e-IRG Workshop Summary

9th of March 2016, Amsterdam

Progress of the e-Infrastructure Commons

For the recordings of the talks and copies of the presentations used at this workshop, please visit: http://e-irg.eu/workshop-2016-3-programme

Workshop objectives

The main objective of this workshop was to determine the progress of the e-Infrastructure Commons[1]:

- **Integration and interoperability among e-Infrastructures** (national-regional-European, academic-industrial, networking-computing-storage)
- **Organisation, funding and revenue models** (national and European)

The Workshop Proceedings - Summary

The following are the main points of the discussion that took place.

Welcome addresses - Opening and Position Statement of e-IRG, Arjen van Rijn/Erik Fledderus [pdf]

Arjen van Rijn welcomed the participants and made an introduction to the Amsterdam Science Park. Erik Fledderus reminded the definition of the Commons and explained its vision. The workshop objectives around the progress of the Commons were presented along with the main sessions of the first day program.

Theme: User perspective - Session moderator: Arjen van Rijn

Wilco Hazeleger - Getting Science out of eScience [pdf]

Wilco Hazeleger presented an outlook of how science should be made in 2020, and what is important for it in the near future. In 2020, there will be much more data produced, the data will be increasingly complex and multidisciplinary and research will have ambitious goals. 'Big science' is driven by 'big societal' challenges and science itself is 'big' and complex. This is why eScience can play an important role, if a suitable complex e-Infrastructure for the increasingly complex and data intensive research questions can rapidly be developed. e-Infrastructures should be made accessible; as an example Hazeleger presented how things work in the Netherlands, where the

[1] http://e-irg.eu/commons

eScience Center plays a central role in enabling e-based science. It combines data and computational sciences along from various research disciplines and uses the e-Infrastructure to produce eScience instruments for enhanced science and services. The working model was analysed and some example cases were presented. In general, eScience bridges e-Infrastructure and scientific research, crucially takes a domain science perspective, helps multidisciplinary scientists and engineers and provides overarching technologies & cross-fertilisation. The themes commonly dealt with are data stewardship, software quality & sustainability, and e-skills (human capital).

Rene Belsø - Researcher perspectives on e-Infrastructure provisioning [pdf]

Rene Belsø presented the perspective of a researcher. The concept that most researchers would support is to have access to "an integrated living ecosystem of resources and services that is open, user friendly and accessible to European researchers and scientists, and continuously adapts to the changing requirements of research and science", just as it is stated in the definition of e-Infrastructure Commons. e-IRG defined the Commons as having three distinct functions: i) Coordination, ii) Provisioning and iii) Innovation
These functions present problems and questions on their own

- Coordination – Who is in charge of strategy, staffing and spending?
- Provisioning – Which e-infrastructures? By whom, for whom?
- Innovation – Who knows how in order to get professionalism and quality?

Can the European Open Science Cloud (EOSC) be used for research? Can it address researchers' problems? National and European e-Infrastructure providers are losing customer base. The "big users" increasingly prefer the ESFRI-style research infrastructure set-up, which does not always use or does not find e-Infrastructure providers to be "useful" and would, therefore, rather do it on their own, while the long tail of science goes to "GAFA" i.e. Google, Amazon, Facebook, and Apple.
There is an increasing need for the re-invention of e-Infrastructure providers. Some suggested solutions are to focus on academic professionalism, commoditisation, and market mechanisms.

Giorgio Rossi - ESFRI view on e-Infrastructures [pdf]

Giorgio Rossi presented the official ESFRI roadmap 2016. The ESFRI Roadmap identifies 29 Landmarks (2 new) and 21 Projects (6 new). All new entries have been evaluated for e-needs, all the Roadmap projects are monitored also for e-Infrastructure aspects and the landscape analysis has been carried out for all thematic areas and for the transversal e-Infrastructures (provided by e-IRG).
The evaluation process for new proposals was to assess both scientific merit and project maturity. e-IRG assisted with the strategy working groups. Out of the 21 projects several are data-intensive projects with special e-needs. Some such examples are EPOS (European Plate Observing System), ACTRIS (Aerosols, Clouds and Trace Gases Research Infrastructure), E-RIHS (European Research Infrastructure for Heritage Science). Of the 29 Landmarks, several are leaders in e-Infrastructure development. For example, PRACE (Partnership for Advanced Computing in Europe), and HL-LHC (High Luminosity Large Hadron Collider). There are also several data intensive RI projects (e.g. ESRF EBS, Lifewatch, and DARIAH ERIC).
The landscape analysis that followed provided the current context of national and international RIs. Finally, the ESFRI mandate on e-RIs was laid out. The Competiveness Council of May 28-29 2015 adopted conclusions on: "open, data-intensive and networked research" and "invited ESFRI to explore mechanisms for better coordination of Member States' investment strategies in e-infrastructures, covering also HPC, distributed computing, scientific data and networks."

