e-IRG Workshop 3-4 December 2012, Amsterdam

Around 100 participants attended the e-IRG workshop organized in Amsterdam on 3-4 December 2012 during the Cyprus EU Presidency[1]. Data, in all its forms and contexts was the main focus of the workshop, a topic which is evident in the renewed e-IRG strategy. The workshop was organised with both plenaries and parallel tracks in the areas of trust, foundations, content, services and governance, including also the recent developments on the Research Data Alliance (RDA).

Tuesday, 3.12.2012

Opening, Peter Wittenburg, (Max Planck Society)

Peter Wittenburg, as the chair of the Program Committee, opened the workshop stating that the e-IRG workshop is coming right on time as importing data initiatives are now forming, namely the RDA and explained the 3 parallel tracks (content, services and governance). He welcomed the large participation.

Welcome and Motivation, Gudmund Høst (e-IRG Chair)

Gudmund Høst, chair of e-IRG, in his turn, warmly welcomed the attendees of the Amsterdam workshop. In his opening talk he stressed the need for one-stop-shop e-Infrastructure services for European scientists. For this a "European e-Infrastructure Commons" is needed along with collaborative data infrastructures and services. He stressed that the renewed e-IRG strategy is focusing on these two aspects, namely coordination among the e-Infrastructure components and data. He then mentioned the related e-IRG efforts and initiatives, namely the invitation of the major e-Infrastructure projects for common actions in the area of e-Infrastructure coordination, the e-IRG roadmap that is being finalised and the work of the common e-IRG/ESFRI Data Policy working group.

Gudmund Høst, closed his speech with several quotes who speak for themselves:

> *Financial Times: "Raw data are like sewage, toxic if not handled properly".*
> *Phillips Executive: "Data is like oil, it isn't much worth until you start synthesizing it".*
> *David McCandless: "Data is the new oil? No: Data is the new soil".*
> *TEDGlobal: "In God we trust, all others must bring data".*

Introduction and topics rational, Wouter Los (University of Amsterdam)

[1] The e-IRG delegates from the Netherlands under the umbrella of SURF and in cooperation with the Netherlands eScience Center (NLeSC) and the Netherlands Organisation for Scientific Research (NWO - e-IRGSP3 coordinator) undertook the organisation of the workshop and second meeting during the Cyprus EU presidency

Wouter Los, made a short introduction about the importance of data and in particular of data services and explained the rational for the selection of the topics.

Trust as basis for Data Access and Sharing, Rüdiger Klein (Royal Netherlands Academy of Arts and Sciences / Former Executive Director of the European Federation of Academies of Sciences and Humanities - ALLEA)

Rüdiger Klein started with some metaphors from recent centuries' history stating that building and balancing trust relationships was also one of the main issues back then. In order to establish trust, first some global actors needed to be established followed by second phase local actors. The "delay of information" in the past and how this was used by different actors is another metaphor with the current data world; dealing with the delay is part of the research (and the publications) system. Then moving in time more personal markets emerged and this is also related to the current Intellectual Property Rights system.

Rüdiger Klein highlighted some issues that are not covered well in data strategy documents:

- The language issue; as an example "Riding the Wave" is ok, while the terms tsunami or deluge are not so inviting.
- The necessity for "Codes of Conduct" as rule setting and how these become "Codes of Contact"! A new culture of collaboration needs to be developed based on new values.
- Trust as the proxy for enabling service openings; if trust is such a proxy, then we need to be serious about this new form of collaboration and develop suitable structures within funding, governance and education.

Other main points highlighted in his talk were the following:

- Trust -and making trust a topic of discussion- introduces risks! Trust is also closely linked with sustainability and this group (e-IRG) has an important role to play in both. Research Data Alliance is a great initiative, however, care should be taken not to overlook and side-line other initiatives, and work closely together with them.
- Public Private Partnerships: In the countries that the 3% of GDP for Research is met, most of the funding comes from the private sector. And the issue of trust is there again!
- Involving researchers that produce data will create a better environment for research. It was stressed that the data producers should be involved in discussions about reward systems for the ones who offer their data to others potentially risking their careers.
- Cloud computing: Care should be taken not to betray the trust of those who produce data and not find ourselves in lock-in situations. Open Access and Open Data is key!
- Lessons from private sector governance: Currently emphasis is given to research data from the public sector. However, data generated in the private sector need to be looked into as well, along with the corresponding governance structures.
- UNESCO case: UNESCO and their ICT-related departments have experience in facing with data related issues and such cases need to be analysed so that we can learn from them.

