e-IRG White Paper 2013

8 July 2013

Final long version

This version is an updated version following the public consultation and it is based on the comments received.

Introduction and Summary

e-IRG sees the integration of services and the interoperability of e-Infrastructures as a crucial aspect to pave the way towards a general-purpose European e-Infrastructure. The e-IRG Roadmap 2012 outlines Europe's need for an "e-Infrastructure Commons" for knowledge, innovation and science to meet the challenges of implementing the EU's 2020 Strategy.

"e-Infrastructure Commons 2020: Integrated Services via interoperable e-Infrastructures" provides an analysis of the situation and provides recommendations on steps to be taken to communalise e-Infrastructures into a public European commons[1]. e-IRG recognizes that these steps are mainly structural and organisational nature and that a strong involvement of user communities is necessary. Integration of services and interoperability of e-Infrastructures are one side of the medal, but coordination of the different stakeholders is perceived as necessity for strategic settings of Europe's e-Infrastructure ecosystem.

The main concern of this e-IRG White Paper is on the integration of service for research communities and the interoperability and coordination of e-Infrastructures. Follow-up aspects are Open Science and the different aspects on data management and the data deluge as well as cloud computing and legal issues that arise from the commercial use of e-Infrastructures.

e-Infrastructures as European Commons is just a step towards the innovation union and the economic exploitation of the innovative capacities of European scientific communities. Opening outcome and processes of scientific research to a broader audience and the commonality is another step to be made to foster innovation and economic growth in Europe. "e-Infrastructure in support of Open Science" depicts current state and future developments for Open Science.

The section "Data policy recommendations for large-scale research infrastructure projects" provides the outcome of a joint e-IRG/ESFRI working group, which was established based on a request by ESFRI to summarize the main policy recommendations from the e-IRG Blue Paper on Data Management. The section aims at stimulating actions timely so that they are in place as early as possible in the development of topical research infrastructures. .

Big Data provides new challenges to all stakeholders of e-Infrastructures beyond the normal issues arising from distributed data storing, preservation etc., which are presented in the e-IRG Blue Paper on Data Management, the high-level expert group report "Riding the wave" and others. In the section "Big Data Access and Standards" two of these challenges are presented and the consequent recommendations are elaborated.

In the section on Cloud Computing the recommendations of the e-IRG Task Force on Cloud Computing are summarized. Furthermore the section extends the full task force report (available on the e-IRG website) on some new insights and developments.

Finally the section "Legal Issues for European e-Infrastructures" depicts six areas where regulations and legislations on different aspects of e-Infrastructures are relevant to commercial use of e-Infrastructures.

[1] Commons: Resources accessible to all members of a community

e-Infrastructure Commons 2020: Integrated services via interoperable e-Infrastructures

Policy Area and Goal

Research and innovation are key elements of the EU's 2020 strategy [2] for European competitiveness in the world.

To meet the challenges of implementing this Strategy, Europe needs a single "e-Infrastructure Commons" for knowledge, science and innovation that is open and accessible continuously adapting to the changing requirements of research and to new technological opportunities. Users, researchers and research communities in Europe, need high quality e-Infrastructure services that are well managed and above all seamlessly integrated from a users' point of view so that they can get on with their business of science instead of spending effort on the various requirements to access those services. The integration of e-Infrastructure services requires the full interoperability of the underlying e-Infrastructures. As a living ecosystem, an e-Infrastructure Commons is flexible and can change dynamically, efficiently and in a future-proof manner.

The key requirements for such a future ecosystem are integration of services and interoperability of e-Infrastructures. They both aim at the removal of existing technical, functional, geographical, institutional and political barriers. The current Internet shows how this works for networking: a common user interface and access mechanism to functionally common services provided by a huge variety of physical networks and inhomogeneous network management domains. The challenge now is to provide this user experience for the full set of needed infrastructure components worldwide.

Such an e-Infrastructure Commons can only be established through a joint and truly common strategic effort between two main actors: (international) user communities -the primary actors- and (international) service provisioning organizations (e-RIs). Next to these two, other actors should be committed: national governments and the European Union, both resulting from their responsibilities for research, innovation and ICT.

Most actors are organized, represented and funded in several different ways. Clearly today there is insufficient cohesion among the different actors to address the challenges ahead. In particular we need more involvement of the user and user communities in shaping the e-infrastructure landscape and its innovation as well as better coordination among the different e-infrastructure pillars. Hence, many barriers towards realising a European e-Infrastructures Commons 2020 are structural and organizational rather than technical. This chapter will further analyse the current situation and makes recommendations on how to pave the way towards the needed e-infrastructure commons 2020.

Context: Current practices - achievements and limitations:

The main e-Infrastructure components and services include networking, high-throughput and high-performance computing, data infrastructures, software/middleware, including authentication and authorisation infrastructures, and virtual research environments that are to be used by international virtual research communities. Today's (and possibly tomorrow's) e-Infrastructures have evolved along different functional, geographic, and type-of-user dimensions, and many of these differences will remain. Also, many stakeholders might find it difficult to navigate in the present landscape of policy-making for e-Infrastructures, as there

[2] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:fin:en:pdf

are several advisory bodies and projects aimed at policy development. Obviously the continuing internationalisation of the user community requires clarification of the roles of local, national, European and global initiatives, especially at the organisational, political and financial levels.

Institutional and geographic borders are increasingly artificial in the "virtual" world of e-Infrastructures. Procedures, and regulations designed for the previous era are not necessarily those best suited for 2020. Organisational models, business models, governance structures, funding models, and regulatory landscapes all face fundamental changes in the face of the development of the vision outlined here: all must be adapted and updated and, where necessary, new models have to be put in place.

Recent policy related publications from several stakeholders (communities) show a growing common understanding of the issues involved and ways forward, notably the GEANT Expert Group report, EGI2020. PRACE Roadmap, EUDAT Roadmap, Riding the wave report as well as recent e-IRG publications, the e-IRG Strategy report and the e-IRG Roadmap 2012.

So the time seems ripe now to turn policy into action. Elaboration of current roadblocks and issues on the road to more effective integration of services will be needed. Chapter 2.1 in the e-IRG Roadmap 2012 provides an initial list:

- insufficient coordination, collaboration, and integration of existing e-Infrastructures services
- legal issues
- hurdles to access and usage of e-Infrastructures
- lack of "Visibility" of e-Infrastructure services, but high awareness by users of borders, interfaces, and technologies of the individual components
- lack of business models based on secure and sustainable funding streams for the use and innovation of e-Infrastructures

This list can be extended with:

- Integration aspects with private users and private suppliers as a consequence of Horizon2020
- the lack of coherence from the user communities, going beyond the well-organized examples, such as CERN & EBI

A better understanding of 'users' of e RI services is needed. What do we mean with users and their usage? Are they individuals or groups? Small groups or large groups? Distributed or centralized? Do we mean flagship user communities such as the ESFRI-projects and the EIROFORUM-labs or the 'long tail' of smaller research collaborations across a few countries? If they have a 'big data' challenge, what do they mean? Many centralized data? Distributed data? Or less 'big' data but very valuable and of high quality? And what exactly is their need to have this big data processed and transported? Answers to these types of questions are needed to develop integrated e-RI services.

Several actions directed towards integration of e-infrastructure services are already on-going, notably actions in the field of data discussed in the Amsterdam e-IRG Workshop or AAI-

activities on European and world level, Persistent Data Identifiers, File movement, Service Monitoring, Service Registry & Discovery, the development of PRACE Competency Centres and similar activities across other e-RIs to help promote uptake of e-RIs and porting to e-RIs. However, clearly more must be done to help service discovery and usage support for end-users.

Proposed approach

All e-infrastructure service providers serve the same community of users. It is clear that today no single e-infrastructure provider provides a full portfolio of e-infra-services needed to the end users. However, it is also clear that users want a single and easy to use interface to all e-Infrastructure services they need. They need services that are coherent, managed and above all integrated so that they can get on with the business of science. But they also need a constant innovation of these services, way ahead of what commercial providers can offer. So we must be careful to not become constrained and stifle innovation in the development, provision and use of these services.

Therefore aiming at a centralised e-infrastructure provisioning is not a solution of the problems. The idea that there will be just "one way" of supplying or using any service through an "efficient" mandated or "voluntary" monopoly has to be avoided. Allowing e-Infrastructures to evolve is important: open competition, collaboration, but also technological bypassing and new distribution concepts of service belong to the 2020 vision as well.

An e-Infrastructure Commons can only be established through a joint and truly common strategic effort between users and primary strategic actors and suppliers. However, striving for a common strategic vision should not, a priori, be taken as a threat or barrier to the continuing innovation and ambition of any of the individual (existing) services. In the 2020 vision, providers have the freedom to innovate and users enjoy the freedom of choice of the services they need and have equal access to commercial services.

So what can be done to achieve this goal of integrated services towards the users if they still will have to be provided by several public and private service providers?

First of all the users need to become much more directly involved in strategy, coordination and innovation activities in each of the e-infrastructure components. Further on users need be prepared and empowered to pay for e-infrastructure services. Only then they will be able to select the best possible providers, including commercial providers. This will not only improve quality and cost for the research users, but also facilitate timely dissemination of innovative services towards a much bigger user base of the society at large.

e-IRG believes this challenge can be met by applying overall organisational principles towards separation of three distinct core functions for each of the e-Infrastructure components, while e-IRG could facilitate a platform for strategy alignment and coordination between the various e-RI pillars to allow for a sound evolution of the e-infrastructure ecosystem. The three core functions to be set up separately are:

1. **Community building, high-level strategy and coordination in Europe:** a single organisation with a central role for user communities, especially involving large, advanced and well-organised user communities at a European level and beyond.

2. **Service provision:** flexible, open, and competitive approach to national, European, and global service provision; with advanced collaboration among the interested public and commercial service providers.

3. **Innovation:** Implementation of major innovation projects through the best consortia including e-Infrastructure suppliers, industry, users and academia with a dedicated management structure comprising the partners per project.

In this approach it is essential that the position of user communities in e-Infrastructure governance will have to be strengthened on four levels:

1. On the **strategic level** user communities will have to organise themselves to drive the long-term strategy.

2. On the **service provision level** user communities will have to learn to use their joint purchasing power in a competitive market.

3. On the **innovation level** advanced users of international e-Infrastructures should support the specification and real-life testing of new e-Infrastructure developments, for their mutual benefit. LHC and eVLBI are already good examples of this approach.

4. On the **standardisation level** user communities should contribute to the process of setting and implementing the international standards necessary to achieve the transition from the current e-Infrastructure service portfolio to the international, service-oriented, e-Infrastructure portfolio envisioned for the e-infrastructure commons 2020.

