



e-IRG Workshop Summary

15-16th of November 2016, Bratislava



For the presentations and copies of the recordings of the talks from this workshop, please visit: <http://e-irg.eu/workshop-2016-11-programme> and <https://vimeo.com/album/4322392>

Executive Summary

The workshop had three main topics.

- The first topic was about data. During the related sessions, members of the Group of European Data Experts gave presentations on relevant data related topics (Session called: Data aspects - View of the Group of European Data Experts - GEDE).
- The second topic dealt with collaborative services for e-Infrastructure Commons, which is an additional step in the realization of the e-Infrastructure Commons (Session called: Collaborative Services for e-Infrastructure Commons).
- The third topic was the input to the e-IRG Roadmap, where a breakout session gave the participants the opportunity to provide their input. Furthermore, there was a session with presentations from various stakeholders, providing input to the e-IRG Roadmap from their perspective (Session called: Input to e-IRG Roadmap).

A plenary talk, by the Secretary General of RDA, preceded these sessions and focused on RDAs position on how RDA sees, treats, and hopes for the future of e-Infrastructures.

The Workshop Proceedings - Summary

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

DAY 1 – 15th of November 2016

13:00 - 13:15 Opening/Welcome addresses: Sverker Holmgren/Ladislav Hluchy

Sverker Holmgren (SH) and Ladislav Hluchy (LH) gave an overview of the workshop session (SH), and technical and logistical details (LH). SH also presented the idea of e-Infrastructure Commons and its current state of implementation.

13:15 - 14:00 Keynote: Mark Parsons [[pdf](#)]

Mark Parsons, being the Secretary General of RDA, gave RDAs reflections on (e-)Infrastructure and RDA. Modern research infrastructure is, or at least requires, e-Infrastructure. It is about data, and usually large amounts of it. RDA scope is to build the social and technical bridges that enable open sharing of data. RDA has members from 115 countries and grows at a very fast rate.



Half of its members are Europeans and 1/3 are from North America. RDA has 46 organizational members and 6 RDA affiliate members. It helps Accelerate Data Sharing and Interoperability Across Cultures, Communities, Scales, and Technologies. It helps in issues regarding the technical parts of the data engine, rules, practices and basic principles, and creates new and better educated (on data aspects) scientists. RDA's firm belief is that open problem solving is key to have good ideas and solutions from the community. Besides Openness, other RDA principles are Consensus, Balance, Harmonization, Community Driven, and Non-profit. RDA recommendations and Outputs were summarized. Indicative recommendations regarding formal technical specifications, issues dealing with governance, training and data publishing were given. Some common themes emerged amidst the difference (Persistent Identifiers - PIDs, certification of trust, conversations, relationships and mediation). Mediators are needed due to the ever increasing complexity. RDA position on trust and sharing and infrastructure: When, or do we even need to certify trust? We must preserve the freedom to tinker. That is challenging in large scale data facilities/RIs. Build in decentralization where possible. Any centralization must be community governed. Trust is built through shared experience, perspectives, reuse and adoption, and sustained performance.

14:00 Session 1: Data aspects - View of the Group of European Data Experts (GEDE)

- [Peter Wittenburg](#): GEDE Intro and state of PID discussion, [[pdf](#)]
- [Margareta Hellström](#): GEDE from Environmental Sciences, [[pdf](#)]
- [Luca Pezzati](#): GEDE from Heritage Sciences, [[pdf](#)]

Presentation by Peter Wittenburg: GEDE Intro and state of PID discussion, [[pdf](#)]

Peter Wittenburg's presentation consisted of two parts. The first part described the current situation (big Data Management and Access temporal costs) and existing problems (interoperability, innovation hampering, high cost for mapping) that formed the need for the creation of GEDE. GEDE aims to use a bottom-up process to drive discussions and form consensus with and within large RIs on general guidelines, core components and data fabric configuration. It was formed in June 2016, is based on a voluntary basis, and already has a Charter and Procedures. The second part discussed about a European Infrastructure for Persistent Identifiers (PIDs). More specifically, the history of PIDs both in science and in the industry, a clear definition and usage, as well as their potential use as anchors for data (just as with IPs for the internet) was presented.