Foundations of Data Organizations, Peter Wittenburg (Max Planck Society)

Peter Wittenburg started with one of the key findings from the High Level Expert Group on Scientific Data declaring the need for a flexible, open, global data infrastructure, that there is no single technological solution, nor a monolithic design to achieve it, and that data trust is a core concept to be taken seriously. He then referred to the Internet analogy and Bob Kahn's statements and stated that data may have a higher degree of complexity, identifying components, interfaces, protocols and objects and properly isolating crucial components. He then referred to the threats for the data future according to the EU and US funding agencies (EC and NSF) and that RDA emerged to act in a practical way and get the data practitioners on-board to deliver concrete results. It was then noted that researchers working with data waste much time for finding and getting access to useful data, checking their correctness and transforming them into different formats, and this is hampering research. A culture of easy sharing in many disciplines is missing, sometimes for proper legal and ethical reasons, or due to lack of awareness, simple technologies or trust. And willingness to share data without any restrictions is still not the dominating trend, while visibility (rewarding) for doing so is not given. The analogy with the Internet was brought up again; identifying data objects as entities to be accessed, managed and shared is crucial and for that a global registration and resolution mechanism for persistent identifiers (PID) is needed, somehow similar to the IP Internet protocol. The importance of PIDs was analysed along with typical access and management workflows. Peter Wittenburg concluded with some basic actions including among others a globally accessible system for PID registration and resolution.

Building Blocks as Basis for a Global Infrastructure - Chris Greer (National Institute of Standardization - NIST)

After reminding the quotes from Gudmund Høst and Peter Wittenburg, Chris Greer brought up similar quotes from the White House Office of Science and Technology Policy and the European Commission on the importance and value of data and big data, and how the opening up of NASA Landsat images contributed to the industrial economy. He then quoted OECD guidelines about access to data and the return for researchers and the general public. He continued by saying that time for action is right: The open exchange of research data has inherent value, 21st century science is global and digital, adequate technologies are available and that a global infrastructure cannot be operated on a voluntary basis. And that a forum for reaching consensus and making decisions and a basis for acting voluntarily on those decisions is missing. He then mentioned the Internet Society and the IETF (Internet Engineering Task Force) as an example of operating a global infrastructure on the basis of cooperation and consensus, and provided more details on its structures and standards. He then moved to introduce the need for a Global Alliance for Research Data Exchange for everyone with a mission to use voluntary cooperation and consensus to enable an open, global research data infrastructure. The characteristics of such an alliance are its community-based type, its openness, its balance of representations, and being consensus-driven, non-profit and harmonisation-oriented. He then introduced the RDA, its steering group and council and its first plenary meeting planned in March 2013.

Data Services for research, Wouter Los (University of Amsterdam)

Wouter Los started also with some main messages from the "Riding the Wave" report, namely that data should be kept accessible, researchers' efficiency needs to be increased and grand challenges need to tackled. He then referred to a set of services that need to be present including core data services (e.g. persistent storage, availability, identifiers, provenance), data upload services (including control, curation and legal interoperability), along with data discovery and access services. The trust that can be built through reporting on data usage among others was then highlighted, while the above-mentioned services were further analysed. It was then quoted that only 4% of the total cost is associated with actually using the dataset, while the biggest percent is about discovery and negotiation to access it, and the remaining about understanding, trusting and manipulating it. He concluded by asking who should be responsible for quality of data, interoperability and standards and misuse.

Assessment of data services: Certification as a means of providing trust, Ingrid Dillo (Data Archiving and Networked Services)

Ingrid Dillo started by analysing the common data and support services layer (as it appears in the "Riding the Wave" report) into a front and a back office, the first usually being a library and the second being a trusted digital repository. She also quoted of studies indicating that cultures of data sharing differ over disciplines and change over time and analysed the reasons for not sharing data, a crucial one being the lack of trust in data produced somewhere else. Trust is thus at the very heart of storing and sharing data. To cope with this important issue the Data Archiving and Networked Services in the Netherlands came up with a Data Seal of Approval with 16 guidelines and transparent procedures. She then summarized related certification standards (DIN 31644, ISO 16363) and the three-tiered European Certification Framework, with basic, extended and formal certification levels. She concluded by stating that trustworthiness of digital repositories is not an illusion anymore (with the above framework), that objective and consistent auditing can be done and that trust by definition implies uncertainty.