In addition e-IRG sees a need for a single e-Infrastructure umbrella forum for strategy setting in Europe, with sufficient user participation for community building, high-level strategy and coordination for the entire e-Infrastructure. This umbrella-forum should be clearly separated from operational responsibilities.
In this forum the different strategy and coordination bodies of the various e-infrastructure components could also jointly address common issues like:

- Expanding the user base of e-Infrastructures:
  - how to increase the visibility of e-RI's
  - how to make e-Infrastructures relevant to a wider user base (better & more broadly appealing services)
- Investigating whether researchers and international research projects are harmed by a Digital (e-Infrastructure) Divide
- Eliminating legal and political roadblocks for international harmonisation of exploitation and innovation of e-Infrastructures.
- Promoting the use of sustainable business models for e-Infrastructuress, in support of exploitation and for innovation of international e-Infrastructuress.
- Clearly separate out the (different?) business models around operation, support/maintenance and innovation.
- Promoting effective structures for international governance and finance giving users their proper role across users of all sizes: from large pan-European ESFRIs to smaller international research collaborations

- Developing common and living functional models and standards necessary to achieve sustainable coordination of e-Infrastructures services.

e-IRG is a clear candidate to facilitate this umbrella forum function and has already expressed its willingness to do so.

Recommendations

International user communities requiring e-infrastructure services need to organize themselves to be able to address the challenges in their future roles:

- driving the long term strategy for their e-infrastructure needs ;
- learn to use their purchasing power;
- participate in and drive innovation of e-infrastructure services;
- contribute to standards;

International organizations of e-Infrastructures need to join forces and share their common challenges towards serving the European user communities, thereby avoiding as much as possible any duplication of efforts in such areas as:

- outreach to and involvement of user communities;
- services registry, discovery and provisioning;
- financial, legal, business development and procurement issues

Both e-Infrastructures and user communities should establish a clear separation between responsibilities and tasks for strategy setting and community building on the one side and operations on the other. They should together strive to establish the e-Infrastructure umbrella forum for strategy setting in Europe, with sufficient user participation for community building, high-level strategy and coordination for the entire e-Infrastructure, with –again- a clear separation from operational responsibilities.

National governments need to:

- provide a basic funding level for (the actors in) their national e-infrastructure, in particular devoted to its continuous innovation, and
- empower and fund their national user communities, enabling them to influence the development and use of the national e-infrastructure;
- remove existing national regulatory or political constraints for accessing public funded e-Infrastructures, in particular for private research or public-private research ventures.
- encourage (the actors in) their national e-infrastructure to collaborate and join forces with their counterparts in other countries and at EU level

The EU should strengthen the actions of the national governments by

- establishing the necessary international harmonised framework for the funding of e-Infrastructure innovation;
- empowering and funding European user communities to influence the development and use of transnational access of the e-Infrastructure;

- promoting the use of Structural Funds for e-Infrastructure development in less favoured areas;
- striving towards harmonisation so that regulatory conflicts can be avoided, both at the national and at the international level, e.g. with the existing regulation for state support or competition;
- providing clear guidelines for 'regulation proof' participation of private research both in the use and in the supply of e-Infrastructure services;
- harmonizing European and international regulatory conditions;
- encouraging a sustainable e-infrastructure offering in Europe.

Existing Service Providers will have to face the continuous challenge of service development, funded through public schemes in its early and precompetitive phases, and through private schemes thereafter.

e-Infrastructures in support of Open Science

Policy Area and Goal

> "Research and innovation benefit from scientists, research institutions, businesses and citizens accessing, sharing and using existing scientific knowledge and the possibility to express timely expectations or concerns on such activities."[3]

The Internet has dramatically changed the way in which research is done, aspects range from data sharing and access to research results to international collaboration in virtual research organisations. The umbrella term Open Science describes best the cultural change how science can be done most efficiently and innovatively, how scientists connect with society and collaborate interdisciplinary. In general Open Science comprises Open Access (access to research publications), Open Data (access to research data) and Open Research (collaboration and interaction with society) for all "levels of an inquiring society, amateur or professional"[4]. Open Science (at least in the meaning of broad collaboration on shared infrastructure and shared data) is naturally given in sciences where measuring instruments are extremely large and/or one research group can study only a fraction of the subject under observation is that large that only once in a lifetime dedicated areas can be inspected (e.g. Large Hadron Collider in the area of high-energy physics or the output of different telescopes in the area of astronomy). Thus data and research results need to be openly available to the research communities to drive science forward.

It is widely agreed that results of publicly funded research should be publicly accessible. This concept of immediate, online, free and available access to research outputs is labelled as **Open Access**. It comprises the permissions of reuse and redistribution (as long as the original source is cited). It can be applied to peer-reviewed journal articles, conference papers, etc. Two main strategies of Open Access are the "green road" and the "golden road". Green describes the self-archived reprints of conventional journals in parallel to a publication by a publisher. Golden describes the shift of publication costs from the readers to the authors. In recent years Open Access made huge steps towards becoming the new norm in research publishing in Europe[5]. Since the beginning of FP7, for all projects Open Access publishing costs are eligible and this is also envisaged for Horizon 2020.

While the term Open Access refers to text, **Open Data** is referring to every other digital representation of data (measured data, images, sound, etc.). In 2007 the OECD published its *Principles and Guidelines for Access to Research Data from Public Funding*[6], which addresses recommendations based on commonly agreed principles to facilitate cost-effective

[3] COM(2012) 392 final, A Reinforced European Research Area Partnership for Excellence and Growth (http://ec.europa.eu/euraxess/pdf/research_policies/era-communication_en.pdf)
[4] http://en.wikipedia.org/wiki/Open_Science
[5] http://europa.eu/rapid/press-release_IP-12-790_en.htm
[6] http://www.oecd.org/science/sci-tech/38500813.pdf

access to digital research data from public funding[7]. The European Commission has communicated[8] the need to publish data publicly to address the societal challenges of the 21th century. Furthermore the European Commission expressed the need for sustainable infrastructures to make these data accessible.

Though broader and less exact than terms like "Open Access" and "Open Data", the more general concept of **Open Research** as sharing methods, software[9], and other variants of virtual infrastructure is seen as a huge step forward in making research results more reproducible, gain transparency, and ease collaboration on existing research data. While Open Access and Open Data enable passive access to and participation in research, in Open Research participants contribute to scientific research and can take creatively part in the research process.

The different aspects of Open Science extend the way to implement the "fifth freedom"[10], the free movement of knowledge across Europe besides the free movement of goods, capital, services and people. Open Science is fundamental to bridge the digital divide and foster economic growth.

Context: Current practices - achievements and limitations

In the overall context of Open Science neither benefits nor beneficiaries are fully known and understood by funding agencies and research institutions. There is little experience about the impact of Open Science neither on science and scientific research nor on individual careers, which is accountable for the weak acceptance of Open Science by the research communities. The example of disciplines such as astronomy, which have been widely using Open Data, shows that this changes the way science is done and gives equal opportunity to scientists, wherever they are, to access and use the best data and tools, wherever they are produced and kept.

Open Science is supported by different e-Infrastructures, which provide services to enable Open Access and Open Data. Open Research on e-Infrastructures is limited by the authentication and authorisation policies of the service providers and will require a cultural change. Open access/data should become the rule, providing that legitimates exceptions are granted (e.g. to protect privacy). This cultural change is starting with Open Access and Open Data.

[7] Recommendations of OECD are adopted when member governments are prepared to make a political commitment to implement the principles (and/or guidelines) set out therein and it is expected that the member states seriously work towards attaining the standard or objective within a reasonable time frame.
[8] COM(2012) 401 final "Towards better access to scientific information: Boosting the benefits of public investments in research" (http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf)
[9] http://arxiv.org/abs/1210.0530
[10] http://register.consilium.europa.eu/pdf/en/08/st07/st07652-re01.en08.pdf

Open Access

In the European Union[11] Open Access is provided by a variety of repositories, journals and national funding mandates. Several projects funded by national funding agencies and the European Commission stimulate activities for Open Access development. The vast majority of the Open Access projects and solutions provide the green way as they have built e-Infrastructures, which provide repositories for publishing the outcome of scientific research.

Although there are strong mandates with substantial effects from the European Union and European states for Open Access, some European countries lack national mandates. The Europe-wide coordination and cooperation of OA e-infrastructures is not sufficient to provide users a seamless and integrated experience. Furthermore issues of copyright and intellectual property have to be addressed European-wide.

It is perceived that a fast move towards Open Access would be a massive intervention into the publishing business so that the promotion of the green road facilitates a moderate transition. The EC is supporting the golden road by funding expenses to the authors, but a general movement could be only global since science itself is globalized and consequences on the publishing process, the budget of laboratories and research organisations in different contexts, scientific careers, etc. have to be fully understood.

Open Data

EC-funded projects are implementing e-infrastructures for a long-term preservation of data (e.g. EUDAT) and are aiming at a long-term perspective, which is at least an issue of financial models. Although e-Infrastructures and thematic data infrastructures for storing and providing access to data are now rapidly emerging worldwide, the majority of the research data are stored in local repositories at universities and research institutions, because the research communities are not confident that funding and financing of storage and support structures is long-lived and will provide long-term solutions. Additionally the technical challenges of preserving large volumes of data remain unsolved, in particular in fields where conditions change constantly. Some disciplines are organising themselves to preserve and share the data they produce and can provide examples and lessons learnt.

Many researchers are reluctant to share data because they expect others will benefit from their efforts and there are no guarantee of reciprocity or proper citation of their work. Furthermore researchers do not want to spend time end efforts to learn how to store data in publicly accessible archives. In some areas of science there is a lack of terminology and standards for achieving and sharing data and/or tools for achieving and sharing data are inadequate. There may also be legal issues that need to be addressed prior to sharing data, such an IPR and privacy protection. Last but not least the system of academic appreciation is currently not able to support data sharing.

[11] http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/access-by-region/europe-and-north-america/

The EC identified the lack of organisation and organisational models and clarity about responsibilities in improving access to and use of scientific data as a major barrier to change. Furthermore current financing models are not suitable to ensure long-term access to stored data. The High Level Group on Scientific Data established by EC in 2010 concluded that diversity will be a keyword of the Research Data Infrastructure and promotes the development of a global "Collaborative Research Infrastructure" including generic and disciplinary elements[12].

With the launch of the Research Data Alliance (RDA) an international forum for "facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability" [13] is available. Although with RDA and the iCORDI [14] support project structures for collaboration and standardisation are in place, interoperability among countries and disciplines remains an issue.

Proposed approach

Although the major challenge for Open Science is the change in scientific culture and therefore not objective to the further development of e-Infrastructures, e-Infrastructures are fundamental instruments for Open Science and their policies, regulations and restrictions must be adapted accordingly to stimulate Open Science.

Strong efforts have been made to promote Open Access. Currently data and the implementation of a Common Data Infrastructure (CDI)[15] receive funding and attention. This CDI has to be interoperable with well-established disciplinary data systems when they exist. To push Open Data forward the implementers of the CDI should be encouraged to support and sustain Open Data to become a European Common.

