EIROforum IT Working Group
2nd December 2014

e-Infrastructure for the 21st Century - *one year later*

EIROforum partners are intergovernmental research organisations – CERN, ESA, EMBL, EFDA, ESO, European XFEL, ILL and ESRF – covering disciplines ranging from particle physics, space science and biology to fusion research, astronomy, and neutron and photon sciences. The partner organisations have a truly European governance, funding and remit, and in many cases share a global engagement. They are world leaders in basic research, as well as in managing and operating large research infrastructures and facilities. The EIROforum collaboration is helping European science reach its full potential through exploiting its unparalleled resources, facilities and expertise. By combining international facilities and human resources, EIROforum exceeds the research potential of the individual organisations, achieving world- class scientific and technological excellence in interdisciplinary fields. EIROforum works closely with industry to foster innovation and to stimulate the transfer of technology.

Abstract:
This document assesses the progress that has been made in achieving the vision of e-Infrastructure for the 21st century and identifies a set of priorities with actions for future work.

Prepared by Bob Jones (CERN) on behalf of the EIROforum IT Working Group.

Executive Summary

The *e-Infrastructure for the 21st Century* document published in 2013 identified several aspects of e-Infrastructures where improvement would lead to efficiencies and increased adoption by a broad range of research communities. Advances have been made in the implementation of a number of e-Infrastructure services. Some of these technical developments have been deployed in production environments and more will follow. Non-technical aspects including e-Infrastructure integration, governance and funding models take longer to address but there are indications that there is a willingness to consider change by several e-Infrastructure service operators and funding agencies. A number of these areas are considered to be high priority meriting closer attention and more determined action if the e-Infrastructure for the 21st century vision is to be achieved:

- Accelerate the pace of integration of e-Infrastructure services
- Open access to data
- Long term data preservation
- Data analysis platforms
- Evaluate new funding models

An e-Infrastructure commons exchange providing an open environment where users can flexibly discover and choose the services and service providers from the public and private sectors that best meet their needs could address the priorities listed above in a holistic and inclusive manner. Specific actions to implement the exchange include:

- Integrate publicly funded e-Infrastructures with commercial services to provide a combined platform on which to build new services. Integration should start with a common catalogue of services and a federated identity management system offering a single sign-on facility to access services across all suppliers.
- Define a lightweight governance model for the exchange to ensure it is operated in a transparent and inclusive manner for European service providers and users from both the public and private sectors.
- Establish a network of public research organisations to implement joint procurement actions with the goal of augmenting the capacity and functionality of the platform.
- The service providers should jointly develop a communication plan to promote the existence of the exchange to audiences including research communities, policy makers and funding agencies.
- Engage funding agencies to sponsor the activities of researchers and SMEs to use and develop new services on the platform.

Introduction

The EIROforum IT Working Group published in 2013 the *e-Infrastructure for the 21st Century*[1] document that presented a vision for an integrated set of e-Infrastructure services to support the growing needs of data intensive science.

One year later, this document summarises the progress that has been achieved in implementing this vision and identifies a number of high priority areas and action points.

Assessment of progress

The *e-Infrastructure for the 21st Century* document identified a number of areas of e-Infrastructures where improvement would lead to efficiencies and further adoption by research communities and opportunities for increasing their impact. This section assesses the progress against a number of those e-Infrastructure areas.

E-Infrastructure Services

Core network

GÉANT continues to extend its geographical coverage and range of networking services. With the help of GÉANT, CERN has been able to put in place the first trans-national 100Gbps link between Geneva and the Wigner data centre in Budapest[2]. A second link has also been established with the commercial provider Deutsche Telekom[3].
GÉANT has been actively engaged in the Helix Nebula initiative where it is working with ESA to help the earth science community better exploit the hybrid commercial-public cloud infrastructure through connections to Japan.

Federated identity management services

Through the work of the Federated Identity Management for Research Collaborations group (FIM4R[4]) initiated in 2011 by EIROforum, a number of research communities have deployed pilot systems many of which are using eduGAIN. Development work has continued on federated identity management but there is no European-wide production system currently in production use primarily due to policy rather than technological issues.
Terena and the network community have proposed a dedicated H2020 project requesting funds to further develop eduGAIN though it is unlikely that a production service will be generally available within 3 years. Federated identity management is also gaining traction in business sectors so it will be essential for eduGAIN to ensure it can engage with commercial identity providers and service providers and avoid isolating the research and education community.