Governance Models and Policies, Juan Bicarregui (Science and Technology Facilities Council)

Juan Bicarregui summarized the work of several policy bodies on governance and policies of research data. He started with the OECD "Principles and Guidelines for Access to Research Data from Public Funding" (2004-2006) including 13 principles addressing openness, transparency, quality, interoperability and sustainability. He then referred to the EC efforts and related EC Communications and Recommendations (2007-2012), the most recent of which being the "Recommendation on Access to and Preservation of Scientific Information". The latter included statements on open access publications of data and papers, and that research data become publicly available, usable and reusable, easily identifiable and linked to each other and the related publications. He continued with the G8+05 group and the related Global Research Infrastructure Sub Group on Data. Its 2011 report stated that by 2020/2030 researchers and practitioners from any discipline are able to find, access and process the data they need in a

timely manner, use and understand data, and that data are managed in a way that optimises scientific discovery, innovation, and societal benefit. He then referred to the Research Council UK principles on Data Policy (2011) including that data are a public good and that data should be manageable, discoverable and protected. He concluded with the UK Royal Society (2012) and its work on Science as an open exercise affirming that publication of scientific theories and data permits to identify errors and to support, reject or refine theories; and that science's powerful capacity for self-correction comes from this openness to scrutiny and challenge.

RDA Governance Model and Policies, Leif Laaksonen (CSC - IT Center for Science Ltd.)

Leif Laaksonen started by saying that community coordination is needed to accelerate data – driven discovery and innovation and repeated the "Just do it" principle that helps communities drive tangible progress. He then summarised briefly the history of the Research Data Alliance stemming from the Data Access and Interoperability Task Force (DAITF) that was initiated by EUDAT and OpenAIRE and the Data Web Forum (DWF) in the US. RDA is developing the building blocks and data bridges which require rough consensus and enable exchange of data. He then specified the types of bridges built (e.g. across disciplines, communities and regions) and that RDA is connecting data and people. RDA is accelerating data sharing and exchange across the globe using as main vehicle Working Groups producing standards, policies, best practices, etc. He then specified the RDA Vision (researchers sharing and using data without barriers) and Purpose (to accelerate data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability achieved through the adoption of infrastructure, policy, practices and standards) and the guiding principles of RDA inspired by the several groups as mentioned in Juan Bicarregui's talk. Leif Laaksonen then went into more details on the Working Groups profiles and "case statements", together with the broader ways of participation through the forum and the plenary meetings. He concluded with the RDA Organisational Structure and the current list of candidate working groups.

General Discussion

In the general discussion that followed it was first asked whether the RDA followed the same governance principles as of IETF. Leif Laaksonen answered with a yes and no, clarifying that it is not strictly following IETF principles, rather some ideas like the bottom up approach of IETF. It was felt that bottom-up needs to be combined with top-down approaches, so that funding is also guaranteed; however others felt that money can also flow through users. It was then stressed that the appropriate culture needs to be developed for RDA to succeed and that e-IRG work needs to be linked with RDA. Gudmund Høst enquired about the active role of the users in RDA and Leif Laaksonen replied that although the governance structure is preliminary, the role of the users is evident and it was complemented that the RDA WGs depend strongly on the users. Chris Greer added that the RDA council will be driven by consensus rather than making decisions. The community will have enough experience to judge when consensus can be achieved and then the Council will take up. So the users are indeed a fundamental building block. It was then mentioned that efforts should be made to get the southern hemisphere

involved, including among others Brazil and South Africa, as well as other parts such as China and India. Relations with other fora like the Global Science Forum or CODATA were also inquired and it was stated that the RDA will invite them and work with them closely. It was mentioned that the goals of RDA will complement the other fora and that is why RDA needs to work on such a consensus level and approach as explained before. We just start with a few funders and in the long run we want to be inclusive. Kees Neggers stated that government funding is not the only way; as an example OSI got all the money, but the winner was IETF. Chris Greer replied saying that government support is more to launch some initiatives that should become self-sustained and community based.

The workshop programme was then split in three parallel tracks, namely **Content, Governance and Services**.

Wednesday, 4.12.2012
The 3 sessions were summarised in the next morning.

Content Working Group - Chairs: Chris Greer and Peter Wittenburg

Contributors to this session were Sayeed Choudhury (Associate Dean for Research Data Management at Johns Hopkins University), Paolo Manghi (CNR-ISTI), Massimo Craglia (Senior Scientist at Joint Research Centre of the European Commission), Daan Broeder (Max Planck Institute for Psycholinguistics) and Wouter Addink (ETI Bioinformatics, University of Amsterdam).

Workforce building and bridges: The session started with a discussion about the users and their relationship to data and data infrastructures. A main question was, how users can contribute to the strategy for RDA; it was stated that a candidate-working group of RDA will discuss the user engagement. The need for training and education of users was addressed. But also the work force, the scientists and technicians for data infrastructures need to be developed, trained, and educated.