To guarantee the long-term perspectives of Open Access and Open Data, project-based funding must be changed into financial models with adequate business models and legal constructs.

Facilitate access for citizen researchers to e-infrastructures, provide training and education capabilities to these people, and encourage the participation in science and research.

Recommendations

To make Open Science a European Common and the natural way science is done in Europe, a cultural change must take place. e-Infrastructures can provide the technical instruments to conduct science in an open way that science and economy will benefit.

[12] « Riding the wave – How Europe can gain from the rising tide of scientific data » (October 2010) http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
[13] http://rd-alliance.org/
[14] http://www.icordi.eu
[15] http://www.eudat.eu

Recommendations from EC, OECD and UNESCO are in place and provide guidance for EC, national governments and e-Infrastructure provides. The stakeholders are encouraged to implement these recommendations.

International scientific communities

The scientific community, as the information provider, has to put in place mechanisms allowing shared data to be properly described, in particular by developing a special trained body of 'data scientists' with the relevant disciplinary knowledge. These functions are different from the generic expertise of librarians, and have to be recognized.

Data publication and sharing have to be fully recognized as a part of scientific activities, and taken into account in the evaluation of individual scientists, teams and laboratories. Criteria for this evaluation (quality, impact, etc.) have to be established.

Funding agencies and national governments

Agencies running large research facilities are strongly encouraged to provide a data archive and to open it for usage by a wider community than the original consortia, eventually after a proprietary period allowing restricted usage to the teams which have played a key role in producing the data.

EC and funding agencies

EC and national funding agencies should engage to sustain relevant e-Infrastructures to provide long-term perspective to Open Science.

EC and national funding agencies should stimulate the coordination and cooperation of the national Open Access initiatives.

EC and national funding agencies should encourage scientists to publish their data in Open Data repositories.

EC and national governments should stimulate the development and promote the use of systematic reward and recognition mechanisms for data sharing, such as citation mechanisms and measurements of the data citation impact.

EC should establish a directive for all e-Infrastructure providers that receive European funding to ensure publication and open access to data about governance, policy and funding of their e-Infrastructures.

Data Policy Recommendations for Large-Scale Research Projects

Policy Area and Goal

Research infrastructures (RIs), such as the initiatives on the ESFRI roadmap, produce and are dependent on rapidly increasing amounts of data. For research and society to take full benefit of the major investments in RIs the data needs to be made openly and easily available for researchers, over wide spans of fields, in sustainable settings. To enable this, the data needs to be managed, stored and preserved in a cost-efficient way, with appropriate quality and safety assurances. Also, access to the data across borders and domain boundaries must be secured. e-Infrastructures provide the versatile services and tools needed for both data management and access, but the development of such infrastructure must be complemented with an effort on RI policy and coordination to accomplish the goals above in practice. This effort must be driven by the needs of the researchers, but many stakeholders need to be orchestrated to complete the build-up of data infrastructures rapidly while still arriving at sustainable and cost-efficient solutions.

Recommendations

In 2012, e-IRG presented a Blue Paper on this topic with a focus on the needs of large RIs such as the ESFRI Roadmap initiatives. This summary contains the main policy recommendations of the 2012 Blue Paper, extracted to provide a comprehensive list of actions that should be taken to arrive at situation where research and society can reap the full benefits of major RIs.

As a fundament for RIs, sustainable e-Infrastructure services for enabling access to, storing, preserving and curating large amounts of data need to be in place. Policy makers are recommended to take action to ensure that:

- Roles (e.g. end users, data owners, infrastructure providers, service providers, and researchers on data management) are identified and, when appropriate, partitioned between different actors to ensure effective and cost-efficient solutions, fulfilling the needs of the end users and data owners.
- Governance and mandates for different actors are clarified and their way of interacting is sufficiently formalised. Actors' specializing on different tasks ensures that synergies can be exploited, leading to cost-efficient implementation of services. Clear responsibilities and formalised relations ensure that the relevant quality of services can be maintained. Funding paths are defined and sustainability for all parts of the e-Infrastructure is secured.
- Costs for different services and procedures are made transparent and that different options for implementing them are investigated.

Also, to ensure that data will be available across borders and disciplinary domains, RIs and e-Infrastructure providers are recommended to take appropriate steps to:

- Ensure that data formats are standardised and contain sufficient information on the data (metadata) to enable global usage within the discipline, across disciplines, and in new research settings that could possibly not be envisaged at the time of creation of the data.

- Build e-Infrastructure solutions consisting of multiple layers, successively adding more specialised higher-level services using standardised interfaces. Here, different actors can provide different layers.
- Adopt a global, standardised lowest-level data infrastructure layer, including e.g. authorisation and authentication and persistent data identifiers. Here, federative approaches could be used to include existing solutions.
- Define and successively move towards a common second-level data storage layer where cross-related requirements between different RIs are identified and utilised to enhance cost-efficiency and quality. Also here, standardised interfaces and federative approaches should be used to include existing solutions.
- Ensure that quality of the e-Infrastructure services and the data security is delivered at a level, which is relevant for the data at hand.

In this process it is important to take into account already existing and used e-Infrastructures and services, which should be integrated with the potentially new infrastructures using well-defined and standardised interfaces. The overall goal is clear: adopt existing e-Infrastructure solutions whenever available, remove isles, and build standards for new challenges which are at the beginning of the development and can be easier integrated in the multi-layer structure.

For the ESFRI initiatives now under implementation some of the mentioned points are partly followed up in work packages in the cluster projects funded by the European Commission. These projects can be seen as a first step towards more coordinated European data infrastructures for some of the new and hopefully also some of the previous established Pan-European ESFRI RIs. The present cluster projects are mainly covering one or two scientific areas (for instance SSH or medicine). There is a need, however, also to secure synergies between the different scientific areas for instance by having closer cooperation between projects within social sciences and humanities on the one hand and environmental projects and/or the medical sciences. Only by coordinating the e-Infrastructure activities and building a multi-layer data structure can we succeed in establishing an RI system in Europe, which will avoid separated isles of infrastructures for a specific scientific domain. The RIs are recommended to

- Continue the work in the cluster projects and to include new incoming ESFRI initiatives under one of the umbrella of one of them.
- Coordinate horizontal activities between ESFRI cluster projects.

This will allow that requirements are synchronised, and that services developed within the clusters are deployed in the early stage of implementation of ESFRI initiatives.

One way for ESFRI to follow up the ideas from the e-IRG Blue Paper is to give advice both to the Commission as well as national funders concerning the need for coordination of data infrastructure activities. This is necessary both of economic and scientific reasons. Another and perhaps parallel way could be to ask e-IRG to play a more active part in organising or give advice based on the many initiatives now coming from different bodies concerning not

only data but e-Infrastructure activities in common. By a better coordination of the activities within the area it will be easier also for ESFRI to follow up on the policy level.

Big Data Access and Standards

Policy Area and Goal

Big data is widely perceived as one of the most challenging items to be addressed in the coming years. Different communities have a different feeling about big data, depending on various factors, among which the reference framework (including, computing platforms, methods, procedures and tools) plays an important role. A general consensus is that big data is about volume, variety, veracity and velocity (often referred to as the "four V's"[16]) of data, in a sense that these attributes arise in a manner which challenges the constraints of current system capacities, even if massive parallel computing instances are used for data processing. To tackle this problem, specific frameworks have been developed, and the need for standardisation of these methods as well as authorisation and access issues to Big Data is of major importance.

The **volume of data** that needs to be processed for analysis in science as well as in industry and economy increases steadily to obtain results, which are more specific, more detailed, or even more accurate. Improvement of the measuring instruments, lowering of the prices for the instruments and also, importantly, the falling prices of storage lead to increasing amounts of data. But, Big Data does not arise solely through the growth of problem-specific data sets --- increasingly, data which tackles the object under investigation only marginally (most notably, social media data) is included in the analysis, thus further exacerbating the Big Data issue.

The use of different data sources is cause for the **variety of data** used for analysis. The data is not always structured, especially through the use of social media semi-structured or unstructured data which are often included in the data-processing pipeline. This variety often mandates semantic interpretations of the data (e.g., extraction of structured entities and relationships), which naturally implies increasing need for computational power. Furthermore, statistical and machine-learning techniques are often needed in order to deal with the **veracity of data**, that is, the inherent levels of uncertainty and possible noise in such semantic interpretations and the data itself.

The fourth factor in the definition of Big Data is the **velocity of data**, which describes the speed in which data streams in, is newly produced or recorded. As depicted beforehand, the use of improved measuring instruments, the inclusion of a broad variety of different data sources and of social media, and the existing very-high-speed data networks enable the gathering of data in a manner that is best described as the data deluge.

Data and data infrastructures are recognized as crucial factors for the solution of the grand societal challenges. To enable the use of data infrastructures open questions concerning preservation, metadata and others must be answered; Big Data additionally brings specific

[16] http://www.ibm.com/software/data/bigdata/

challenges. In this section of the e-IRG White Paper the access to Big Data and standards for Big Data will be addressed.

Where does "Big Data" originate?

Data originates everywhere around us, for instance, via web services, social networks, applications, measurement instruments, sensors, infrastructures, in principle via anything that generates log information. Digital data can now be collected, replicated, moved, and processed more quickly than ever before, and with constant developments in information technologies, sensors, and communication networks, the volume of scientific data will continue to grow. Big Data emerges in places where this data is being aggregated (data warehouses). Data has become such a significant part of science that data and information sciences have advanced into complex disciplines with numerous areas of specialization. As the scope of Big Data expands beyond the Petabyte and Exabyte levels across engineering, physical, and social sciences, the data deluge will give the scientific community compelling reasons to embrace new scientific methods, paradigms, and attitudes toward data sharing.

The level of noise/uncertainty inherent in the data is not consistent and primarily depends upon the accuracy of data sources generators. In order to use this data stemming from different sources in a consistent and effective manner this level of accuracy should be part of the metadata associated with the data. Without this information results from data obtained via very different data sources (e.g., climate sensors and humanities data) cannot be compared and usage of this data in cross-disciplinary studies becomes impossible to handle. Another aspect to consider before using Big Data for research and industry purposes is related to the normalization factor of the data. Normalization of data has been a common practice in relational databases since the early 1970s when due to lack of physical resources the need to keep the duplication of data and the amount of tables as low as possible was apparent. Within the forming Big Data landscape normalization refers similarly to preserving the data warehouse design with performance and optimization issues in mind.

What do ‘standards’ mean for Big Data and why they are important

Standards for data have become an important item since many years. Although for small or medium size user communities compliance to standards can be of little interest, whenever there is the need for sharing or exchanging data, data-openness becomes one of the key issues to be addressed. This is the reason why there are a lot of important initiatives related to open data or open linked data all over the world. The European Commission has showed a growing interest in open data that resulted in the Open Data Strategy for Europe[17] launched in December 2011. The main goal is to ease the reuse of the public sector information and this is achieved also through the Open Data Hub of the European Union, recently released in beta version[18]. In the research arena, compliance to data standards is considered important by many user communities and this has led to various standardization activities, although most of them are specific to a scientific domain (e.g. medical imaging, etc.).