[1] http://zenodo.org/record/7592
[2] http://geant3.archive.geant.net/Media_Centre/News/Pages/CERN_first_user_of_GEANT_terabit_network.aspx
[3] http://www.telekom-icss.com/pressreleases/161616
[4] https://indico.cern.ch/event/301888/

Cloud services

The stimulus provided by publicly funded research organisations (CERN, EMBL and ESA) via the Helix Nebula initiative has led to the creation of its first product: the *Helix Nebula Marketplace* (HNX[5]) with considerable investment by the commercial cloud service providers and the active engagement of the publicly funded European e-Infrastructures GÉANT and EGI.

HNX is still in its early phases but it has shown the value of such public-private partnerships where, driven by the common procurement needs of research organisations, Europe's IT industry has shown it is willing to invest. The demand-side has continued its deployment testing and procurement activities with the Helix Nebula Marketplace (HNX) product during 2014. With approximately 30% of its membership being SMEs, the Helix Nebula initiative is providing a channel by which innovative cloud service companies can work with major IT companies and public research organisations. Helix Nebula has published an update to the strategic plan[6] that launched the initiative three years ago and a roadmap for future developments[7]. The intention is to spread the scope of the Helix Nebula initiative to become a forum between the supply-side and the demand-side where issues of common interest (such as procurement models, contractual frameworks, service platforms etc.) can be addressed. The interest in the use of commercial cloud services and the hybrid cloud model continues to grow among the EIROforum members. ESO, ESRF and ILL have joined the Helix Nebula initiative. ILL and ESRF are investigating how the initiative could help the photon and neutron science community to provision big data analysis services to treat the increasing data volumes created by improved sources, detectors, and software. Building on the success of EC projects including Pandata[8], such services would be offered via a common platform to users from domains including physics, life-sciences, chemistry, material sciences and cultural heritage that make use of many of Europe's synchrotrons and free-electron lasers.

Data Services and links to existing European e-Infrastructures

Data services have provided an opportunity for the European e-Infrastructures to work more closely together.

The ESFRI cluster projects (BioMedBridges, CRISP, DASISH and ENVRI) have jointly produced a document that identifies common challenges in data management, sharing and integration across scientific disciplines[9].

EGI has produced a document[10] stating its support to an open science commons. EUDAT has pursued the implementation of several data management services and a first public release has been made in 2014. A number of these services depend on technologies that have already been developed and deployed by the High Energy Physics community for large-scale production usage, notably Invenio[11] for the simple data store and FTS3[12] for transferring large data collections. PRACE is considering whether it can make use of data services from other e-Infrastructure providers to offer data storage beyond the current short-term staging for users of the tier-0 machines that are selected through its calls for proposals.

[5] http://hnx.helix-nebula.eu

[6] http://www.helix-nebula.eu/publications/deliverables/d92-strategic-plan-scientific-cloud-computing-infrastructure-europe-three

[7] http://www.helix-nebula.eu/publications/deliverables/d91-roadmap-of-future-developments

[8] http://pan-data.eu/PaNdataODI

[9] http://dx.doi.org/10.5281/zenodo.7636

[10] http://www.egi.eu/news-and-media/publications/OpenScienceCommons_v2.pdf

[11] http://invenio-software.org/

[12] https://webfts.cern.ch/

The Zenodo digital repository powered by Invenio and operated by CERN as part of the EC co-funded OpenAIRE series of projects, has been extended with important features that greatly improve data sharing. In particular, Zenodo now offers persistent identifiers for data objects so datasets can be cited and includes interfaces permitting metadata can be harvested. EMBL-EBI is using the Zenodo service as a means of storing material from the BioMedBridges FP7 project. The earth observation community is using the service to manage data and software for Envisat ASAR data. A collaboration involving ESRF is using Zenodo to manage material for the development of the Vispy high-performance interactive 2D/3D data visualization software[13]. EFDA-JET collaborators make use of Zenodo as a document repository for publications.
The impressive rate at which the *'long-tail of science'* have adopted data services such as Zenodo that offer reliable, long-term data storage shows there is a clear need and willingness to share data. The EC has recently extended the funding of OpenAIRE until 2018.

Distributed Infrastructure

Distributed infrastructures are now benefitting from advances in cloud technology. Many of the grid sites within EGI are now making use of cloud software suites, such as OpenStack and OpenNebula, to manage their resources and offer high throughput computing services. The rate of adoption of open source cloud software stacks, in particular OpenStack, means they are rapidly becoming de-facto standards in both the enterprise and public sector domains. EGI has continued to develop the prototype EGI Fed Cloud[14] which has been tested with several user communities and integrated into the Helix Nebula Marketplace (HNX) product described above. The integration has been tested with flagship applications from CERN and ESA. The experience gained from this work has shown that integrating publicly funded e-Infrastructures with commercial services to provide a combined platform on which to build new services has a clear value for the users. Two developments would provide the basis for further integration:

- a common catalogue of services including services provided by all suppliers present in the exchange,
- a federated identity management system offering a single sign-on facility to access services across all suppliers.

