the e.nventory service to the e-Infra community including 7 tools, 45 benchmarking indicators, 10.000 figures and 18 months of stakeholders’ feedback. Indicators include performance/capacity, usage/utilisation, budget/cost mainly for networking and computing), and other indirect or broader ones such as scientific impact, expenditure for R&D, unemployment, Digital Agenda indicators and others. Finally he demonstrated the e.nventory platform available at www.enventory.eu through a video.

Assessing the Costs of the European e-Infrastructure: Mission Impossible?, Matti Heikkurinen (e-FISCAL project)

Matti Heikkurinen provided an overview about the activities of the e-FISCAL project. The e-FISCAL project aims at an assessment of the costs of European computing e-Infrastructures. He explained the challenges to get the costs of computing e-Infrastructures for research and compare them with commercial equivalents. Furthermore he explained that not only the assessment of costs is challenging, but also that a prediction of the future based on gathered data is difficult since multiple factors influence the costs of ICT. The e-FISCAL methodology does not need the detailed costs of all the items in the past like the Full Cost Accounting one or their depreciation in the future like the Total Cost of Ownership one. These methodologies can work well for a single organisation that has recorded all such line items. e-FISCAL is rather based on estimated operational and capital costs per year. This works better for multiple organisations that don’t necessarily keep track of all the detailed costs. 26 computing centres from all over Europe have taken part in the first round of the survey and the first findings have been presented. It was shown that the cost per core may vary from 0.03 Euros to 0.08 Euros (median and average values for year 2011), PUE values are around 1,5 (median), while the cost for operations is more than half of the overall cost. Comparing to Cloud is still challenging, as benchmarking efforts are still not completed; a first comparison between the Irish HPC center ICHEC and Amazon HPC on the Cloud with the NAS benchmark shows an average performance loss of around 40% in the latter. More concrete results are expected in September 2012. Finally,

it was clear that when moving application from research-owned e-Infrastructure to commercial Clouds a part of the administration and adaptation effort will remain.

The Strategic Case of Commercial Hosting of Academic High Performance Computing, Kolbeinn Einarsson (Advania Thor Data Center, host to Nordic-HPC)

Kolbeinn Einarsson is a business manager of Advania, a merger of 9 IT companies in the Nordic countries, which provide support in ERP, hosting and related services. In terms of hosting Advania provides HPC resources located in Iceland. In the question whether to host in the Cloud or not Einarsson proposed to consider power price and availability of Green IT/carbon emissions. Network connectivity and the operational environment are also important. The benefits for outsourcing rely on the economies of scale achieved. Furthermore researchers want to focus on science, rather on operation. He stated that scientists have reservations of outsourcing their IT due to various reasons, including price, security/privacy or speciality of solution. In order to show the benefit to the users it is necessary to have the appropriate metrics, have clarity on the costs and find ways for a correct comparison. HPC centers can be operated more efficiently using green energy and naturally free cooling, which is secure, expandable and gives a clear operating costs. Compared in terms of risks and free cooling, Iceland has the lowest power prices. The Advania Thor Data Center is the first real data center in Iceland, a Tier3 centre consisting of 284 Racks with a PUE of 1.16, and an ISO 27001 certification. Advania is pioneering as cloud provider to higher education.

Access to European High Performance Computing - A PRACE User Perspective, Aake Nordlund (Computational Astrophysics, Niels Bohr Institute)