An ESFRI Working Group (WG) on investment strategies in e-Infrastructures has been created, and the WG mandate is to draw also on the competence of the e-IRG and provide input and draft recommendations to ESFRI on how to best address the Council conclusions.

For further details, refer to http://www.esfri.eu/roadmap-2016

Panel Discussion: Giorgo Rossi, Rene Belsø and Wilco Hazeleger
Introducing the concept of 'Interplate' - the flat and the generic e-Infrastructure that should form the electronic basis of all the RIs.

Questions/Comments and Answers to them:

- Q: Is there any hope for a generic e-Infrastructure for all RIs?
  - A: There are generic e-Infrastructure parts for the ESFRI RIs, but much more is needed. And there is no need to separate the GAFA from the other parts of the e-RI, especially in the human capital side.
  - A: Further innovation is needed to develop this concept.
  - A: The services provided should rather focus on innovation for ESFRI projects.
  - A: There is a missing item – to be handled by ESFRI: big data storage, i.e. data produced by a project to be reused by another project or community. We need to understand what data various communities produce. We also need an e-Infrastructure (software) using their analysis tools, but also picking up some data for reuse.
  - A: Policy-makers could take advantage of the data outputs
  - A: The human capital and resources of all e-Infrastructures should be shared by all
  - A: The needs of the various infrastructures should define the flow of money spent to finance them. The user communities should be in charge of the funding (and not the e-Infrastructures themselves)
  - A: The e-Infrastructure providers really need to know the needs of the users.
- C: There is a disconnection between what we call long tail and the 'existing path' e-Infrastructures.
- Q: How can we drive research communities towards this e-Infrastructure?
  - A: We need to incentivise research communities to invest in developing their own domain services on top of the common e-Infrastructure. We should be looking at the hybrid model (researcher + policy makers) – not only investing in the e-Infrastructure, but also in the higher-level services.
  - A: The specific RI is part of the e-RI.
  - A: The national level uses the European horizontal e-Infrastructure.
- Q: Regarding the commercial infrastructure (GAFA) and the statement of the long tail. Are e-Infrastructures, which are necessary for e-science, a step ahead of the commercial infrastructures?
  - A: They are. If they see that they fit, they go for it.

Theme: Industry perspective - session moderator: Erik Fledderus

Tony Hey - Open Science in an ever changing e-Infrastructure landscape [pdf]

Tony Hey presented the bottlenecks preventing the e-Infrastructures (or e-RIs) from enabling excellent science. These can be summed up in the following way:

- Much of science is now Data-intensive. Astronomy, Genetics and Environmental Science all need the four "Vs": Volume, Variety, Velocity, Veracity.

- Open Access, Open Data and Open Science. Budapest and Berlin declarations and White House Memo in the US defined them. If used widely, they may enable “Reproducible Research” and will help sustainability of data,
- There are specific network requirements for data-intensive research. TCP and end-to-end performance play important role for research production, Science “Demilitarised Zones” (DMZs) and “Superfacilities” can offer solutions. The Science DMZ model provides the framework for building a network infrastructure that is more loss-tolerant. Europe needs to build Science DMZs and European Superfacilities (US already has started doing so).
- Industry, data scientists and e-Infrastructure are correlated. Data scientists are needed in the industry. Data engineers, data analysts and data stewards each have a discrete role. It is science that needs and often produces data scientists.

An important remark by Tony Hey was that Europe must become competitive to the US.