Software services and tools

Dropbox-like services that allow researchers to integrate e-Infrastructures with their everyday activities and personal devices have flourished over the last year. A number of the established services, both commercial and publicly funded, now offer dropbox integration to simplify the import and export of data.

Building the data continuum

It is important that underlying datasets, and the software used to analyze them, are treated as first-class objects that can be cited so the researchers that produce them are given credit for their work. The CERN-hosted repository Zenodo has been extended to allow source code from the popular software development site GitHub to be preserved and cited.

[13] http://vispy.org/

[14] https://www.egi.eu/export/sites/egi/infrastructure/cloud/fedcloudflyer2.pdf

Funding models and the engagement of funding agencies

The provision of services via supplier funded resources is the foundation of national and European e-Infrastructures. The concept of an exchange as a marketplace[15] with the ability for users to choose from a range of services and suppliers can offer a practical implementation of the concept of an e-Infrastructure commons. The exchange approach coupled with the pay-per-use model, as championed by Helix Nebula, is being considered by a number of e-Infrastructures.

Public procurement of cloud services is promoted as a means of encouraging innovation and reducing the time to market for new products and services. Building a network of public research organisations that can procure e-Infrastructure services will attract the interest of commercial service suppliers as well as national, regional and European funding agencies.. The EC's co-funding for pre-commercial procurement of innovative cloud services in H2020 is a positive development. The majority of this funding will be directed to commercial service providers and the approach has the advantage of permitting the procuring organisations to choose which services and suppliers receive these funds and thus represents a change to the established funding model for IT services.

The hybrid public-private model for e-Infrastructures is perceived as both a threat and an opportunity by service providers. Commercial service providers see the market potential of selling their services to the public research sector but fear their investments in pay-per-use services will be undermined by publicly funded service providers offering similar services at no cost to the user. Publicly funded service providers fear that commercial services will replace their existing services operated on their in-house infrastructure.

Through the work of initiatives such as Helix Nebula it has become clear that it is essential to separate the roles of *end-user* (the researcher making use of the services) and *customer* (the organisation sponsoring the consumption of the services by the end-user) and ensure that services, both commercial and publicly funded, are offered as *free at the point of use*.

Publicly funded e-Infrastructures are investigating a new brokering role to facilitate access to commercial services for their user-base. A key attraction of the brokerage role is that it offers a new revenue stream and while these business model innovations should be encouraged it is important that competition between brokers does not lead to renewed fragmentation of the e-Infrastructure commons. An aspect of brokerage which is under-estimated by the publicly funded e-Infrastructures is the necessary financial engagement and liability. Experienced financial brokers from utility markets could help ensure the good governance of the exchange and take on-board some of the financial risks from users and suppliers to accelerate the expansion of the market.

Policy bodies, such as e-IRG and ESFRI, are working more closely together for the benefit of e-Infrastructures and the EC's work programme for 2014-2015 made cooperation between e-Infrastructures and ESFRI Research Infrastructures a key theme.

The EC has taken steps to ensure funding for GÉANT over the full duration of H2020 by introducing 'Framework Partnership Agreements' (FGA). The FPA model represents a more long-term engagement that could encourage Research Infrastructures to integrate them into their computing models. If the FPA approach for networks is successful its adoption by other pan-European e-Infrastructures could establish the basis for the European Research Area's digital commons and lead towards Science 2.0[17].

[15] https://cdsweb.cern.ch/record/1709709/files/HelixNebula-MISC-2014-001.pdf

[17] http://ec.europa.eu/research/consultations/science-2.0/background.pdf

Governance

The on-going plans for restructuring collaborative activities through TERENA and DANTE will contribute to a rationalisation of the publicly funded networking domain. If such a restructuring proves successful the model could bring additional benefits by being extended beyond the network domain to encompass a wider range of publicly funded e-Infrastructures.

EGI is preparing a new governance model while the EC has highlighted the value of public-private partnerships through structures such as Contractual Public-Private Partnerships (cPPP).

GÉANT, EGI and EUDAT have submitted a proposal to the EC requesting funds to organise a joint annual user forum. In all cases the focus remains on the service providers while the role of the user remains purely consultative.