Research data lifecycle changes: It was then remarked that the research data lifecycle has been changing in the recent years. Accessing data (e.g. remotely) and working laboriously on metadata were two examples. New methods and better tools are required for efficient data management.

What is data? And see beyond that… It was observed that working with data helps to better understand the user needs and this ultimately helps with the sharing of the data, especially if you go beyond disciplines.

The very special relationship between libraries and the scientists (e.g. traditionally from the humanities sciences) was underlined, due to the fact that the libraries preserve the primary research data of the scientists. It was mentioned that this also applies to data infrastructures in general and it will lead to a similar relationship between the scientists and data infrastructure

providers. A new role may have to be defined for libraries handling special collections and advanced services.

Fundamentals: More of what is data and metadata. Is there a common triple? And what is a PID? Peter Wittenburg proposed to define a data object as a triple consisting of a bit sequence (the data itself), a persistent identifier (PID) and metadata. Although the discussion if PID is a kind of metadata and which metadata must be included arises, the majority of the participant in the session agreed on this definition. It was mentioned that already the imperative of a PID would be a tremendous success. And it was agreed that a PID should have global presence.

Enhanced metadata: The need for accurate and rich metadata, which was created during or nearly after the data creation process, was identified. Today metadata are mostly domain specific; to enable multidisciplinary use of these data it is essential to bridge the semantic gap in the metadata. The quality of metadata and the ability to improve this metadata quality were discussed intensively. A feedback mechanism within the metadata and a rewarding system for metadata was seen as appropriate. Furthermore the need for data catalogues, disciplinary and inter-disciplinary ones, which provide access to metadata was stated. Chris Greer further stimulated the discussion on metadata by differentiating between contextual (understanding the context) and relational (relationship-indexing) metadata and the question for a minimum level of metadata was posed. Although this discussion had no result, to raise awareness to create metadata was seen as an essential waypoint.

Interoperability to enable interdisciplinary science: Another topic, which came up in the discussion was the inter- and multi-disciplinary use of data and how to address the various obstacles on all layers that impede the discovery, usage and interpretation of data from one discipline by another. Especially the interoperability on the semantic level was identified as one of the main obstacles. This concerns the data but especially the metadata, which describes the data; how they are automatically processed, structured, interpreted etc.

Standards: Finally, one topic, which spanned horizontally the whole discussion, was the urgent need for standards to enable communication and interoperability in the data area.

Governance Working Group – Chair: Leif Laaksonen

Fulvio Marelli from ESA talked about harmonizing digital preservation policies for Earth Science data. He stated that data is more valuable when combined together and that preservation of data is useless without preservation of the knowledge associated with it. It was then mentioned that accessibility to archived data needs to be ensured, enhanced and facilitated (allowing to combine data from different sources and to perform more complex analyses) and that coherency of approaches among different Earth Science providers need to be ensured.

Kees Neggers from SURF talked about the way to govern an ecosystem. The Internet history shows that no grand design, nor central management is needed; rather an evolutionary model, and that innovation is driven by the advanced requirements of the science community. The

lessons learned include that a shared control plane is required (and not a centralistic model) creating loose cooperation between domains. Basic principles are to keep it simple, that the architecture is based on openness and diversity, multiple domains are connected via open standards, and that bottom-up development together with users (with opposition from incumbents) is required. e-Infrastructure innovation will be driven through competition, cooperation and flexibility; it needs openness, neutrality and diversity as guiding principles and must take account of the global context. Three core functions need to be distinguished: (i) community building, high-level strategy and coordination; (ii) (competitive) service provisioning; and (iii) innovation. Finally, cooperation remains essential for the new internet and e-Infrastructures.

Francoise Genova spoke about lessons learnt from building the International Virtual Observatory Alliance (IVOA) that focuses on development of standards and encourages their implementation. She mentioned that IVOA goals are similar to RDA but for a single discipline. The membership and structure were explained along with the formal procedure for the acceptance of recommendations. The policy Executive Committee, the Technical Coordination Group and the stakeholders/participants were explained complementing nicely all together.

Jamie Shiers from CERN talked about the harmonisation of policies for High Energy Physics (HEP) data. He elaborated on the data management activities required for HEP data such as storage, access and preservation. Data preservation for long-term reuse is an important use case with clear links to other dimensions of the data domain. There is strong motivation to address both technical and non-technical issues in an international / multi-disciplinary environment. And that an effort to profit from this motivation is required.