[17] http://europa.eu/rapid/press-release_IP-11-1524_en.htm?locale=en

[18] http://open-data.europa.eu/open-data

The e-IRG 2012 roadmap foresees, for the evolution of the current e-Infrastructures towards Horizon 2020, a common data infrastructure integrating a set of coherent data services exposed to users by means of an interoperable set of underlying e-Infrastructures. This Collaborative Data Infrastructure, described in the "Riding the Wave" document[19], and depicted in Figure 1, is a key element to allowing user communities to get on with the business of science, since the services they need could be provided by a variety of different actors.

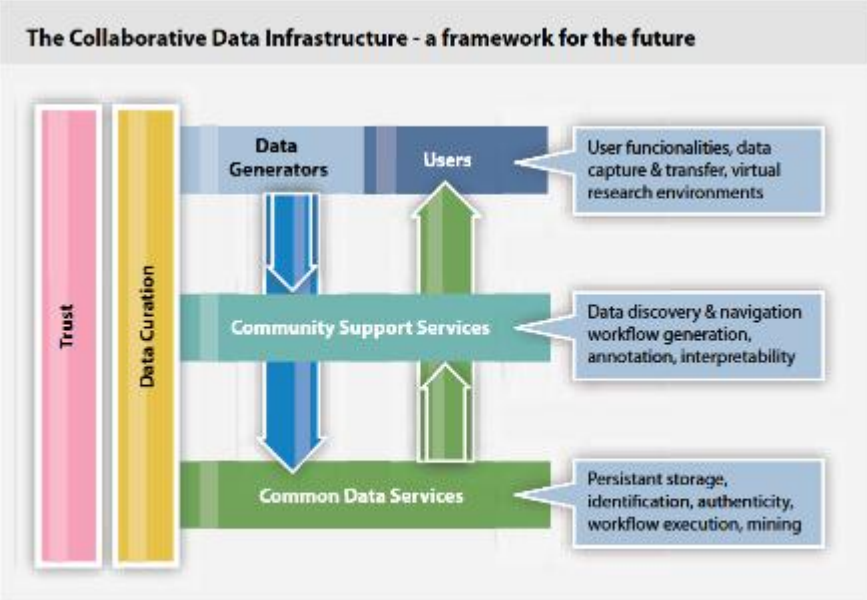

Fig. 1: The Collaborative Data Infrastructure as envisaged in "Riding the Wave"

To reach this goal, standards (either de jure or de facto) are a key element since they are crucial to easing the aggregation of services and even more the e-Infrastructures interoperability. Either the definition of new standards for Big Data or the extension/adoption of existing ones should be driven by the four main features highlighted in the previous sections; however, the variety of data seems to be the feature that will mostly impact these activities. Data originates from a lot of different sources: some data are very well structured, although with a different level of accuracy (e.g., scientific data), while other data (e.g., data produced by social activities and tools or some data currently in use within the public sector) may not be structured at all. So, in order for standardization of data to be achieved, the issue of how to measure different data and their accuracy and how to normalize them should be addressed; this also leads to the need for metadata standardization.

Why and when to use Big Data

Within the industry sector, Big Data analytics can reveal insights such as peer influence among customers (e.g., by analysing shoppers' transactions, social and geographical data). Being able to process data in a reasonable amount of time removes the troublesome need for

[19] Van der Graaf, M. and Waaijers, L. (2011) A surfboard for Riding the wave. Towards a four country action programme on research data. A Knowledge Exchange Report http://www.knowledge-exchange.info/surfboard

sampling and promotes an investigative approach that produces near real-time useful and insightful results that are of major importance within decision-making processes. Furthermore, predictive analytics based upon analysis of Big Data can be utilized and projections are also made possible. With respect to research objectives, new interdisciplinary fields will emerge among distinct fields of science as correlations and coherences within Big Data collections will be sought for by different actors and stakeholders.
Organizations (e.g., CODATA[20]), institutes, and individuals will strive to link together various national and international data seeders through their common interest in and reliance on scientific and technical data. It is a vital part of an agenda aimed at achieving a truly global, coordinated scientific community, solving problems for the benefit of society with research that cuts across borders and disciplinary boundaries.

Ethics of Big Data

Personal Identity, Authority, Privacy, Ownership and Reputation within data are key aspects to consider when discussing ethical issues related to the Big Data landscape. For instance, the offline and online identities of a person and the linkage of these two through analytics has produced an on-going debate on how we should deal with Big Data and how it should be used. Other aspects to take into consideration in the case of sensitive data are authoritative access and ownership issues (i.e., who should control access to the data, who owns it, and what are the potential obligations of people that either produce or use this data). As far as Reputation is concerned, the key aspect to consider is the trustworthiness of data (i.e., how can a person using the data know its origin, the process that was implemented in generating the data and so on). All these considerations showcase that interested parties should work upon a common landscape for metadata attributes.
Another ethical parameter that is very important has to do with personal information that is included in large databases. The practice is to use such data sets but initially filter out all private information that may possibly be included. According to the legislation of most European countries such practice is legal and allowed. Thus, an anonymisation procedure takes place in which all names, addresses, e-mails, etc. and any other personal information of the particular node are erased and at the same time the nodes are assigned only a node number, which contains all the network characteristics but no personal information at all. There exist standard procedures for such transformations. Typically one lets a computer program give every unique identifier a randomly selected unique number generating a code key. The code key will then be used to recode the identifiers. The Code key will then be deleted immediately. Such anonymisation procedures (and several other techniques) are standard nowadays and accepted in European legislation.

The future of Big Data

Currently, the common belief shared in the academic world is that people should contribute to Big Data by fostering openness and raising awareness of the importance of collecting, maintaining, and distributing high-quality, validated, standardized datasets to scientists of any background. Scientists should prepare themselves to handle huge amounts of data, as the flood of observations from numerous scientific experiments, sensors, and numerical

[20] http://www.codata.org

simulations is ever-increasing. There is great potential for discovery within the existing data, and scientists collecting extremely large datasets should concede that they alone might not be able to thoroughly use that data to its full potential. Scientists are also starting to realize the need for novel systems and interfaces to unify, organize, and distribute data and information. New technical solutions to deal with 'extreme' data, socially mediated content and new connected devices are considered the key element by most enterprises to derive business opportunities from big data; about 15% of the enterprises plan to adapt their information technologies abilities to reach this goal, according to Gartner[21].

Context: Current practices - achievements and limitations

Access to Big Data

Data access, in particular authentication and authorization methods, and the need for a cross-interoperable data infrastructure along several e-Infrastructures are of paramount importance. The ability to associate a person to a service endpoint is currently being covered distinctively by e-Infrastructure providers and several technologies are being used (x509 certificates, VOMS services, SSO tickets/tokens etc.). Currently a large investment is made in overcoming such obstacles through providing researchers across Europe with a federated identity management service.

Another issue to consider is related to Open Data access policies. Although the need for making data deriving from the public sector and several scientific experiments and investigations open and accessible this should not be true for all data collections (e.g., medical data) which, even if anonymised and stripped off personal sensitive information, should be accessible by authorized persons only.

Big Data often resides within Data warehouses. These are places where data is stored for archival and analysis purposes. Depending on the aggregation mode of data a data warehouse may be characterized as offline, in which case data is updated frequently (i.e. daily, weekly, monthly etc.), or real time, in which case data is aggregated synchronously upon generation. Data warehouses are also referred to as integrated as long as they support tools that allow access by other systems. Data residing upon a data warehouse may be structured (i.e., a database) or unstructured (i.e., data files on a filesystem level).

The advantages of using data warehouses is that data are kept clean, safe and secure and that through integration tools provided multiple options for query processing are provided.

The major current shortcomings of data warehouses is that control over the data is partially lost from the user perspective (i.e., ownership, responsibility and accountability may not necessarily be inherent in the data), it is very difficult to introduce changes to already existing data (i.e., schema changes, data types and ranges) and it is difficult or impossible to monitor changes in the data (as in some cases data is treated as static).

[21] http://www.gartner.com/technology/topics/big-data.jsp

Computing Tools and Methods

The Big Data movement has been driven so far by web companies such as Google, Yahoo!, Facebook, Amazon, and Twitter. They have shown how highly innovative technology combined with a clever business model can turn a start-up into a world-leading enterprise in no time. They have developed many of today's best known tools to manage Big Data, including Yahoo's Hadoop and S4, Facebook's Cassandra, and Twitter's Storm (often inspired by technology first developed at Google). And, most importantly, they released many of these technologies to open-source, making them available to the rest of the world.

Current Big Data architectures excel at the ability to ingest and process Petabytes of data. The first wave of big data systems was dominated by the Hadoop and the MapReduce approach. Google, Yahoo, and other companies have exemplified how to process massive amounts of data in a batch system using just commodity hardware. The use cases have been built for web based systems interacting with humans: e.g. analysing incoming Twitter streams, handling customer's orders at Amazon, managing social networks at Facebook, or indexing web pages at Google. The paradigmatic Big Data use-case is loading, storing, processing and indexing of data from the Internet. This data is unstructured or semi-structured at best, and predominantly persistent, i.e., it is required to be stored for relatively long periods of time.

As a result, Big Data technologies to date mainly focus on batch processing of data stored on distributed file systems; implementations are designed to be as general as possible, employing simple and general key-value semantics. NoSQL databases such as HBase or Cassandra gained prominence in such use cases, since they allow for data to be partitioned upon different machines, do not require a fixed relational schema, and provide better scalability, resilience to failures, and often superior performance compared to more traditional relational database systems.

But, over the years, the limitations of the original MapReduce approach became apparent. The second wave of big data systems tries to work around some of the problems. In a first avenue, some of the limitations have been fixed within the original design, e.g., providing an improved resource management framework (YARN) for Hadoop. A second avenue provides alternative implementations to Hadoop, e.g., trying to avoid Hadoop's excessive use of the file system, e.g., for storing intermediate results, or reuse of Java jobs in MapReduce to avoid start-up overheads, reporting large-factor improvements for some cases.

Other MapReduce limitations, however, cannot be fixed within the basic assumptions of MapReduce. The most important limitation is that MapReduce is a batch oriented approach that has a lack of support for ad-hoc, real-time queries. Many of the players in the Big Data field have realized the need for fast, interactive queries. Thus a third avenue for improvement is to utilize the Hadoop distributed file system, but bypassing MapReduce. An example for a SQL-like solution is Impala from Cloudera or Apache Drill (both inspired by Google's Dremel system). By directly working on the distributed file system they eliminate the need for starting many Java jobs. These developments promise to evolve Big Data management architectures beyond the original use cases, allowing low latency ad hoc queries using a familiar SQL-like interface. These systems are often advertised as real-time processing solutions; however, this is in the sense of reducing the latency of ad hoc queries, and not in the sense that they allow real-time stream processing. None of MapReduce, Impala, or Drill allows for real-time processing of data streams --- they work on data which is already there,

and they process it in batch mode. Thus, a more fundamental approach is to complement MapReduce-processing with true, real-time stream processing.