Presentation by Margareta Hellström: GEDE from Environmental Sciences, [[pdf](#)]

Margareta Hellström's talk on "GEDE from environmental Sciences" focused on GEDE and e-Infrastructures from the point of view of environmental RIs. She presented ENVRI plus (Goals, WPs, products and services) and explained the ENVRIplus Themes. The expectations and interest of environmental RIs on GEDE were presented. Specifically, the ability to map out commonalities and differences, as well as the possibility for environmental RIs to test new advances and gain from them, suggesting in turn solutions (recommendations) to and with RDA, were mentioned. The expectations of environmental RIs with respect to e-Infrastructures were also presented. These consist of expectations for help in many of the technical services, but not in pressure to force the environmental RIs to change and fit generic services, as well as consultancy and involvement on technical design, both in planning of new RIs and in consolidating existing ones. Examples per major e-Infra were given. Overall, dialogue between the RIs and e-Infra's is starting. However, training should not be focused only on the RIs to use existing e-Infra provided solutions, the RIs experience might prove useful for e-Infra as well.





Presentation by Luca Pezzati: GEDE from Heritage Sciences, [[pdf](#)]

Luca Pezzati's presentation on GEDE from the ESFRI project E-RIHS (European Research Infrastructure for Heritage Sciences) point of view begun with an overview of the current landscape in Heritage Science. Heritage science is highly interdisciplinary and requires research infrastructures for interpretation, preservation, documentation and management, while also trying to serve a cross-disciplinary community and establish a common data infrastructure. With this in mind, E-RIHS sees GEDE as an opportunity for Heritage Science (HS) to i) join the discussion on common solutions, ii) to help single out best practices and solutions and to adopt them, iii) to connect to e-infrastructures services, and iv) to give visibility to and to allow the sharing of critical issues in HS data management (e.g. long-term preservation). Also, HS expects to speed up the process towards digital literacy in HS, and to better integrate HS in the “digital ecosystem”.

15:30 Session 2: Collaborative Services for e-Infrastructure Commons (Ladislav Hluchy, chair)

- Joint service catalogue ([Sergio Andreozzi](#), EGI.eu), [[pdf](#)]
- Federated services and data management ([Giovanni Erbacci](#), PRACE, CINECA), [[pdf](#)]
- Providers community outreach activities ([Jan Bot](#), SURF), [[pdf](#)]

Presentation by Sergio Andreozzi: Joint service catalogue, [[pdf](#)]

Sergio Andreozzi started by describing the reason behind the need for a joint service catalogue for research (JSCR), which is now being designed by EGI. Potential benefits of a JSCR are: (i) a simplification in the discoverability of services, (ii) a better understanding of their relevance, (iii) the identification of similar offerings or gaps (for customers & users), (iv) a provision of shared language for service descriptions and an increased visibility of service offerings (for the service providers), (v) an improvement in the communication of what they support, (vi) a support in the evaluation of policy impact (for funding agencies). However, this catalogue is a major challenge due to the differences in service offerings by various e-infra providers. Andreozzi then continued by presenting the work done so far and the approach used, as well as the end result in which point the service model was thoroughly explained. A small set of open questions on scope, governance and management were laid out. Finally, the e-INFRACentral (which includes 5 major e-infras) project's mission and objectives were given, while the schedule to achieve the goals was summarized.

Presentation by Giovanni Erbacci: Federated services and data management, [[pdf](#)]

Giovanni Erbacci made an introduction to the partnership of PRACE (members and infrastructure). He then went on to present the statistics that show how successful a partnership PRACE is, and presented the main methods of access to the infrastructure for scientists.

The project access method (which has cut off dates twice a year) requires 5 months preparation/selection and provides access up to 3 years, while preparatory access gives shorter access times and has cut off dates 4 times a year. He then talked about SHAPE, a programme designed to raise awareness and assist SMEs in taking advantage of HPC. The educational activities of PRACE were also presented (PRACE holds seasonal schools and holds various courses).





Also issues on services and security related to PRACE and its infrastructure were analyzed (Operational security, AAI, WISE). The closing slides presented the current state in PRACE collaboration with other major infrastructures (EUDAT, GÉANT, EGI).

Presentation by Jan Bot: Providers community outreach activities, [\[pdf\]](#)

Jan Bot presented the outreach activities of the Dutch e-infrastructure provider SURF. After a brief presentation of what SURF is and does in each of the 4 major key components (SURFsara, SURFnet, SURFmarket and e-Science Center), he gave an overview of the research service portfolio of SURF. The main challenge for SURF, which is to support the entire Dutch scientific community (and not just a subset of it) was shown with a simple graph. SURF's main idea is to rely and engage with local (i.e. institutional) ICT support and to further tailor the Dutch national e-infrastructure to the needs of scientists. SURF initiated a Support4Research project based on a train the trainers model. The outcome of the Digital Infrastructures for Research - DI4R - session was that it is possible to find a path to EU wide e-infrastructure collaboration. What is needed is a complete service portfolio, carried & advertised by all e-infrastructures, a shared training portfolio and training activities, a shared helpdesk/no-wrong-door policy, and the participation in user community conferences, training provision, e-infrastructure highlighting etc.