Michael Symonds - Industry Angle [pdf]

Michael Symonds presented the angle of industry on the workshops topic. Using Helix Nebula as an example, which started off as a technical challenge and was based on an EC-funded project, he mentioned the learnings to date, and provided a prognosis on the future. Helix Nebula aims, among others, to form a hybrid Cloud Computing Marketplace, by integrating European cloud providers and existing e-Infrastructures. It was quickly resolved that scientific research is a greenfield market for IT Cloud service provision. In fact, most researchers need a combination of at least three services. i)ad-hoc and “bursty”, on-demand processing facilities, ii) large-scale, long-term data storage and archiving, and iii) a platform for deriving valuable information on top of these infrastructure services. Additionally, it was determined that data storage provision and location are more of an issue than processing capacities.
Some important issues have arisen from this assessment concern: i) Governance and cloud services, ii) Data protection and iii) Cost calculations.
In order to address them, currently, suppliers are trying to add more value. For example, for Copernicus access platforms there is effort to provide a basis for the provision of a full end-to-end information as a service (INFOaaS) ecosystem and derive valuable and beneficial information from the available data. By working in a federated, competitive environment, and by reducing time-to-market, European industry can provide both the services required and also considerable further benefits. Europe would be endowed with high added-value assets, provided by European actors, rather than non-European ones.

Bob Jones - HelixNebula-Science Cloud - Pre-commercial procurement pilot [pdf]

Bob Jones started by presenting Helix Nebula's (a European cloud public-private partnership) solid strategic plan: i) Establish multi-tenant, multi-provider cloud infrastructure, ii) Identify and adopt policies for trust, security and privacy, iii) Create governance structure, and iv) Define funding schemes.
The Helix Nebula initiative has brought together research organisations, data providers, publicly funded e-infrastructures and European commercial cloud service providers to develop a hybrid cloud model with procurement and governance approaches suitable for the dynamic cloud market. HNSciCloud Joint Pre-Commercial Procurement plans to procure a joint science cloud platform for the European research community to combine services at the IaaS level into an environment that supports the full lifecycle of science workflows.
Helix Nebula expects commercial cloud services to play an increasing role in the computing models of Research Infrastructures. A hybrid cloud model leverages the investments made in both

the public and private sectors while ensuring trust and continuity. Mature technologies exist but integration, policy and governance requires careful attention. Changes to the procurement process in the public research sector are necessary to benefit from a dynamic Digital Single Market (See PICSE call to action http://www.picse.eu/). A pay-for-usage model can contribute to the sustainability of services by supporting their operational costs. The PCP/PPI funding approach can be a means of developing and deploying innovative cloud services. HNSciCloud is the first in a possible series of EC co-funded projects.

Panel Discussion: Tony Hey, Michael Symonds and Bob Jones
Questions/Comments and Answers to them:

- Q: What is the challenge about computing in US?
  - A (Tony): US leading facilities towards Exascale computing.

- Q: Erik: What is the role of the e-Infrastructure providers? E.g. Roles of GEANT, EGI?
  - A (Michael): EGI has been working with HelixNebula on integrating various of the EGI cloud in a federated manner and are looking also in identity management. Also came up with a good way of managing the required "inter-supplier" support, which is essential in a federated environment.
  - A (Michael): GEANT are supporting the technical integration via the networks and you can access the infrastructure resources.
  - A (Bob): There is a number of central horizontal services. GEANT gives access to resources to the whole research and education community. There is also AAI with eduGAIN and we would like to see this urgently deployed as a production service. A: EGI has mechanisms for the competence centres for providing human consultancy in capital for working in different disciplines, centralised helpdesk which would be needed to link both services from the public and private sector.
  - C: Bob comments the question of what is the role of e-Infrastructures: The real question should be what is the role of the IT departments on each organisation. We can act as IT departments and ensure the users have access to services coming for many suppliers. They do respect policies for data access, for security, etc.

- Q: Should we think about a procuring agent in Europe that would make sense to those cash flows and the policies for the different e-Infrastructures?
  - A (Bob, Michael): There is value of having some organisation doing it. Increased volume, can have an impact on how we can procure services.
  - A: It can also help centralise the terms and conditions for the procurements.
  - A: it is good for us if the education/ research sector can identify common requirements, terms and conditions.