An example of such cluster-based stream processing systems is IBM's System S, which lead to the commercial Infosphere Streams product. Infosphere Streams is a dedicated stream processing system, where the processing of events is distributed among a dedicated cluster of machines. Depending on the hardware infrastructure and use case, millions of events can be processed per second. As another example, Twitter's Lambda architecture combines batch and stream processing in a single architecture and has thus a great potential for a unifying architectural scheme. Thus, it has recently become the object of much attention both in the academic and the commercial world. Twitter uses Storm for real-time processing and indexing of a stream of incoming Twitter messages, so that the view of the system remains up-to-date in-between the execution of two MapReduce jobs executed on the full static data set.

At the same time, there exist several significant, and ongoing, research efforts from the data-management community in the area of efficient data-stream processing algorithms. The focus here is on designing and implementing novel algorithmic tools that can enable effective processing and analysis of massive, rapid-rate data streams in small space and time. ("Small" here refers to quantities that are significantly sublinear (e.g., logarithmic or poly-logarithmic) in the data stream size.) Such guarantees are typically provided through the use of appropriate sketches (or, synopses) of the stream that can summarize key statistical properties of the streaming data within a limited memory footprint. More recent work also has also factored in the issue of "small communication", when processing massive distributed data streams (e.g., within an IP-network monitoring architecture, where probes are typically distributed within the production IP network). The combination of such algorithmic tools with cluster-based stream-processing engines is an area that will likely gain prominence in both research and practice over the coming years.

While there are today a number of implementations of stream and event processing systems, both commercial and in the open source area, Gartner analyst Roy Schulte estimated in July 2012 that around 95% of the event processing applications are built using ad-hoc programming and do not use existing frameworks[22]. Two of the main reasons are the difficulty to think in terms of event driven architectures which are asynchronous in nature, and the relative complexity and, often, the lack of maturity of existing tools, making them difficult to deploy and maintain in a production environment, as well as impractical and inaccessible for business users. Another point to consider is that, as several of these tools are often based on open-source contributions, optimization of the underlying technologies and quality assurance metrics are often not accounted for in the development life cycle.

Standards and Relationship between Big Data and Cloud Computing

e-IRG's Blue Paper on Data Management 2012[23] describes the most important European projects where data standardization, at least within the framework of homogeneous scientific areas, is carried on. There are similar activities also in the US and Australia as well.

However, these important efforts will take quite some time to deliver results but anyway they address only partially the problem since some important issues are related to the

[22] http://www.complexevents.com/2012/07/25/does-anyone-care-about-event-processing/

[23] http://www.e-irg.eu/images/stories/dissemination/e-irg-blue_paper_on_data_management_v_final.pdf

interoperability between domains and this depends, among the other factors, on the increasing data scale and the analytic complexity.
It is worth investigating how the Big Data main features mentioned in the previous section (volume, variety, veracity and velocity) will add complexity to the known issues of open data; this activity has already started within some "harmonizing" initiatives, such as the Research Data Alliance[24]. It is also important to take into account the results of the existing domain-specific activities on this item such as the BioMedBridges[25] cluster project..
An interesting experience from CODATA is the Group on Earth Observations (GEO) Data Sharing Principle. GEO is a voluntary partnership of governments and international organizations coordinating efforts to build a Global Earth Observation System of Systems (GEOSS). Since 2006, CODATA, as an interdisciplinary committee of ICSU, has played a lead role in supporting the implementation of the GEO Data Sharing Principles, which call for the "full and open exchange of data, metadata, and products shared within GEOSS, recognizing relevant international instruments and national policies".
GEOSS is a good example of how Big Data made its mark on the scientific data landscape. There was first the recognition that there was so much data out there that something needed to be done. Then, it grew into scientists talking and eventually organizing efforts to put a plan in motion. The result was a truly global effort and a tangible data portal to organize and distribute data. Something like GEOSS would not exist though without Big Data and a shift in attitudes that Big Data has brought to science.

Data portability and replication are general issues and their importance grows when the computing power needed to process data is so big that it can't be provided by either a single or a small number of data centres. Whenever distributed, and often heterogeneous, computing infrastructures are needed, compliance to standards becomes more and more important. This is the reason why some important initiatives, such as the Joint Cloud and Big Data Workshop[26] organized by NIST in January 2013, have recently started investigating the relationship between Big Data and cloud computing, the most promising computing technology currently under development. These two major challenges, Big Data and cloud computing, are not totally independent: not only because Big Data may require huge computing power, but also because Big Data could represent the killer application for clouds. This is the main reason why it is important that their relationship, their specific issues and also their mutual opportunities (e.g., how can big data find value on the cloud) are taken into account.

Proposed approach

In a recent communication document published by the European Commission[27] one of the key objectives that are being considered in relation to Data Access is the development (in cooperation with ESFRI, e-IRG and other stakeholders) of a Charter of Access setting out

[24] http://rd-alliance.org
[25] http://www.biomedbridges.eu/workpackages/wp3
[26] http://www.nist.gov/itl/cloud/upload/CCBDW_agenda_final_11413.pdf
[27] EU communication COM(2012) 392 final "A Reinforced European Research Area Partnership for Excellence and Growth"

common standards and harmonized access rules and conditions for the use of Research Infrastructures. In addition, member states are invited to work on harmonizing access and usage policies for research and education related activities on e-Infrastructures and provide complementary digital research services enabling consortia of different types of public and private partners.

Currently there are some projects and activities underway, which facilitate the establishment of standards and the creation of services for data-oriented communities. EUDAT[28] is an example of a project that provides a platform to coordinate the activities between data e-Infrastructure providers and research communities. It brings together data providers and users to facilitate the design of services for the data communities.

Among the various initiatives dealing with data sharing and exchange it is worth mentioning the Research Data Alliance, which is addressing the complexity of data and Big Data by means of opportunities and tools (such as fora, mailing lists, working groups and events) available to a wide research community. This initiative aims at accelerating international data driven innovation and discovery by facilitating research, data sharing and exchange, use and reuse, standards harmonisation for specific communities and across scientific disciplines. The RDA launch event, sponsored by the European Commission, the U.S. Government, the Australian Government and various important leaders in the data community, took place in March 2013 in Gothenburg[29]. RDA is supported by the European Commission also through the iCORDI[30] FP7 project. RDA has set up several working groups addressing the most important and challenging items about data. The PID Information Types working group aims to provide a framework for information types (i.e., information associated with a persistent identifier) and to define some initial essential information types while the Data Type Registry working group aims to develop and run a registry where data types can be defined formally. The Data Foundation and Terminology working group aims to provide a broad base of shared terminology relevant for data-exchange as sharing a terminology is the first step towards communication. The Practical Policy working group aims to gather running policies used for data ingest, curation, provenance, exchange, etc. as they are currently being used in a number of large data centres around the world. In a second stage they plan to offer a number of best-practice practical policy sets for several policy frameworks that comply with various standards. This way the current data centres can improve their policies and it will also make it easier to start a new data centre.

The Legal Interoperability working group has a wider scope since it is a joint RDA and CODATA interest group. They plan to look at legal issues with respect to interoperability and to provide recommendations and licenses to deal with these issues. Specific issues and licenses will be handled in separate short-lived working groups.

[28] http://www.eudat.eu
[29] http://rd-alliance.org/programme
[30] https://www.icordi.eu

For these initiatives to be successful, it is important that the main actors, in particular the user communities, be involved and contribute to investigate the issues and work out solutions. A bottom-up approach, based on real use cases, together with the investigation of the possible generalizations could be the better way to proceed. This approach has already proven to be quite successful in some different fields, e.g., data with CODATA and the EUDAT project, HPC with the PRACE project and HTC with the European Grid Initiative (EGI). To be successful also for Big Data, and even more for cloud computing needed for Big Data analysis, it is important that the funding agencies, the national governments and the European Commission promote and strongly support this path.

The need for standards for Big Data highlighted in the previous section can be addressed in at least two complementary ways:

- promoting and supporting all the parties involved in the activities towards standardization (user communities, projects, standardization bodies)
- promoting the creation of an abstraction layer to interface the different 'proprietary' solutions for data formats

These approaches, although both valid in principle, have a different scope: the adoption of a new and general standard for (Big) Data, when it will exist, will likely be suitable for future data producers (experimental equipment, applications, services, etc.) while the abstraction layer approach is more suitable for already existing data.
Both approaches are challenging and they will have a chance to be successful as long as users are engaged in the process and strong support is provided by an independent body, with the role of promoting and coordinating these efforts, where all the stakeholders (including the funding bodies) are represented.

Recommendations

In order to address the issues and challenges described in the previous sections some actions involving different actors could be envisaged.

International scientific communities

Contribute to the development of standards as well as tools and methods to facilitate Big Data exchange and interoperability; participate to the RDA Big Data Analytics working group.

Promote a bottom-up approach by identifying relevant use cases to be used for pilot activities. The generalization of specific solutions could be more effective than general solutions or architectures (top-down approach) developed from scratch.

Promote the usage of Data Warehouses for archiving and maintaining data collections.

Standardisation Organisations

Keep on promoting the strengthening of the existing standards and the development of new ones if necessary; promote a stronger involvement of user communities and support pilot activities; promote Open Data policies especially for data that do not contain sensitive information.

EC, Funding Organizations, National Governments

Promote and support events and initiatives where the relationship between Big Data and cloud computing is investigated and cross activities are carried on in order to exploit the potential of these two approaches together to provide effective services to a wide variety of users and facilitate innovation.

Support innovation and knowledge transfer on Big Data research via public-private sector partnership activities.

EC, Funding Agencies

Sustain projects and activities to investigate the big data features in order to identify and address the shortcomings and to force the adoption of the appropriate standards.

Sustain the development of new tools and methods provided they could enable the semantic processing of social media data and other complex data.

Sustain and support quality assurance metrics in the development of new tools and methods

Existing Services Providers

Adopt standards or open interfaces in order to make services and infrastructures interoperable and promote data sharing.

International organizations of se-Infrastructures

Adopt standards or open interfaces in order to make services and infrastructures interoperable and promote data sharing.

National governments

Encourage users and research communities to participate in and use European e-Infrastructures and achieve competence in Big Data research.

Cloud Computing

C1. Policy Area and Goal

Cloud computing got and still gets a lot of attention from the commercial world, the European Commission and the research world alike. Cloud computing is often mentioned as a major factor for economic boost, for saving IT expenditure for IT customers and bringing new income to IT vendors and IT service providers, while at the same time leaving room for innovation for all involved parties. Despite all this positive publicity for the cloud it is not yet clear if this is the solution for scientific computing including storage of scientific data.