In the discussion that followed the following topics were addressed:

Usability of the „Internet" model for the data community: It was agreed that we cannot just copy what was done before because the world today is much more complex than how it was when internet started. It will not be possible in only a few years and the challenge is to make it faster than was done for internet and with more different people.

Lessons learnt from the creation of internet: It was clear that Internet was not invented. People worked in parallel and together based on sound principles. Top-down approach for investments was not always efficient.

Governance: Only the main features that have to be in place to create governance need to be worked out at this point, and not all the details. Major impact is expected from the working group level. This will create links with governance in different places.

Incentives for coordination of all the different data worlds: Progress should be made one step at the time and there are things than one community knows better than others. The result from connecting communities horizontally is not clear.

International aspects: e-IRG can become more international and play an important role as it wants to be a coordination platform for discussions among e-Infrastructure components with special attention on data area. The main participants are now in the Northern hemisphere and Australia. The time is right now for strategic initiatives in the Southern hemisphere (some have started).

Impact and how to measure it: It was clear that small bridges are essential. If they are in place there will be traffic, this traffic will grow and the bridges will be enlarged. The working groups will be enablers.

How to bring the knowledge of other communities into RDA: It was deemed useless at this point in time to have interoperability groups with communities that do this interoperability already. Liaisons with these organised communities need to be created. They could be interested in participating in technical groups, while partnerships should be based on added value. The purpose of RDA is to create the connections between communities that create such added value. Audits should be exercised to see if best practices are used. Finally RDA could help to adopt something that is common to everybody which will be beneficial.

Recognition issues and motivation: It was clear that publications contribute to the prestige of scientists and a similar mechanism should be developed in the data area. Metrics, evaluation and peer reviewing of data are important and related goals need to be defined. Credits can be used to recognize data contributions in research, acknowledging the data providers and creating specific data journals. Still, it will take a long time before the academic community accepts recognition of data contributions.

Role of private research and their meaning for governance structures: Three cases have to be distinguished: public, private and co-funded research. RDA will have at some stage value that will interest companies but this has not been discussed in detail yet.

Standards in data community: The data community leadership has to accept that they will make mistakes and that there will be a say five year period before there are results.

The first steps to be taken and who should be doing what: It was recognised that some communities are advanced in creating a global network where exchange of data will be easier. They could have impact and it is a small step on the way. Then RDA should work bottom-up except for things like looking at gaps and overlaps. At some stage there has to be some kind of planning that the output of the working groups have some common denominators (and are not diverging). Perhaps a task for the advisory committee. Brilliant ideas and solutions are generated by individuals, so it needs to be ensured that the RDA WGs construction make it possible that such ideas are developed. Finally, best practices in data need to be recorded.

Services Working Group, Chair: Wouter Los (University of Amsterdam)

The following contributions and interventions were made: Norbert Meyer from the Poznan Supercomputing and Networking Center reported on views on data services: EUDAT and the e-IRG Blue Paper. Andrew Treloar from the Australian National Data Service spoke about providing data services to researchers outside the core community that produced the data. Rainer Stotzka and Marcus Hardt from the Karlsruhe Institute of Technology – Institute for data processing and electronics talked about facilitating data upload, access and deployment. David Giaretta from the Alliance for Permanent Access talked about data preservation and a proposal for an RDA working group. Finally Peter Baumann from the Jacobs University of Bremen and Rasdaman GmbH talked about multi-dimensional big data: from data to service stewardship.

PID and what is persistent: In this Working Group a big part of the discussion was devoted to the Persistent Identifiers (PIDs). There was an effort to delineate the word persistent (e.g. next 5 or 50 years) and whether a long term commitment would be sustainable. It was also debated whether the PID refers to the object, the actual dataset or the service.

User requirements: Then the user requirements in term of data services were presented from the e-IRG Blue Paper on Data Management and a long list was revealed including among others restricted access, federated AAI and accounting, metadata structuring, provenance, integration, interoperation, preservation, search, etc.