17:00 Breakout session - Input to the e-IRG Roadmap

DAY 2 – 16th of November 2016

09:00 Session 3: Input to e-IRG Roadmap (Erik Fledderus, chair)

- GEDE representative ([Carlo-Maria Zwölf](#), Observatoire de Paris) [\[pdf\]](#)
- User community ([Andy Götz](#), ESRF) [\[pdf\]](#)
- User community ([Dimitrios Koureas](#), Natural History Museum London) [\[pdf\]](#)

Presentation by Carlo-Maria Zwölf: GEDE representative, [\[pdf\]](#)

Carlo-Maria Zwölf presented some comments collected from the Group of European Data Experts in RDA (GEDE-RDA) team as input to e-IRG Roadmap. GEDE-RDA has representatives from e-infrastructure and European chairs from RDA groups and is supported by the RDA-EU project. It is based on a bottom-up process and its goal is to form consensus about core components and data fabric configurations. An overall comment from GEDE focuses on e-IRG's existence which is important in order to help discuss issues and policy at the EU level. GEDE, having no mandate to make policy statements about the Roadmap, could be used for collecting individual comments from e-infra practitioners. A basic difficulty for large computing and data centers is that EU involvement is often small, and there are other agendas of higher priority. Service orientation may be a good chance since practitioners will know what real services and contributions are. However, service orientation makes sense in a competitive scenario with open and direct interaction on quality of services, minimal policy involvement, and the possibility to stop a service. Services need to be evaluated and funded by the users who must be willing to pay as a community. This ensures data and services are useful for the given community. Difficulties such as interoperability in data driven science may be solved by specific communities related to the data. Quality, continuity and support are crucial for services, and today there are many factors that limit the trust in continuity. Open forums (preferably organized by the users) are needed to discuss and help mid class services to continue to exist.





As for the Commons, a common layer is recommended in order to have a normalization between technical and policy aspects. A supplementary brokering level should be introduced and focus on few points, be neutral for generating trust, and have clear governance rules for interaction between service officers. The European Commission decided to fund a project for establishing open forum for services. This project should promote an open forum for gathering community comments, and be open to all actors. The traditional split between hardware and virtual layer is wide, although some national data centers have ongoing efforts to shorten it. Summarizing, scientists would like to see prices in services, and have clear packaged solutions, with no further development being required. The Roadmap may highlight different strategies for implementing commons in the case of e-infrastructures, data infrastructures, clouds, etc.

Presentation by Andy Götz: Photon and neutron data analysis as a service, [\[pdf\]](#)

Andy Götz presented some comments collected from the **Photon and Neutron Data (PanData)** community as input to e-IRG Roadmap. PanData is a community with great variety within it. Its size is about 40K active users from all over the world (not just EU). It handles both experimental data (about 3/4 of the times) with typically short processing times but typically large amounts of data (and difficult to move outside the site), and theoretical (about 1/4 of times) with small amount of input and output data but long processing times. A main challenge for PanData is data management and policy on the increasing data volumes, since new large experiments will require several PBs per week. Other challenges include helping (new) users to analyze their data, the development of Data Analysis as a Service, the integration of private and public clouds and several more. A major scientific and infrastructure software problem for PanData is the data bottleneck (data acquisition rate, size, analysis). PanData has started to move towards an open data policy since 2011. The associated costs are paid for by the research infrastructures with own funds. EOSC is proposed to be a federation of services, but it is not clear what services will be useful for the Photon and Neutron source users. The Commons is already happening for the PanData community at almost all aspects, since best practices have already been adopted. Finally, despite the fact e-IRG proposals are good, they are at a very high level and e-infrastructures have little or no interaction so far (except for GÉANT indirectly). It is not sure how scalable e-infrastructures are to cater for PanData community and all the others, it is easier for PanData, according to its own analysis, to get IaaS from commercial companies, while little or no mention exists of PanData community in roadmaps/white papers.

Presentation by Dimitrios Koureas: Virtual Research Environments as integral to e-Infrastructure Commons, [\[pdf\]](#)

Dimitrios Koureas presented an overview of how virtual research environments (VREs) can help science by presenting some successful use cases. On average, scientists in Europe are reluctant to engage with 'open science' practices. Especially in some sciences where technology uptake is relatively low (i.e. Humanities), there is a socio-cultural difference that requires some community engaging mechanisms to be created. VREs can be seen as a solution for the 'last mile' challenge for research e-infrastructures. They facilitate the research data lifecycle and can combine data collection and generation, with curation, publishing and analysis.