Cloud computing comes in many flavours (Saas, Paas, IaaS, DaaS, etc.) and the corresponding services are often offered by different service providers. Hence should we see "the cloud" as a range of services that can be implemented on existing e-infrastructures rather than as a separate e-Infrastructure. How does cloud computing compare to supercomputing, HPC and HTC or to other services such as outsourcing?

Public (commercial) cloud computing services are available since a few years. While some cloud services, mostly in the Software as a Service arena, have a tremendous success the more compute and big data oriented services have met with some reticence. The barriers that are hampering the take up are manifold.

Cost of cloud computing is also an issue. Migrating to public computing and or data services needs a well-reflected decision supported by the necessary financial data. But practice reveals that making comparisons between computing and data services is not a trivial task.

And where is the user in the cloud debate? Does the type of e-infrastructure he has to use is of any importance to him. Probably the user wants to see his computing and storage needs solved in a transparent way, with a single user friendly interface.

C2. Context: Current practices - achievements and limitations

C2.1 Each new computer paradigm that evolves to a new e-infrastructure raises the discussion if the new e-infrastructure is the ultimate solution that will make all the existing ones redundant. This has always been the case with cloud computing. In a first instance cloud computing was the user-friendly e-infrastructure that was the answer to the complicated grid computing that rooted well in the scientific environment but did not really succeed in the commercial world. A different business model might have been at the basis of the acceptance or non-acceptance of the concerned e-infrastructure. Grids and clouds have a lot of common characteristics and can house almost the same type of applications. Differentiations are in the basic business model, the security approach (simpler but less safe in the cloud, complex but safer in the grid) and often the way users get access to the infrastructure. Grid and cloud generally support the same type of applications but often the grid is used for more complex applications with a large volume of data while clouds are currently used by less complex ones. Supercomputing and HPC in general are considered to yield services for complex applications that will explore the characteristics of the hardware to use that expensive infrastructure as efficiently as possible and to obtain maximum performance. Access to the HPC infrastructure is most of the time not possible via the web or user friendly interfaces. Cloud computing does normally not offer the possibility for applications to fine-tune on the

hardware or a specific hardware configuration but rather aims to hide the hardware reality from the user. However if we look at the cloud as a large range of services it becomes more and more clear that many existing e-infrastructures services could be offered in the cloud model while holding their specific characteristics and this to the advantage of the whole user community. While this option is most probably technically possible, financial barriers may stop this evolution.

C2.2 Massive uptake of cloud computing has not happened and this is for several reasons. Security issues (risks on data protection, risks of errors in data management, break-ins, etc.) belong to the main barriers for the use of cloud services in both the commercial and research environment. This is especially important for data stored in the cloud. Other barriers include the fear for vendor-lock-in. Data and application interoperability and portability are almost non-existing or certainly not guaranteed or come with a high associated cost. An absence of cloud computing expertise can be a significant barrier. Given the perceived user-friendliness of clouds it is often assumed that its use is straightforward for non-technical users and that no training is needed. But this is certainly not the case and significant expertise needs to be acquired to adapt a specific application to the cloud. The research world has often an extra barrier in the mismatch of the funding and the business model of the cloud. Barriers can also be related to legal issues as cloud deployments can involve more than one country and therefore have to do with laws that are country dependent.

C2.3 Cost issues related to the choice for public cloud services require a special attention and at several levels:

- Comparison of costs: Recent European project like e-FISCAL and Helix- Nebula show that it is very difficult to compare costs for the use of different e-infrastructures or different cloud services. Comparing the cost of installing and maintaining a local computing/data infrastructure to the cost of the corresponding cloud services is very difficult. Comparing costs between several cloud service providers hurts the same level of difficulty. This difficulty arises from the fact that hardware offered in the different organisation is not the same and/or services are based on virtual machines in the cloud while this is not the case on a scientific e-infrastructure. SLAs might be non-existent or too different to compare. Quality of service (including availability and reliability) might not be defined for the cloud service, …

- Type of research funding: Already mentioned as a barrier above, there might also be a cost issue at the user level, a researcher who wants to use some computing cycles on a pay-per-use base on a public cloud might not have the funding that permits such expenses.

- Investments in equipment at the side of the cloud service provider and uncertainty at the side of the customer: as mentioned before cloud computing is until now not really used for complex computing and a very large amount of data. To be prepared to offer cloud services in that domain the provider has to invest a large

sum without being sure that the service will be sold. This leads to a denial of investment or to a high price for the service often accompanied with a high switching cost (cost to for moving an application to a new provider). Long-term contracts would alleviate this burden. However the customer does not want to sign long term contracts with one provider as this can lead to a loss of money when the IT market changes and a lot of flexibility to change to anther provider that gives a better or more adapted service.

C2.3 Standards could also help to remove some of the barriers to cloud computing. However the standardisation field is complicated and not always adapted to real life. According to the Vice-President of the European Commission responsible for the Digital Agenda, international standardisation efforts will also have a huge impact on cloud computing. *"Open specifications are a key in creating competitive and flourishing markets that deliver what customers need. Europe can play a big role here.* The EC Strategy for Cloud computing unveiled this summer (2012) emphasizes the importance of standards: "Cutting through the jungle of technical standards so that cloud users get interoperability, data portability and reversibility; necessary standards should be identified by 2013". But the hype around cloud has created a flurry of standards and open source activities leading to market uncertainty. SIENA is the first initiative to bring to the same table standardization bodies to support the analysis of open standards-based interoperable grid and cloud computing infrastructures.

C2.4 At the end of 2012 the status of cloud e-infrastructures at the European level can be summarised as follows. Work done by the e-Infranet project in 2011 and a recent questionnaire about cloud computing issued by eIPF (e-Infrastructure Policy Forum) unveiled that very few countries have a roadmap to deploy a national cloud infrastructure for research (only 2 from the 19 respondents). However in most of the countries there are on-going pilot projects on cloud computing initiated by universities, research institutes or already established e-infrastructures. All the countries welcome collaboration at the European level to establish a (federated) cloud for research in Europe; however some reservations apply including that the user need for this kind of infrastructure is clearly stated, that there is a clear funding mechanism, or that the required services cannot be delivered by existing private or public services.

C2.5 And where is the user in this cloud story? Looking down to the history of computing infrastructures and more recently the national and international e-infrastructures, the end-user or the researcher in the case of the research environment has not had much to say in the whole set-up. Computing infrastructures are/were set up by ICT professionals that opted for the most performant environment or the best value for money or the maximum equipment for money, sometimes losing sight of the user. Some e-infrastructures are set up by user communities that fine-tuned the infrastructure to their needs but never thought about the usability by other communities. The cloud computing services were set up by the ICT industry as an innovation to services that were already available. Also in this case the demands of the researcher or the end-users in general were not taken into account. These errors committed in the past lead to a fragmentation of effort and funding in the set-up and maintenance of e-infrastructures and often lead to rivalry between e-infrastructures.

Proposed approach

Sections C2.1 and C2.4 show that the integration of cloud services with existing e-Infrastructures at national and international is a preferred scenario and also follows the

interest of the countries to collaborate at international level. Existing e-infrastructures and new e-infrastructures should collaborate to find the best formulas that give the best service to the end-user. Inclusion of services provided by commercial e-infrastructure providers has to be considered. Section 2.5 urges that the end-user is involved in this collaboration.

Work on standards has to continue but needs the contact with the existing e-infrastructures and has to proceed as quickly as possible. Standards could also reduce the barriers mentioned in C2.2 and the financial issues described in C2.3. Work on standards should be much more supported, also financially, by the EC and the national governments. Too often standardisation work has to be done by people that already have a fulltime task elsewhere.

As mentioned in C2.2 and C2.3 funding mechanisms at the national and European level should be adapted to the changed computing environments to enable use of cloud services in a hybrid environment (mix of private and public cloud) or the use of the best suited e-infrastructure service at any moment.

C2.5 shows that attention has to be given to end-user involvement in the set-up and subsequent operation of an e-infrastructure. This is not an easy task. Where are the users? Do we have to work with user communities representing their end-users? And how is the coordination done. All those questions have to be answered at the national and European level.

Recommendations

The recommendations below distinguish between infrastructure-related recommendations and support actions and both are addressed to national governments and funding agencies and to the European commission.

Infrastructure-related recommendations at the national level

National governments should:

- Support the integration of cloud technologies in existing e-infrastructures
    - Adapt national e-infrastructures and ensure the participation in the European e-infrastructures in order to be able to exploit upcoming European cloud resources
    - Promote and financially support the innovation and evolution of national public e-infrastructure providers
- Stimulate the integration of several e-Infrastructure components at national level, so as to facilitate a single point of access for European researchers.

Infrastructure-related recommendations at the EC level

The European Commission should:

- Evaluate the use of commercial cloud resources as part of a hybrid community cloud environment during Horizon 2020 and support the development of related business models for the procurement of such commercial resources
- Evaluate the integration of several e-Infrastructure components at European level, so as to facilitate a single point of access for European researchers
- Stimulate all member states to participate in the European e-infrastructures, including cloud initiatives, to develop innovative and interoperable services

Recommendations for support-actions at the national level

The national governments should:

- Establish the necessary policies, rules and legal framework including SLAs allowing the funding and use of public cloud resources for research activities and work on the control procedures of such resource usage
- Participate actively in the standard creation process and ensure that user communities have the possibility (financial) to contribute to the standardisation process
- Promote the establishment of repositories of standards and applications running in the cloud to support the reproducibility of the research experiments and allow the take-up by the commercial sector
- Support the provision of training activities for new technologies such as virtualization and cloud technologies in cooperation with related European activities and industrial entities.
- Ensure that the end-user and user communities are involved in the definition, set-up and exploitation of national e-infrastructures

Recommendations for support-actions at the EC level

The European Commission should:

- Promote the establishment of the necessary and harmonised policies, rules and legal framework including SLAs for the use of cloud resources for European research activities and work on the control procedures of such resource usage
- Invest in research, methodology and development that ensure the elimination of the vendor-lock-in problem and promote interoperability among commercial and research-owned clouds and grids
- Invest in the standard making process
- Invest in research about management, provenance and privacy of the data in cloud environments
- Ensure the involvement of the end-users and user communities in the definition, set-up and exploitation of transnational e-infrastructures
- Support the provision of training activities for new technologies such as virtualization and cloud technologies in cooperation with national and regional activities, also involving industrial entities.

Facilitating Use of State-Funded e-Infrastructures by non-State-Funded Parties: Legal Issues

Policy Area and Goal:

Existing state-funded e-infrastructures are mostly used by state-funded researchers in universities and other academic research institutions. There are also significant potential benefits for both organisations and society if non-state-funded research and development could also use data-intensive computing and storage resources, high-performance hardware, and specialized networks to connect them. For example in e-Health access to, and processing of, genomic information could benefit public-private collaborative research and development or R&D by private companies. Similar opportunities are likely to exist in very many other fields of research and industry; these are already being exploited on a small scale in some countries. Facilitating use by a wider range of researchers as part of the e-infrastructures' existing mission will be essential to achieve Horizon 2020 goals.