Aake Nordlund, a Danish PRACE user, talked about his experience using the infrastructure. He started by stating that PRACE has changed the conditions for supercomputers in Europe to the better. With his 40 years of experience with supercomputers his citation record would have been much smaller if only Danish resources were used. PRACE resources are the same size as the Danish resources but can be used in 10 days. As an example in solar modeling the magnetic field in a solar corona was shown as controlled by the magnetic field at the surface of the sun. To this aim, the velocity spectrum of the sun is computed. This example showed the importance and impact on visualization. A 3D visualisation with time animation is crucial, since it complements quantitative and statistical analyses by providing high cadence fluid snapshots and sub-sampled particle snapshots. However, these simulations also have strong requirements in high I/O bandwidth for data. Aake Nordlund underpinned his point by showing impressive visualizations of a record breaking simulation with 135 billion particles, what emphasizes his statement "the higher the resolution, the better it is". He pointed out that there are still some constraints which should be overcome like the initial restriction of PRACE projects access to only one year. Furthermore he pointed out that the European Tier-0 resources must be complemented by national Tier-1 and local resources, which is equally essential as an intensive training of the users. Another challenge he was facing is the rule that in PRACE data has to be

moved within 2 months and archived locally. For example, at FZJ there was no analysis machine available with a large enough memory. Furthermore there is also the necessity to give access to collaborators. The analysis of the data should be possible for other researchers without the need to copy it to their institution. If researchers would be given direct access would save a lot of time and network bandwidth.

Complex Computing and Data Infrastructure Challenges - the Bioinformatics Case, Kristoffer Rapacki (Center for Biological Sequence Analysis (CBS), Technical University of Denmark)

Kristoffer Rapacki reported about a dramatic paradigm shift in biology: from gene-by-gene analysis to all genes of an organism, proceeding further to entire genomes and finally studying genomes of entire populations. This was caused by a dramatic drop in costs for sequencing in 2008-2009. However, the growing CPU power, disk capacity and data transfer could not keep up with the need for sequencing. Since sequencing is a disruptive technology with a higher increase than Moore's Law, novel types of infrastructures are needed. And there are more disruptive technologies coming up in DNA sequencing that lead to an exponential growth in data. The ESFRI project ELIXIR tries to solve this situation, but the costs for preparing the infrastructure lack very much behind the expenditures for sequencing. Another challenge is the mismatch of tools, formats, data, etc. One purpose of ELIXIR is to align this data. From the caBIG infrastructure, which was an initiative similar to ELIXIR, one can learn that an infrastructure should not be designed top-down, since caBIG was terminated because of its highly-centralized and very normative way that turned out to be a mistake. Another important and very critical issue in the human genome project is the data security of person-sensitive data.

Architecture and Governance in Large IT Infrastructure Projects, Paul Brand (Stratix Consulting, Netherlands)

Paul Brand from Stratix Consulting gave an introduction in the architecture and the governance of large IT infrastructure projects. Mr Brand explained the different governance models; centralised versus decentralised and project versus line organisation. He said that governance needs to ensure a good balance between different interests, inducing tensions in any collaboration. These tensions exist between delivery and innovation, vision and execution, or central and distributed control. According to Mr Brand a decentralised control will promote innovation. Mr Brand adapts these theoretically aspects on today's Internet and the NRENs, which are not centrally controlled and driven by user requirements, only held together by centrally agreed and coordinated standards. Finally Mr Brand proposed open lightpath exchanges between instances of all layers of the existing architecture and argues for more distributed control to promote innovation.

Regional Development in the European Context - A Case Study for South East Europe, Ognjen Prnjat (European and Regional e-Infrastructure Manager, GRNET)

Ognjen Prnjat started by presenting the EU vision and the related model that GRNET has deployed for bridging the gap between South Eastern and the rest of Europe, i.e. through

regional infrastructure and support projects preparing for and complementing the European efforts. A series of projects were presented in the networking, computing but also on the policy level, such as the SEERA-EI one covering the whole area in the region including in some cases the Southern Caucasus countries. The regional benefits were highlighted getting all the countries up-to-speed and sharing experiences and know-how. Projects besides SEERA-EI included SEEREN/BSI/SEELIGHT, SEE-GRID and HP-SEE. The status of each of the project was presented and an advanced infrastructure was revealed including optical fibers in many countries along with cross-border ones, advanced Grid and HPC infrastructures and related tools. The regional cooperation on policy level was also highlighted, guaranteed through the signed MoUs setting the Network-Grid-and-HPC strategy until 2020. A joint pilot call between ministries on cloud computing has been also recently launched with a budget of 700K procuring public cloud services from industry. Overall, the regional cooperation brought significant progress and contributed to the development and sustainability of the national e-Infrastructures and the synergy, interoperability and integration with EU actions.