The role of RDA

The Research Data Alliance (RDA) plenary events have grown to become popular fora for data practitioners, but not necessarily the end-users, from many research disciplines. RDA now has over 1,600 individual members from over 70 countries. RDA Europe has received additional funding from the EC to cover the period 2014-2016[18]. EIROforum has joined RDA as an Organisational Member and CERN will host the second RDA Europe science workshop in 2015 with a group of scientists from a variety of disciplines selected by RDA Europe. The participants are expected to give their opinions on RDA's activities and the value of the results of the first batch of RDA working groups (including Data Foundation and Terminology[19], Data Type Registries[20], PID Information Types[21] and Practical Policy[22]).

At the 4th RDA plenary, held recently in Amsterdam, the Interest Groups were seen as being just as valuable as the Working Groups, and there has been at least one example where an Interest Group has led to a H2020 project proposal (on harmonization of standards for certifying digital repositories: CTRUST). A new Interest Group is currently being setup in the area of Reproducibility of data and results and this is seen as potentially important as reproducibility is increasingly required by funding agencies.

Our initial assessment is that RDA as a forum can help establish new collaborations and provide input on policy issues. The potential impact on the implementation and deployment of data services will take longer and so RDA should manage the expectations of its stakeholders. RDA is a new organisation that is still developing its governance model and it is important that the decision making processes becomes more transparent and inclusive.

RDA has produced a draft report[23] analysing the data practises across a large number of research disciplines and makes a number of recommendations. The document states that researchers do not trust companies with their data. This lack of trust is probably limited to data services in the later phases of the research lifecycle because many of the instruments researchers use to acquire data are commercially produced along with the quasi-totality of IT equipment on which it is subsequently transported, stored and analysed. Commercial data services are relatively new and do need to earn the trust of researchers even if most researchers have no issue in using commercial data services for many aspects of their private lives.

[18] http://cordis.europa.eu/project/rcn/191517_en.html

[19] https://www.rd-alliance.org/group/data-foundation-and-terminology-wg.html

[20] https://www.rd-alliance.org/group/data-type-registries-wg.html

[21] https://www.rd-alliance.org/group/pid-information-types-wg.html

[22] https://rd-alliance.org/group/practical-policy-wg.html

[23] https://rd-alliance.org/sites/default/files/Survey%20of%20data%20mangement%20needs.docx

It is essential that the research communities have a more collaborative and open attitude to working with the commercial sector. Companies can provide expertise and innovative data services and will be prepared to invest in developments if there is the opportunity to sell services to the research communities. Trust can be encouraged by imposing standards in the public research sector and by defining a clear and European-wide legal basis for the management of data.

RDA, as a forum bringing together representatives from many research disciplines, should recognise this situation and work with the research communities to evolve the attitude towards commercial data services as well as the funding models for research. This will be essential to encourage other stakeholders, including the commercial IT sector, to invest in data management services that can serve research communities.

CERN openlab

A positive example of collaboration between research communities and the IT industry is the joint developments that EIROforum members have undertaken as part of the CERN openlab[24].

CERN openlab is a unique public-private partnership between CERN and leading IT companies. It was created in 2001 in support of the ambitious computing and data management goals set by the construction of the Large Hadron Collider (LHC) and detectors. The underlying principle behind the successful history of CERN openlab is the mutual benefit that CERN and the industrial partners derive from the collaboration. CERN gets early access to new technologies and the companies have the unique possibility of testing their upcoming products on CERN's very challenging IT infrastructure.

In order to define the long-term technological context in which joint research activities can take place in the next five years, CERN, ESA, EMBL-EBI, ESRF, ILL and the Human Brain Project, have set ambitious challenges covering the most crucial needs of their IT infrastructures as documented in the whitepaper on *Future IT Challenges in Scientific Research*[25]. The identified areas, or challenges, are data acquisition, computing platforms, data storage architectures, compute provisioning and management, networks and communication, and data analytics.

The whitepaper is the result of many discussions among IT experts and scientists and is a first step in a strategy to address the big data challenges in scientific research. It forms the basis for the fifth phase of CERN openlab which will start in January 2015 engaging the European laboratories, international scientific projects and leading IT companies for 3 years.

CERN openlab and the projects it will support represent a prime example of a self-supporting public-private partnership which could be expanded via the Horizon 2020 work-programme to accelerate the innovation cycle and increase the impact of scientific research on the economy in Europe.