Concerns have, however, been expressed that legal and regulatory barriers may hinder these wider uses of e-infrastructures. This paper reports the conclusions of an investigation of relevant areas of European law to determine whether they do, or could in future, create such barriers.

Six areas of law and regulation have been identified as relevant to use of state-funded e-infrastructures by non-state-funded parties: state aid law, public procurement law, network regulation, data protection, terms of use of e-infrastructure providers, and software licenses.

Context

As the detailed sections on individual areas below indicate, the investigation concluded that current European law should not prevent the use of e-infrastructures for research and development by non-state-funded parties, subject to the State Aid Block Exemption's limits on the contribution that the state can make to such activities (typically 50%) and the requirement that any state aid create an incentive effect. However lack of clarity about the application of State Aid and Data Protection law, and differences between national implementations, may create actual or perceived barriers because of difficulties in ensuring compliance.

Expanded use may, however, be prevented by the current terms of use of e-infrastructure components and by licence terms of the software that those components rely on. If e-infrastructure procurements were too narrowly drawn there may be a risk of challenge if the scope of use is subsequently expanded. Infrastructures wishing to support new uses will need to review, and possibly amend, these. Specific procurement and taxation exemptions for public sector infrastructure operators may also limit the overall amount of private sector involvement they can accept.

Finally, new laws in the areas of State Aid, Data Protection and Network Regulation are all being developed at European level. The Commission's proposal to modify State Aid law explicitly mentions and supports the wider use of e-infrastructure. Those developing other legislation may well be unaware of its potential effect on e-infrastructure use, even though the impact could be severe. Data Protection legislation intended for commercial cloud services could make the law even harder to apply to research infrastructures; Network Regulation that requires particular services or implementations could be incompatible with

flexible, innovative, high-performance research services. These, and other, unintended consequences of legislation need to be identified and avoided.

State Aid

e-Infrastructures are generally supported by state funding, which will confer a benefit on users, operators and providers of the infrastructure. Some or all of these may be acting as economic undertakings. It is also possible that e-infrastructures might be used in ways that distort existing competitive markets. This could apply particularly to GEANT which has the power to reach across national borders at extremely high capacity. European State Aid law is therefore likely to apply either directly or through the conditions imposed on any relevant exemptions.

A block exemption for research, development and innovation allows the state to contribute funding or support to projects by economic undertakings, up to a specified proportion (typically 50%, with the permitted contribution reducing as activity gets closer to production) and provided that the contribution results in activities that would not have taken place under pure market conditions. Some formalities are required in calculating the respective contributions and to demonstrate the incentive effect, though the latter requirement is waived for Small and Medium Enterprises (SMEs). These requirements may create a perception that applying state aid law is difficult: the Commission's paper on modernizing State Aid suggests that the formalities might be improved in the next version of the block exemption after 2013, and specifically mentions e-infrastructures as an area to be supported. Clarification and simplification would be helpful for non-state-funded use.

More challenging problems may arise in calculating the value of the state contribution. Normally this is done by comparing against a market rate, but for many of the services provided by e-infrastructures there may be no relevant market to compare against. Assigning a value to new intellectual property or to long-distance, high-performance network services that do not exist outside research and education may be particularly difficult. This might be simplified by the distinction already recognised in some countries between innovation costs and operational costs. Guidance on acceptable ways to value state contributions, and on revenue sharing or other ways to deal with intellectual property, could reduce uncertainty for both e-infrastructure providers and their commercial partners.

Offering fully commercial services using an e-infrastructure would not be compatible with the Block Exemption, so would need to use other mechanisms to avoid challenges under State Aid, and possibly Competition, Law. These may well require new organisational structures – in other areas state assets are made available through trading companies that buy services from the state and act as economic undertakings in their own right.

Procurement law

e-Infrastructures are likely to involve large procurements by public authorities, which will be subject to EU procurement laws. Pre-Commercial Procurement (PCP) and Public Procurement of Innovation (PPI) procedures may be appropriate, particularly for the development and prototyping of innovative technologies or applications. As these areas are not well understood by practitioners at the present time, this may require particular attention.

Current procurements for shared public sector e-Infrastructures may benefit from the *Teckal* exemption, which can help National Research and Education Networks (NRENs) develop and provide services to public authorities without the need for each authority to run a separate public tender exercise. If tenders were required, publicly-funded NRENs might be prohibited

by State Aid law from responding to them. The exemption is only available where the relationship between the NREN and its customers involves both structural control and economic dependency. However the draft Public Procurement Directive suggests that this will only apply where at least 90% of the NREN's activities benefit the public-sector community. In some countries taxation arrangements for public-funded infrastructure operators may have similarly high thresholds. These requirements might well limit non-public-funded use of e-Infrastructures to a low level, thus conflicting with the desire of State Aid law and the Horizon 2020 goals to expand such use.

Procurement law requires the purpose of the procurement to be specified. Provided future e-infrastructure procurements include the possibility of non-state-funded use this should not be a problem. However if procurements of existing e-infrastructures contained statements that ruled out expanded use there may be a risk of challenge if the change of scope would have affected the bids made. This possibility will need to be reviewed by e-infrastructure operators considering extension to non-state-funded use. On occasion it may be appropriate for e-infrastructures to consider public procurement of innovation (PPI) where a product or service required by the e-infrastructure is not currently available on the market. Since this is likely to involve state funds underwriting some of the risk of developing a new product it may raise the same issues identified in the previous (State Aid) section of pricing and managing the benefits of creating newly-created intellectual property.

The investigation also considered possible roles for pre-commercial procurement (PCP) relating to e-infrastructures. According to the Commission's guidance, PCP involves funding research, rather than buying a product or service. If suitable research questions arise during the design or construction of e-infrastructures then, provided there are sufficient candidates to compete for the work, a PCP competition might be an appropriate vehicle. However existing mechanisms for providing research grants should also be considered. Alternatively, access to e-infrastructure services might be offered as part of a PCP competition in some other field of research, in which case the grant of access would be covered by normal State Aid provisions discussed above.

Network Regulation

e-Infrastructure components may be covered by two separate areas of European law. Networks and connectivity are likely to be classed as Electronic Communications Services, covered by the Telecommunications Framework Directive and associated legislation; data storage and processing services may fall within the definition of Information Society Services in the e-Commerce Directive and others. Assigning regulatory duties to specific parties within an e-infrastructure that comprises connectivity, storage and processing components under the control and management of several different organisations may prove difficult. However the current laws regulating private communications services and information society services do not appear to present significant problems for non-state-funded use of e-infrastructures.

Were the status of National Research and Education Networks (NRENs) as private electronic communications services to change there would be much more, and much less harmonised, regulation to accommodate. In particular public networks could be required to implement particular designs and technologies that would restrict the ability to provide flexible advanced communications facilities, such as bandwidth on demand, that are critical for high-performance e-infrastructures. NRENs should therefore ensure they continue to offer service to demarcated groups of users, in order to keep their private status.

A recently proposed Directive on Network and Information Security illustrates the risk of legislation having unintended consequences for e-infrastructures. The draft Directive creates a special category of Information Society Services, known as "market operators", that could be required to implement specified processes for preventing and responding to security and privacy breaches. Although the Directive is aimed at payment services, blogging sites, etc. it is possible that an e-infrastructure service might fall within the definition or a market operator. Since the duties to be imposed are designed for consumer platforms, it is unlikely that they would be compatible with the very different design and user relationships of an e-infrastructure service.

Data Protection

Where e-infrastructures are used to process personal data that use will be subject to Data Protection law. Information about the users of the infrastructure is also likely to be regulated. Both may raise new issues when processing is done across an e-infrastructure provided by multiple organisations rather than by a single data controller; even the Article 29 Working Party were apparently unable to assign the critical roles of data controller and data processor to the various components of a 'research grid'. If e-infrastructures cross national borders then the problems are worse as national implementations of the European Directive differ, and are sometimes contradictory, on questions as fundamental as what constitutes personal data and what formalities are required to process it.

Problems of interpreting and complying with data protection law already limit the use of e-infrastructures by public-sector researchers. In e-Health it is not clear (and national laws may differ) whether the research exemption may be used, or whether explicit consent is required from every person whose data may be processed. The absence of clear, authoritative guidance on these questions is likely to delay significant health benefits. Non-state-funded use is unlikely to alter the problems, though private sector organisations may be more concerned about the resulting regulatory uncertainty.

A new Data Protection Regulation is currently being debated. As a Regulation it should reduce differences between Member States, however the implications are far from clear as more than 4000 amendments to the Commission's original draft have been proposed in the European Parliament and Council. Depending on their eventual definition and implementation new policies such as the rights to be forgotten and to data portability could significantly affect the e-infrastructure model. The new law seems likely to favour approaches such as Privacy Enhancing Technologies and Privacy by Design: e-infrastructures should consider how these can be used to reduce both privacy and regulatory risks.

Terms of Use of e-infrastructure providers

Existing research networks and services have been established under different legal and political bases and with different rules for what types of organisation may connect to them. Few research networks currently provide direct connections to commercial organisations. This may not be a barrier if commodity internet connectivity is sufficient to reach and use the e-infrastructure service, but may prevent the use of services requiring high-performance or specialist network connections. In particular, unless terms of use are harmonized internationally, it is likely to be difficult to provide users outside the traditional state-funded research and education community with high-performance connections to international e-infrastructures. Many research networks are now considering how to connect commercial service providers where this would be of benefit to research and education; these discussions should also consider whether there is a case for allowing connection by commercial users of

services at academic organisations. Many network policies prohibit charging for access to the network, which may conflict with requirements under State Aid law to at least account for network use on a full economic basis.

e-Infrastructure services seem somewhat more likely to have policies that permit non-state-funded participation in R&D projects, though there is considerable variation both between services and between projects. Several only allow use by academic researchers. In the short term it may be sufficient to amend the policies of individual projects, however if existing e-infrastructures are to be linked into a general-purpose facility there will need to be national and international harmonization of at least the basic rules permitting access.

Software licences

Processing, storage and communications components of the e-infrastructure all rely on software that is subject to licences. e-Infrastructures for academic use may have obtained licences for standard software that are limited to non-profit use, to particular subject areas or to particular groups of users. Such licences may need to be extended, replaced or re-negotiated to permit different types of use, such as those involving non-state-funded partners. This may involve an increase in the licence fee.

Where bespoke software has been developed for an e-infrastructure, particularly if this was done during research projects, there is a possibility that no licence or intellectual property agreement was made with those contributing. In the worst case this can result in a situation where it cannot be determined either whether the software can be used for new purposes or who are the owners of intellectual property in the software whose agreement would be needed for such use.

e-Infrastructures considering expanded use should identify the licences they have and whether there are any that do not permit non-state-funded use. Where bespoke software is being developed, rights to the intellectual property should be agreed before development starts. To reduce the possibility of state aid problems if publicly-funded software is subsequently used by economic undertakings, licences should be as open as possible to avoid discrimination between undertakings.