Requirements grouping: At the end the requirements were grouped in four key data functionalities, namely:

1. *Find* (and one should not try to change a researcher's searching behaviour)
2. *Assess* (for which procedures for trusted repositories such as the Data Seal of Approval might come in handy; but is certification scalable if 'everybody' becomes a data producer?)
3. *Access* (least restrictive access should be the target but that cannot not always be open)
4. *Reuse* (which can be especially challenging if data are reused in other communities than the original one)

Actors for data services and a user-centric perspective: It was then argued that data producers (researchers) are not good at providing data services and that separate dedicated organisations are needed for that, i.e. databanks. A researcher still wants to know everything about the data, so it was deemed sensible to let him query freely. However, in this case not all data will be included in the databank. The session concluded with the declaration that a technology-dominated perspective should be avoided and a user-centric one should be always the case.

e-IRG Blue Paper on Data, Norbert Meyer (PSNC)
Norbert Meyer summarised the e-IRG Blue Paper sections that included the requirements from different communities (including several ESFRI cluster projects such as CRISP, DASISH, ENVRI and others like DC-NET, ITER, and PanData), a general data e-Infrastructure section and several

specific topics, namely reliability and replications, metadata, unified access and interoperability, security. The identified stakeholders around data were: End-users, owners (producers), infrastructure providers, service providers, policy makers, as well as computer science researchers (on data and database management).

Norbert Meyer then focused on key recommendations from the e-IRG Blue Paper namely:

Data e-Infrastructure; identifying business cases and requirements from ESFRI projects and defining generalised cross-border requirements; define specialised roles for the different stakeholders and ask the service providers to deliver a sustainability policy;

Metadata: establish metadata service managers and give them a greater role in supporting users and especially newcomers; enable easy and standardised metadata and establish federated data catalogues.

Security: evaluate the use of a federated authentication process for the community and check the cost of the transition phase; implement data encryption as a way to increase data protection, confidentiality and integrity in transit and at rest; influence the EU data protection directive which is under revision.

Blue Paper follow-up by e-IRG-ESFRI: Norbert Meyer then explained that the e-IRG Blue Paper work has been taken up by a joint ESFRI-e-IRG working group that looks into the policy implications of the document, specifically what would be requested from the Member States to accommodate the recommendations, what changes are needed at European level, what recommendations to Research Infrastructures in Horizon 2020 can be given and what aspects from the report should be taken into account in future assessment of Research Infrastructures.

Open issues: Norbert Meyer concluded with the list of open issues from the e-IRG Blue Paper that can act as items for further investigation and future work, namely:

- *Business model*: who should pay for what in order for data to be preserved and managed properly
- *Standards*: what is the motivation to use them, who enforces them and who pays for them
- *Common interfaces*: who should define them, especially as the market has interest in doing otherwise, and who pays for them
- *Trust network*: who are the trusted providers of authentication and authorisation information, can it be decentralised, should there be some kind of “passport” office?

Discussion that followed: In the discussion after the talk, the first remark was made by Steven Newhouse on the lack of participation of projects in the e-IRG-ESFRI follow-up work. Norbert Meyer clarified that the e-IRG-ESFRI follow-up work was more about the processes of implementation (not implementation per se) and the policy implications. Gudmund Høst further clarified that the ESFRI forum are mainly policy makers and the intention of this joint e-IRG-

ESFRI group is to dig out what this would mean at the policy level. He also said that e-IRG is more of an expert group and will be working with ESFRI policy makers to increase communications and understand policy implications.

Peter Wittenburg raised the issue of missing the word *simple* (i.e. keep it simple) in the security section of the Blue Paper; Norbert Meyer replied that to make it simple it is much complicated!

Others comments raised were about the need for specialisations in the different roles to be able to cope with the complex data environment, the need for "e-Science integrators" and other data-specific profiles (also for professors) that will advocate and take care for important data management tasks and finally the push towards standards.

Then the issue of competition vs. collaboration, e.g. between the RDA Working Groups, was discussed and the experience from the Internet. Kees Neggers stated that sometimes there is a quick consensus and only one Working Group, while in other areas there is competition and multiple Working Groups. An example for the latter was the area of IPv6. In such cases, the IETF Architecture Boards can also intervene. It was stated that in all cases users should be allowed to choose and that it is the plenary that creates the consensus and takes decisions. On the other hand, it was noted that a rough consensus test is what people will use. So there may be parallel activities to standards that might be adopted in practice and standards not used.

Patrick Aerts reminded that the word 'data' (plural of datum) stems from the Latin word 'dare' meaning to give. So data are given!

Data is the currency of modern science, Thierry van der Pyl (Director DG CONNECT/C, EC)

Thierry van der Pyl started by stating that he will express viewpoints from the policy maker perspective; such a language is needed to convince politicians to allocate budgets. He noted that at present we are in the middle of discussions for Horizon 2020 for Research and Innovation. He referred to the "urbanisation of the digital world" and the significant data challenges. The fast circulation of knowledge enables the creation of new processes such as crowdsourcing and more recently crowd-funding! Citizens are tax payers and it is important to have their saying in how public money is spent. And research is not an exception. It is thus important to show the benefits for citizens.