- Q: Why is procurement an instrument that funding agencies would use to expand the RI and will RI comprise of commercial suppliers, and public funded ones depending on the model?
  - A (Bob): There is no reason why that can't be. It is difficult to understand the different rules from each organisation. But if we have a common model, it would be easy to implement it.

- Q: How much of the computing at CERN will be done at this cloud environment? Are there some investments from USA?
  - A (Bob): They have a production environment and there is pilot running in parallel. They want to expand the group of procurement; at the moment it is European (funded by H2020)

- Q: Are commercial providers far from understanding the research needs?

- A (Tony): It is just now maturing. Microsoft is learning the needs of the researchers.

- Q: Atos is participating in Helix Nebula and Indigo project. Is Atos considering the business model during the pre-commercial procurement?
  - A (Michael): Yes, Atos documented and identified two business models.

Panel 3 - Theme: Policy perspective - Session Moderator: Cees de Laat

Kees Neggers - The development of the e-Infrastructure landscape in the Netherlands [pdf]

Kees Neggers presented the history and current state of the e-Infrastructure landscape in the Netherlands. After a short overview of the history from the 60's until the 00's, and the 10's (i.e. '11 creation of NL e-Science Center, '13 merge of SURF and SARA, '14 restructuring of SURF), the current situation of SURF, its organisational and governance structure was explained in detail. In this historical review it was explained why e-Infrastructure governance can be a complex matter, since there is i) often a large number of stakeholders, ii) high upfront investment cost, iii) a possibility for large spill-over effects, and iv) dynamic nature and many interdependencies. In the Netherlands it was proven that e-Infrastructure Commons is an organisational rather than technical challenge. SURF 2.0 acting as a cooperative and thus it has several unique characteristics:

- More transparent and streamlined governance structure
- Increased sense of ownership and control for member institutions
- Better fitted for insourcing
- Reinforced managerial involvement
- Core business remains providing a national infrastructure for research and education

The current state of NL e-infrastructure Commons is good, but still requires work, it improves involvement of all stakeholders, it speaks with one voice towards funders, and it strengthened the role of SURF as national umbrella, while enabling users to play a more prominent role.

Barend Mons - European Open Science Cloud [pdf]

Barend Mons presented in a clear way the European Open Science Cloud (EOSC). The key challenges for EOSC to succeed were summed up in just five sentences. i) The lack of widespread awareness of the value of data and of incentives for data sharing is still dominant, ii) There is a lack of common standards to ensure inter-operability of data, iii) There is not enough hardware capacity for scientific computing, storage, connectivity, iv) there is fragmentation and lack of coordination over different scientific communities and countries, and v) the need to translate recent changes in privacy, data protection and copyright rules to the research data domain is dominant. It was explained that Open Data is more than disclosure, Open Data must be FAIR (Findable, Accessible, Interoperable, Reusable). The frame in which EOSC developed is based on i) trusted access to services & systems, ii) re-use of shared data, iii) across disciplinary, but with social and geographical borders, and iv) federated environment, across Member States. EOSC functionally and structurally adopts the "Internet approach" and supports Open Science and Open Innovation, and long term data stewardship. It needs to overcome a plethora of organisational and (inter)national obstacles. To do so, it must also establish new modes of communication and reward practices in the scientific community and train data experts. Data Management Plans will gradually become mandatory comprising around 5% of the total budget.

At the policy level it is suggested to i) take immediate, affirmative action in close concert with Member States, ii) close discussions about the 'perceived need', iii) build on existing capacity and expertise where possible, and iv) frame the EOSC as supporting Internet based protocols & applications.
At the governance level; i) aim at the lightest possible, internationally effective governance, ii) provide guidance only where guidance is due, iii) define Rules of Engagement for formal participation in the EOSC, and iv), federate the Gems across Member States.
Finally in an implementation level it is suggested to i) turn into an EC approved White Paper to guide EOSC initiative, ii) develop, endorse and implement a Rules of Engagement scheme, iii) fund a concentrated effort to locate and develop data expertise in Europe, iv) install a highly innovative guided funding scheme for the preparatory phase, v) make adequate data stewardship mandatory for all research proposals, vi) install an executive team to deal with international coherence, and the preparatory phase of the EOSC

Edward Seidel - Trends in Scientific Computing and Data [pdf]