Proposed approach

In order to reduce both perceived and actual barriers to the expanded use of e-infrastructures, action is needed at two levels: clarification and, where possible, simplification of the legal position by national and European regulators and legislators; combined with removal of barriers within existing e-infrastructures and components. Successful exploitation of e-infrastructures is also likely to require an investigation of issues relating to research content: some that were identified during the investigation of barriers are noted here but this should not be taken as a comprehensive list.

Regulatory clarification

Although the investigation concluded that the current Research, Development and Innovation Block Exemption does permit non-state-funded parties to use state-funded e-infrastructures for research and development, there is considerable uncertainty over how to achieve this. The Commission's paper on Modernising State Aid intends to facilitate the use of e-infrastructures: we believe this can be achieved by providing clear and easy to use guidance that provides certainty for funders, public and private participants. In particular we consider that clearer guidance would be helpful on assessing whether a project satisfies the requirements of the exemption in terms of aid intensity and incentive effect, and the

formalities required to report on applications that are granted. There appears to be particular uncertainty over applying the exemption to larger organisations. Authoritative advice on how to value services and results for which there is no obvious market price – including intellectual property, developed software and international, high-performance network connections – would increase confidence among both providers and researchers that their use of e-infrastructures is not exposed to legal challenge.

In procurement law there is a concern that including possible non-state-funded use in the scope of procurements might be open to challenge on the grounds that definitions are too wide and uncertain. Guidance on appropriate scope definition would be helpful. Inappropriately low thresholds for the *Teckal* exemption or taxation arrangements could also create a barrier to increased use. Guidance on how Public Procurement of Innovation can be carried out in compliance with State Aid laws might increase confidence in the use of this mechanism. Pre-Commercial Procurement of research and development might be used to develop future e-infrastructures if existing grant-funding mechanisms are inadequate, however the apparent complexity of the approach indicates that this would best be done in collaboration with a partner organisation that was familiar with the mechanism.

In future both e-Infrastructures and projects using them may be conducted by public-private partnerships (for example between a public genomics database and a private health provider). The requirements for such partnerships to comply with both state aid and procurement law appear unclear, which may act as a barrier to this type of development.

Uncertainty over the application of data protection law is limiting the use of national and international e-infrastructures for both public and private sector research in socially important areas. Varied definitions of personal data and formalities for handling it, and the difficulty of assigning the roles of data controller and data processor, are particular problems for international e-infrastructures. If the current revision of the law does not provide greater clarity and harmonization and give clear policy guidance on how appropriate research can be performed within the law, then European research in social and health sciences will not achieve its potential. The investigation concluded that, apart from Data Protection, current legislation does not create significant barriers to expanded use of e-infrastructures, though there are risks that future legislation on networks and networked services may do so. Impact assessments for such legislation should include the effect on e-infrastructures so that unintended consequences are identified and avoided.

Development of e-Infrastructures

Expanding use of e-infrastructures is likely to require changes to current access policies for both networks and services. Where discussions are already taking place about allowing use of networks by private-sector service providers, these should also include consideration of use by private-sector service users. However there are considerable benefits for e-infrastructures in remaining within the definition of a private communications service and this may be essential for providing flexible, innovative services. Expanded access policies should therefore avoid placing this status at risk, for example by ensuring that the organisations able to connect to the network are sufficiently demarcated rather than left completely open. Building international e-infrastructures will require that access policies be harmonized across both networks and services, to avoid creating policy barriers to interoperability. Policies should at least adopt a “country of origin” principle, by respecting the access policy of the network or service where a user first connected. Greater policy harmonization should be encouraged as part of work on end-to-end services.

e-Infrastructures considering expanded use will need to identify the software licences they have and whether these limit the uses that can be supported. If necessary, plans (and budgets) should be made for expanding key restrictive licences. These inventories and plans should be updated as new software licences are obtained, though organisations should avoid limited licences unless the benefits clearly justify the limitations. e-Infrastructures should have clear policies on the ownership and licensing of bespoke software, to avoid the risk that uncertainty over intellectual property rights will make this unusable in future. Past procurement documents should be checked to determine whether expanding use creates a risk of these being challenged.

These procurement, licensing and policy development activities should help to provide data for the service costing models that will be required for State Aid compliance.

Finally, all e-Infrastructures should consider the use of Privacy Enhancing Technologies (PETs) and Privacy by Design (PbD) approaches both within Europe and by international partners. Federated access management, a recognized PET, is already used for authentication and authorization by some e-infrastructures. The application of PETs and PbD to protecting research data within e-infrastructures is itself a valuable research area.

Content related issues

The investigation noted two areas of regulation and policy that may in future affect the content processed within e-infrastructures, and where there is a risk that inappropriate laws could create unnecessary barriers. A full investigation of content-related barriers was not undertaken, as it would require subject specialists' knowledge. However these areas should be monitored in future.

As the use of e-infrastructure increases, the reliability and integrity of the information that is processed, and ultimately relied upon, may be subject to the legal system. For example in the case of drug discovery and DNA sequencing, the integrity of the data produced could be subject to health and safety regulation; or in safety-critical research questions of liability for failure might arise. Future laws regulating modelling and simulation will need to be aware that these types of research will not necessarily be done within the boundaries of a single organisation.

The developing Open Access agenda may affect both the information available for processing in e-infrastructures and the products generated by them. While open access should, in general, support the work of e-infrastructures, licences and agreements that give significance to the location where information is held, processed or created may become increasingly hard to apply.

Recommendations

Recommendations for support-actions at the EC level

Recommendations to the European Commission for State Aid and procurement laws that promote the use of e-infrastructures:

- Ensure that the new exemption for research, development and innovation is clear, easy to use, and provides certainty for funders, public and private participants. In particular

- Provide guidance on the necessary formalities, demonstrating incentive effect, and valuing contributions from state and private sources (particularly when those are not SMEs);
- Provide guidance on determining reference prices for products and services where there is no effective "market price", such as intellectual property, bespoke software, specialized services and international networks;
- Clarify any risks arising from the use of international state-funded networks to access services;
- Ensure that a revised Teckal exemption does not create barriers to wider use of public-sector e-infrastructures.

Recommendation to the European Commission to facilitate the development of e-infrastructures by public/private partnerships:

- Provide guidance on the application of procurement and state aid law to the establishment of e-infrastructures by public/private partnerships.

Recommendations to the Commission, regulators and national legislatures on Data Protection law that facilitates the use of e-infrastructures:

- Harmonise and clarify the application of key data protection concepts – controller/processor, personal data – to international e-infrastructures;
- Ensure that rules and formalities for processing personal data are either harmonized, or at least based on a clearly-defined, single country for each processing activity;

Recommendation to the Commission and national legislatures on other laws relating to networks and networked services:

- Include effect on international e-infrastructures in the impact assessment of new legislative proposals.

Recommendations for e-Infrastructure Providers

Recommendations to e-infrastructure operators to facilitate expanded use of e-infrastructures:

- Network operators should consider how to extend their access policies to support wider use, but
- Network operators should ensure their policies comply with the definition of a private communications service, for example demarcating the communities to which connection is available;
- Network operators should harmonise access policies internationally to facilitate the provision of end-to-end services and avoid creating barriers to international research use.
- Infrastructure services should consider how to extend their access policies to support non-state-funded use
- Infrastructure services should harmonise their policies across projects, services, and countries, to facilitate inter-operation of their e-infrastructures.
- e-Infrastructure operators should assess whether existing and future software licences limit use and, if necessary, plan how to extend those licences

- e-Infrastructure operators should ensure they have clear policies for ownership and licensing of any bespoke software they may develop or use.
- e-Infrastructure operators should assess the risk of past procurements being challenged if use is extended to non-state-funded research and development.
- e-Infrastructure operators should develop models for costing their services under State Aid exemptions.

Recommendations for support-actions at the EC level

- e-infrastructures should consider the use of Privacy Enhancing Technologies (PETs) and Privacy by Design approaches both within Europe and by international partners.

Annex I – Editorial Responsibilities

Introduction:
Jan Wiebelitz

Chapter 1, e-Infrastructure Commons 2020: Integrated Services via interoperable e-Infrastructures
Champion: The Netherlands, Editors: Arjen van Rijn, Kees Neggers
Contributor: Frank van Iersel

Chapter 2, Open Science
Champions: Germany and France, Editors: Gabriele von Voigt, Anne Decrouchelle
Contributors: Francoise Genova, Lambert Heller, Jan Wiebelitz

Chapter 3, Policy Requirements for ESFRI Projects
Champion: Poland and Sweden, Editors: Norbert Meyer, Sverker Holmgren
Contributors: Gelsomina Pappalardo, Bjørn Henrichsen, Lorenza Saracco

Chapter 4, Big Data
Champions: Italy and Greece, Editors: Luciano Gaido, Panos Agyrakis
Contributors: Jan Wiebelitz, Paschalis Korosoglou, Minos Garofalakis

Chapter 5, Cloud Computing
Champion: Belgium, Editor: Rosette Vandenbroucke
Contributors: Maciej Brzezniak, Radek Januszewski, Fotis Karagiannis, Dana Petcu, Marcin Plociennik, Imre Szeberenyi, Paweł Wolniewicz

Chapter 6, Legal Barriers to Commercial Use of e-infrastructures
Champion: England, Editors: Andrew Cormack, Bob Day
Contributors: Paul Lewis, Sandra Oudejans, Willemijn Waisvisz, Eirini Kontrafouri, Nikolaus Forgó, David Foster, Panos Louridas

Annex II – Glossary

CERN	European Organization for Nuclear Research (originally the Conseil Européen pour la Recherche Nucléaire)
Commons	a resource management principle by which a resource is shared within a community (refer to "Infrastructure: The Social Value of Shared Resources", 2012, B. Frischmann)
e-Infrastructure	an environment to share research and educational resources (e.g. network, computers, storage, software, data) so that these resources can easily be accessed and used by academia, researchers and scientists as required
e-IRG	e-Infrastructure Reflection Group, the European international advisory group for e-Infrastructure related policy makers
EIROFORUM	a partnership between eight of Europe's largest inter-governmental scientific research organisations that are responsible for infrastructures and laboratories: CERN, EFDA-JET, EMBL, ESA, ESO, ESRF, European XFEL and ILL. (http://www.eiroforum.org/)
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
ESFRI	European Strategy Forum on Research Infrastructures (http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri)
GÉANT network	The pan-European research and education network[31] (originally: Gigabit European Academic Network)

[31] The term GÉANT can be used in different meanings, as in: "GÉANT network" for the pan-European network backbone, "GÉANT project" for the GN3+ project and its predecessors, and "GÉANT2020", as used by the GEG for a GEANT2020 European Communications Commons.