Thierry van der Pyl continued stating that Horizon 2020 is currently being elaborated and there is a strong focus on societal challenges; there will be three main pillars interacting: science, industrial leadership, and societal challenges. But the latter is crucial. A second main element is the second package on Open Access. The purpose there is to make clear that research funded by public money will have as an effect that publications should be open access; and this will be an obligation. Open access has three main angles: for benefit of science (but also as an incentive to publish), for innovation (benefit from data) and for citizens to make science more transparent and participatory. There will also be a pilot on open access to data. On this area, a balance should be kept between research and innovation, aiming not to prevent any exploitation of

results. Once again the three ingredients -value for science, citizen and innovation- are crucial, especially for politicians. So all stakeholders need to be brought together to discuss around these. He brought up that smart cities and health are two examples in which everything is about data.

Thierry van der Pyl mentioned that data e-Infrastructure is the “new kid in the block”. And that this area is not short of challenges! Some of the challenges are the data deluge, their preservation and quality. The picture that we have in the EU is very fragmented; fragmented by country, by domains, by institutions. On the other hand this fragmented environment also creates and opportunity for an open data infrastructure. And the EU member states are important to accomplish this open infrastructure. Yet, this is not enough, as science is global. EU should however be part and if possible lead global discussions.

He continued mentioning that what we propose is a framework for action based on commonalities. We cannot afford to go discipline by discipline and there is need to go beyond the selfish disciplinary path, striving for cross-disciplinary actions and interoperability. What we need is a combination of expertise, involving the communities, but also expertise of the ICT communities, especially on the limits of technology. The culture of scientists need to change also, working more on incentives and not because of law. So good reward systems are needed.

Other important areas are that of trust and security, including authorization. And likewise the development of skills and specialists. Despite the high unemployment rates, there is still a lack of skills. So we need to think carefully about that and deliver the right professionals for the future. Further issues include soft vs. hard standards, bottom up vs. top down approaches and public vs. private funding. All these are important for funding and sustainability, highlighting also the role of private money.

Thierry van der Pyl continued with the Research Data Alliance (RDA) where EU is currently working close together with US and Australia. He stated that EC believes that the RDA will be very helpful and bring those who know best together, in a model inspired by IETF. Proper balance between top-down and bottom up is being sought. The scientific communities organize themselves around that, while iCORDI is very instrumental and we are very thankful. The same holds for US and Australian sides. RDA is not a club rather an open initiative support interoperability. And it should not be only a talk show, if we want to see tangible results. He gave his wishes on make it happen and deliver results.

In the last part of his talk, Thierry van der Pyl referred to the Horizon 2020 programme, about inventing the future and that research and innovation is in the centre of it. The Horizon 2020 and structural funds are key and they will be leading to growth and jobs. And that everything will be seen with that in mind, i.e. not science for science, but rather lead to growth and jobs. He closed by stating that e-Infrastructures need to enforce working beyond silos and support academia, researchers and citizens.

Discussion: In the discussion that followed Steven Newhouse commented on the aforementioned innovation aspect and that one of the challenges to embrace commercial cloud providers is the usage policy on research traffic. Thierry van der Pyl stated that both academic and industrial research are acceptable, however this is the limit. It should be made clear that competition cannot be distorted and if this is respected then the path will be found. Steven Newhouse reasoned that there are many structures and rules that prohibit integration of commercial providers. Thierry van der Pyl agreed that there are strong rules not distorting EU but also competition that are being followed by DG Competition.

Peter Wittenburg commented on the required change of culture in research and the corresponding reward system and that the carrots and the sticks need to work in parallel, asking about the EC actions on the latter (stick). Thierry van der Pyl stated that the EC is good in directives and that the stick is the obligation of open access publications. For data publication a pilot is planned and its definition is not straightforward. On the other hand, he stressed that they will work closely with research organizations for a reward system on the data publication.

Panel: How to move ahead, Wouter Los, Chris Greer, Leif Laaksonen, Thierry van der Pyl and Andrew Treloar

The panellists started by presenting their views on how to move ahead. Andrew Treloar from Australia stated that we need to elevate the state of data, making data a first class object and not only using primary data for generating the graphs and the tables for publications. Primary data should become first class object for reward, open access, etc. Wouter LoS referred to previous statements and conclusions, namely that data is the infrastructure along with all the related processes and that the Blue Paper calls for a common approach. Leif Laaksonen said that we are in the beginning and that a more coherent and full picture is still missing. So that is why more people teaming up and a more coherent picture is needed on the future. Thierry van der Pyl agreed that we all need to work together, doing concrete things, not necessarily reaching the perfect spot. And that complementarity is important. Chris Greer stated that it is a journey, not a single step. We have to have a discussion about public and private infrastructures and made an analogy with highways and bridges, and that if they are not used, then the value is not realised. He also stated that the ones who run the infrastructures will implement the RDA recommendations. And that's how we can achieve a sustainable e-Infrastructure, i.e. making value and involving the private sector.