In essence, VREs enable researchers to organically change their modus operandi by increasing confidence, agility and marketing visibility of their work. However, VREs are built in isolation, and often develop redundant technological solutions for storage, authentication, and computing. VREs need to be re-invented as integral part of the e-Infrastructure Commons. e-Infrastructure Commons consider community engaging mechanisms as integral to its operation.





The development of EOSC does not reduce the importance of VREs. In fact, it strengthens it. As they become the needed mechanisms that enable communities of practice to engage with core services. The EOSC reduces the operational costs of VRE up-keeping and hardwires interoperability minimizing fragmentation and redundancy. The recommendations for the integration of community specific services to the e-Infrastructure Commons can be summarized as follows: User communities need to be able to articulate and communicate their community-specific needs in regards to data and services, and translate these needs into clear functional requirements that will drive the development of VREs. VRE operators need to look beyond the ephemeral timeframes of project-based approaches, build public-public and public-private partnerships that ensure sustainability, and link with existing underlying e-infrastructure and build on top of available backbone services. Funders need to further acknowledge the pivotal role of VREs in support of user community engagement, and develop, keeping in mind long-term sustainability, dedicated VRE funding programmes with targeted calls to discipline-specific communities.

11:00 Round table (Ladislav Hluchy, chair)

- **Rob Hoofst: GEDE perspectives,**
- **Birgit Schmidt: GEDE perspectives,**
- **Dimitrios Koureas: User community perspective,**
- **Per Öster: Provider perspective (EU and national),**
- **Dorte Olesen: Provider perspective (EU)**

The round table members presented shortly themselves and informed the workshop participants on their past and current positions.

According to **D. Koureas** the community now understands the potential of the e-Infra commons. Yet, it is difficult to articulate within the user communities, on how and why to invest on the commons. It is not that we should not do it, it is simply difficult to make the argument clear. Investment is needed for example in the EOSC, thus, the main point is to articulate the benefits and the value of the EOSC. This has to be in lame's terms.

B. Schmidt added that we need to fill in the gap with all brokers besides ICT, e.g. librarians or support. We must start training these people so that they better engage with the end users (researchers).

R. Hoofst gave ELIXIR's point of view, under which life sciences have growing computing needs. He mentioned that one of ELIXIR activities is already an e-infrastructure. There is also an active discussion with the other e-Infrastructures. There are talks with the BBMRI ERIC to be able to relate to its users' needs. This means that brokers are needed. According to Elixir, all of these layers have a role, not only the bottom layer(s), i.e. EOSC.

P. Öster said that according to his experience, ESFRI has a very broad spectrum of RI communities. Some are very mature and very advanced such as Bio and EMBL, EBI. And other are less mature, e.g. ENVRI. When e-Infrastructures are mature and have good services to offer, it is always the case that the communities are interested. e-Infrastructures must be conscious of the user requirements and be made sustainable, having a business model. From a national perspective, this requires a dialogue. Only then can one use the best technology, come up with strategic solutions and follow the trends in the community.

D. Olesen's stated that she will focus on networking, both from GEANT and the end-user perspective. We must help these two to fit very well together. The network is very mature and reaches all institutes and collaborations, via NRENs, and LANs to campuses. Eduroam is an example of pervasiveness (i.e. AAI) which works extremely well, and leads to huge economies of scale.





Procurements are nowadays taken care of in an exemplary way and they serve well both big users and the long-tail. The short message is that the providers must get their act together and collaborate and create a successful EOSC. The EC also needs to help by creating a better environment, suitable for long term planning.

L. Hluchy summed up until now, and mentioned that in his view, in order to attract the users, we need sustainable infrastructures.

In the process of developing a new research infrastructure **D. Koureas** mentioned that the process of procurement of services is critical. As long as people are unwilling to pay for a service, the service is essentially worthless. The service needs to be maintainable by the provider and useful for the user.

Elaborate and drive their needs, participate in the innovation of e-Infrastructure services and contribute to standards and data caring. Several providers are doing this already, especially the big ones. e-Infrastructure providers must work closely together with the users.

D. Koureas mentioned that progress is good up to know, but we need to continue. Turnkey solutions such as eduroam are needed for the researchers! Research infrastructures are ultimately responsible for providing these turnkey solutions. National governments need to get their act together and engage in discussions with research infrastructures and end users.

12:30 Closing remarks (Sverker Holmgren, e-IRG Chair)

S. Holmgren, summarised the workshop. The main points were that the e-Infrastructure Commons have built the foundations for RIs. However, recommendations should not be focused only towards large research infrastructures or to the providers, but they should be joint ones to help them raise their level of discussion. The purpose of e-Infrastructures is to provide to the users. The situation today is that the e-Infrastructure Commons is the foundation for the EOSC. The chair thanked the audience and participants in the workshop for taking part in these interesting and active discussions.