Scientific Cloud Computing Infrastructure for Europe, Maryline Lengert (European Space Agency (ESA))

Maryline Lengert talked about the establishment of a cloud computing infrastructure for Europe, named “Helix Nebula – The Science Cloud”, an initiative that started in 2010. The motivation has been to spend research funds in Europe to build a scientific cloud based on commercial services from multiple industrial providers. The initiative was initiated by ESA and later joined by CERN and EMBL and later on partial EU support was sought. HelixNebula tries to establish a sustainable cloud computing infrastructure for ERA, and to create a repository of scientific data to serve as a unique opportunity for scientists in Europe to comprehend major challenges. The “Science Cloud” shall consume and analyse data, and provide the results by using a layer that “normalizes” the input/output data so all the participating disciplines can use the results. All this shall be gained without vendor-locking commercial cloud provisioning. The project is a public-private partnership involving public resource providers as well as commercial European companies, and it shall be opened to other partners after the initial proof of concept has been implemented (this summer). Nevertheless, the pilot is slightly delayed because it was much more work to get a running infrastructure than expected. It was emphasized that the project is aligned to the e-IRG White Paper 2011.

Complex Computing and Data Infrastructure Challenges - the Case of Language Based Materials, Bente Maegaard (Centre for Language Technology, University of Copenhagen)

Bente Maegaard is vice director of the CLARIN ERIC, the Common Language Resource and Technology Infrastructure, a research infrastructure for the humanities and social sciences. The mission is to create a European federation of digital archives for language-based data, which comprises text, audio or video or a combination of these. Services include not just metadata search and retrieval, but also content search and advanced language and speech technology

tools operating in a service oriented architecture. CLARIN faces challenges that are typical for e-infrastructures in general and especially data infrastructures; these are the heterogeneous user base, long-term preservation of data, the persistence of identifiers, interoperability standards and the long-term funding guarantee. Legal, security and other issues include heterogeneous access and licensing systems, differences in national IPR legislation, and that in general access to language resources and archives is too limited. Computing is also getting more important, while knowledge sharing and education are also vital, especially given the varying level of expertise and that teams are scattered. Finally collaboration is key and the CLARIN ERIC is working in this getting as many countries together.

e-Infrastructure Needs for the Arts, Culture and Humanities, Antonella Fresa (Digital Cultural Heritage NETwork (DC-NET))

Antonella Fresa, from the Ministry of Culture in Italy gave a presentation about the e-Infrastructure needs for the Arts, Culture and Humanities. The amount of digitized material in the European Cultural sector is growing rapidly. This generation of data is accelerated by Europeana that is fostering the European cultural institutions. The needs of DCH include high quality information technology management, access facilities and interoperation. Fresa underlined that it is not a new infrastructure but a new approach based on national and regional interoperable systems using existing resources. The main actions required in the DCH area are to improve awareness, promote trust, establish priorities, consult stakeholders, and to promote international cooperation. The three main projects to take care of these actions are DC-NET: joint programming for DCH e-Infra implementation, INDICATE: international cooperation user case studies, pilots and DCH-RP: developing and validating a roadmap for digital preservation. DCH-RP focuses on the storage phase, which includes both long- and short-term preservation. All three projects are part of a wider process, which started 10 years ago among cultural institutions. This process is entering a new phase and time is ready to start working towards an Open Science Infrastructure for Digital Cultural Heritage in 2020. All stakeholders need to be active; cultural institutions providing digitised and aggregate data, e-Infrastructure providers together with user communities deploying the e-Infrastructure and developers including industry that will create new services and products.