Edward Seidel focused on the current trends in computational and data-enabled science and the need for coordinated e-Infrastructure. The fact that collaborations grow as problems become more complex is well known; less known is the fact that the same goes for systems and data sizes, too. Instruments, HPC and data science made possible (just this year) the proof of a 100-year-old Einstein's prediction on the existence of gravitational waves. Yet, HPC is only part of greater ecosystem! Breakthrough HPC systems support breakthrough science. Similar steps have been made in materials understanding and design, turbulence, and many more. It is real world science and engineering that require the integration of compute and data. More and more data is produced and shared in scientific communities. The conclusion is that we need e-Infrastructures that better support convergence and allow interdisciplinary science. A key challenge will be to create a deeply integrated compute- and data- environment that supports complex problem-solving for academia and industry. But there is a fundamental problem, long term planning horizons are needed before construction! Is this possible? Data must become an enabler for collaborative work. In summation, research is changing dramatically. Complex problems require collaborations at new scales and computing and data capabilities grow at unprecedented rate. There are several major infrastructure projects which are highly computing & data intensive, in which data services are needed for support, and a comprehensive, integrated approach would better serve science, and reduce costs. Overall, data sharing supports international cooperation.

Panel Discussion: Kees Neggers, Barend Mons and Edward Seidel
Questions/Comments and Answers to the presentations:

- Q: How do the communities' specific services fit into this picture in the US cyberinfrastructure? What is the US strategy around such specific services on top of generic e-Infrastructure services?
  - A (Ed)**:** With my experience (not speaking as NSF), in the US the community-based repositories and services are doing well. We are trying to connect them via the necessary national structures to be able to share their data. And a lot of the communities are not well-organised yet and do not have models to share data. So a lot of our services are for the long tail of science.
- Q (Cees): Is there too much policy in Europe or is it too little?
  - A (Kees): There is no lack of vision and it is easy to sit together and refine the policies. The challenge is to get the community together and act together. And this needs leadership.
  - A (Barend): We have to find a way to respond quickly to the new opportunities.

- Q: In a Dutch survey among 10.000 researchers only a quarter of them is willing to share their data. And more than half haven't heard of the terms software sustainability, data stewardship or research data management. Should we involve more the end researchers in these developments?
    - A (Ed): In the US after 2010, NSF implemented a rule that a proposal cannot be submitted without a Data Management Plan. And maybe initially we got lousy DMPs, but now everybody knows what this is. So this is a powerful policy.
    - A (Barend): Similar in the EU, but we still need more. The current impact factor in publications is not a good system. Additional measures are needed, namely rewarding for sharing data, data should be citable, etc. But researchers should be educated.
    - A (Kees): Why did we call e-Infrastructure Commons in the e-IRG documents: Because it is a shared responsibility to keep the e-Infrastructure alive. And I agree that education is important. And you also need an organisation that represents all the stakeholders (such as SURF in the Netherlands) to bring people together to act together towards a common interest.
- Q (Barend): Do we really need centralised computing and storage and exascale computers, or possibly a distributed approach of interconnected devices (e.g. 7 billion smart phones) is good enough?
    - A (Ed): Special machines can be built for special purposes and they are fantastic for the specific purpose, but then they cannot be used for other purposes and they fade out. There will always be a balance among centralised and distributed approaches and they are all important modalities. And exascale has specific areas/purposes that fit well.

Closing: Arjen van Rijn/Erik Fledderus

Erik: We need to understand whether there was progress in the e-Infrastructure Commons:

- Rene mentioned what a Commons means, redefining the role of e-Infrastructure providers, and this challenge was picked up and continued in the second theme speakers. And we want to go even further. And there are e-Infrastructure projects that can already provide good services for some big users like the ESFRI projects. And it was interesting to note from Ed that US needs such good services, while Tony was saying that Europe has to learn from US with regards to the Science DMZs. And both are right, so both regions have to learn from each other.
- Commercially and publicly funded resources and services should be combined and we will see how these will evolve in the coming years. Kees showed how the Dutch ecosystem has picked up a number of the challenges presented, being worked out under the SURF cooperative. One of them is the "SURF Market" marketplace.
- I would like to thank all speakers for their presentations and inputs!