Thierry van der Pyl stated that from the EU perspective it is clear that all activities are creating value aiming at making EU more competitive. So once again it is about growth and jobs and this is the vision that we have included in all our activities. In the European Research Area (ERA) there is one important element being that academics can move. However, there is no such "fluidity" in EU between research and innovation, i.e. to help the academics to become entrepreneurs. This does not work so well in Europe. We should not always expect money from public sources and business model for innovation are needed and this has to be taken into account.

It was stated that both EU and US claim that the private sector should be involved and asked for comments and suggestions. Neil Geddes reacted saying that the first recommendation in the e-IRG Blue Paper is on business models and that requirements are not enough. Progress should be made towards business models. A follow-up question was whether there is any commercial involvement in RDA. It was answered that there are some representatives from industry who want to check if and how it will work. If they see value they will join.

On the lack of innovation in Europe it was remarked that there is a risk involved in pursuing innovation and we need to understand how USA make people take such risks. One possible explanation is that there are not so many people trained with this in mind and there is a hindering effect. So people can't see what's coming. Wouter Los stated that "data is the goat" referred to a biodiversity example and a related business case; they go and submit their data in a provider under the condition that they can use some data (e.g. pictures) and this is similar as depositing money to the bank (that can be re-used and invested and also the one who deposits gets an interest).

Thierry van der Pyl stated that there are many small companies exploiting the data in an innovative way. Data are creating opportunities and we need more (of such) entrepreneurship. He said that the EC has issued a cloud strategy but there is almost nothing on research really, rather about the boundary conditions on moving data. And SMEs and other EU companies are exploiting the cloud.

Andrew Treloar said that there is a great potential for collaboration and that we all need to work together to make a bright data future involving library, IT people, and research offices.

Then the discussion turned towards RDA and whether it provided added-value and it was felt that it was the right structure. RDA could also be expanded to respond more to societal needs. Leif Laaksonen stated that not only the research community but also a broader community will be using the data. Societal challenges can only be solved by collaborations. He added that e-IRG makes recommendations and the RDA WGs implement them.

Chris Greer stated that every bit of data increases the value of the global infrastructure. He added that the technology changes rapidly: 60% of devices nowadays use a specific operating system which is only 4 years old! He observed that lightweight organisations work much better. This is the direction that RDA needs to take.

It was also mentioned that the translational aspect is just as important as the provision of additional services and solutions. Thierry van der Pyl noted that we like to think that the Internet is only bottom up, but there are also top down constituents e.g. neutrality. A common vision is needed and openness is good from the EU perspective. Establishing such an open data infrastructure will have an effect on easing multi-disciplinarity and also change the way research is done.

The volatility of data formats was raised and that we need to be cautious of implementing standards; data producers and consumers need to be trained on data practices.

Peter Boumann made some observations, namely that core database researchers should be involved (as they have earlier experience that can be used now), that establishment of companies needs to be supported (including the necessary matchmaking) and that collaboration with industry is not what researchers want. Regarding industry privacy issues were also quoted and that we need manage down expectations.

In their closing points Chris Greer remarked that it makes sense to work on a common infrastructure with EU and others and optimal alignment of investments is required, while Thierry van der Pyl remarked that we need to work beyond silos, that we cannot afford to do business as usual and that we should align investments in a sensible way.

Closing, e-IRG chair and local host

Gudmund Høst thanked all the groups involved in the organisation and the participants that made the workshop a success. He thanked the local organisers for the unforgettable moving dinner and then quoted some of most important liners raised like what would be the first concrete step, what role, who are we and to make it simple it so complicated! He reminded that the aim is to make every workshop better than the previous, and this workshop had no advertisement of initiatives or projects, besides RDA and iCORDI. He invited everybody to the next e-IRG workshop in Dublin on the 22nd and 23rd of May, where discussions will continue. He thanked once again everybody.

Patrick Aerts, on behalf of the local organisers thanked SURF, SARA, the e-IRGSP3 secretariat and Rossend Llurba in particular, EGI and himself (Netherlands e-Science Center)! He wished that a similar workshop is organised during the Dutch presidency and that everybody spread good words about e-IRG and the Netherlands!