National Archives - Strategic Challenges within Storage Cooperation, Kirsten V. Kristmar (The Danish State Archives)

Kirsten Kristmar reported about the strategic challenges within Storage Cooperation between the members of the Danish Heritage Institutions, like libraries, museums, and archives. The talk started from the Danish national archives, but also covered similar issues in other countries and on the European level by generalizing the presented issues: summarizing, even if the heritage institutions have different objects, the same tasks have to be done by all institutions. However they are using different formats and technologies. She concluded three main requirement groups that are essential during the daily business in the Danish Heritage Institutions: the support of the long-term preservation process which is using the OAIS model (Open Archival

Information System model), security aspects like ensuring integrity, and other requirements like the amount of data to be processed, I/O bandwidth, mass processing of data, avoidance of data conversion, and data searching. She outlined the increasing importance of European and International cooperation in the field of long-term preservation of data and the drawbacks of using different formats, and technologies. Problems are similar to other sciences with the need to handle the growing amount of data. The main advantage the Danish Heritage Institutions using in-house solutions is the full control of the system and the data.

Closing Debate - What are the Greatest European e-Infrastructure Challenges - Policy; Legal; Economical; Technological?

The closing discussion focused on whether the publicly funded e-Infrastructure services can be obtained by commercial providers such as cloud ones and how should the home-grown e-Infrastructures evolve in the future (and whether people vs. hardware should be subsidised). The data management challenges were also brought up several times.

Diverging views were raised that can be summarised as follows:

- The need to move from a communications infrastructure to a combined e-Infrastructure incorporating multiple e-Infrastructure components was raised. In other words, more interaction between the e-Infrastructure components is needed. The end-users need to have integrated and as easy as possible services and innovative tools and services need to be developed to assist their work. We should not reach the other extreme though, i.e. a single e-Infrastructure service monster.
- As soon as equivalent (to the current research ones) industrial commodity services emerge, they can be used by the different communities. These services can be mixed with innovative services from the research community. And such a cycle can continue. So, from a top-down shared e-Infrastructure, a hybrid owned/outsourced infrastructure scheme needs to be developed, mixing the best of the two breeds.
- To use such industrial commodity services, there is certainly a need to agree on interoperability standards and protocols. Of course in the beginning, there are no standards, maybe de-facto standards. After some time, effort on standards and interoperability needs to be dedicated, either adopt de-facto standards or define interoperable protocols.

On the other hand there were the following views:

- The collaborative nature of science that drives people working together requires shared infrastructure services such as EGI. So innovation should not dominate the discussion and the proper balance between innovation and a common service level needs to be kept to make e-Infrastructure services productive and effective. The benefits of interoperability and standardisation are obvious and should not be given lower priority in favor of innovation.

The issue of GEANT, EGI and PRACE projects portrayed as “monopolies” was also iterated.

- From one side it was stated that such projects get the majority of the funding in the area using it to have a common denominator service around Europe, something that hinders innovation. This side proposed to reduce their funding and challenge them stimulating competition and promoting innovation.
- On the other hand, it was stated that loads of projects are funded and that funding such projects are not controlled by the flagship e-Infrastructure projects. So there are not such monopolies at all and that there is a good part of innovation in these projects. Furthermore, the fragmentation in too many small projects is not good and sometimes there is no critical mass. In H2020 this needs to be addressed properly.
- On the last point there was consensus that using funding wisely is vital, and that interoperability and standards enable global collaborations and portability of applications in different e-Infrastructure environments.

The increasing importance of adequately managing scientific data including open access was also evident across all discussions. Diverse user requirements need to be understood and taken into account, and then appropriate governance structures will need to be developed. Interoperability and global standards are also vital in the data area; something similar to IETF in needs to be developed in the data world. Some kind of bureaucracy needs to be accepted and if the data field can do that, others can do it also (computation world).