Research Accelerator Hubs

Two prototype Research Accelerator Hubs (ReAcH) were described in the *e-Infrastructure for the 21st Century* document. The development of these prototypes by EMBL-EBI and CERN has continued. EMBL-EBI has extended the range of services available via its Embassy-cloud to

[24] http://openlab.cern.ch

[25] http://press.web.cern.ch/press-releases/2014/05/cern-openlab-publishes-whitepaper-future-it-challenges-scientific-research

tenants for commercial use and the subscription scheme, terms and conditions as well as the service quality criteria have been refined. The first commercial customer is now using the service, which continues to be refined through this use and with other research collaborations. CERN has grown the installed capacity at its data centres in Geneva and Budapest (Wigner). The Zenodo digital repository has been adopted by many research communities and several commercial journal publishers have expressed interest in using the service as the basis of their data management plans. CERN currently does not have the equivalent of EMBL's Enterprise Management Technology Transfer (EMBLEM[26]) affiliate and commercial arm but is exploring the most appropriate mechanism that would allow CERN to offer services.

[26] http://www.embl-em.de/

Conclusions

Advances have been made in the implementation of a number of e-Infrastructure services. Some of these technical developments have been deployed in production environments and more will follow. Non-technical aspects including e-Infrastructure integration, governance and funding models take longer to address but there are indications that there is a willingness to consider change by several e-Infrastructure service operators and funding agencies.

Priorities

The *e-Infrastructure for the 21st century* document stated:

> *In order that such an effort be sustainable and permit maximum flexibility across domains, as well as being able to fulfil the goals of working together with industry and key global players, it is essential that the future service infrastructure and tools be fully based on open standards, open software, and promote open access to the data.*

The following aspects are considered high priority meriting closer attention and more determined action if the e-Infrastructure for the 21st century vision is to be achieved.

Accelerate the pace of integration of e-Infrastructure services

Higher-level services tailored to the needs of the user increase the adoption of e-Infrastructures by user communities representing *big science* and the *long-tail of science* alike. Such higher-level services rely on combing multiple services and datasets from a variety of sources. Encouraging existing e-Infrastructure service providers to work more closely together will facilitate the creation of innovative higher-level services.

Open access to data

As the Scholarly Publishing and Academic Resources Coalition (SPARC[27]) states:
"Funders invest in research in order to accelerate the pace of scientific discovery, encourage innovation, enrich education, and stimulate the economy – to improve the public good. They recognize that broad access to the results of research is an essential component of the research process itself. Research advances only through sharing of results, and the value of an investment in research is only maximized through wide use of its results."

The potential benefits of providing open access to data are still under-estimated by many research communities. The Horizon 2020 open access data pilot will encourage projects to consider open access to data and will also lead to the realisation that it represents an additional cost which has not yet been taken into account.

Long term data preservation

Data is recognised as the truly valuable asset of a data-driven economy and hence must be preserved. Today the process of data preservation is not clearly understood by the majority of research communities but its value will become more apparent as the quantity and rate at which scientific data is produced increases.

[27] http://www.sparc.arl.org/

Data analysis platforms

Data Services and open access repositories have demonstrated their value through their rapid adoption by the *long-tail of science*. A range of additional services are being added to these repositories which improve the potential of the stored data by establishing and identifying relationships between objects as well as measuring their usage and impact. Such services bring true added value but require significant compute capacity.

In parallel the adoption of cloud technology is rationalising the organisation of data centres, increasing their efficiency and the cost effectiveness of the installed hardware base.

Linking data services with cloud computing capacity to offer on-demand data analysis platforms will present users with a comprehensive environment supporting the full lifecycle of science workflows. This will require significant investment in core software components capable of scaling to handle big data challenges.

Evaluate new funding models

The innovation potential of e-Infrastructures is under –utilised. Introducing mechanisms that can increase the pace of technology transfer from research to the economy and society as a whole will ensure better exploitation. Promoting the innovation potential through channels such as public-private partnerships will attract other funding sources and encourage entrepreneurship.

Financial incentive schemes, such as innovation vouchers and joint procurement co-funding, are promising approaches that will encourage a wider sharing of the investments necessary to increase technology transfer and the uptake of e-Infrastructure services. The EC has recently funded a support action (PICSE[28]) led by CERN to establish a cloud services procurement network of publicly funded research organisations across Europe which will build on the activities of the Helix Nebula initiative.

[28] http://www.helix-nebula.eu/picse/home

Actions

Taking into account the priorities identified above, the e-Infrastructure commons exchange, defined as an open environment where researchers can flexibly discover and choose the services and service providers from the public and private sectors that best meets their needs, has the potential to advance the vision of the e-Infrastructure for the 21st century in a holistic and inclusive manner:

> *"Researchers across Europe are looking for cost effective and sustainable IT services that can be combined to accelerate their work and increase its impact. Europe has a wealth of public and private sector service providers and when brought together they can create a ground-breaking open platform for innovation.*[15]*"*

Specific action points to implement the exchange and address these priorities are:

- Integrate publicly funded e-Infrastructures with commercial services to provide a combined platform on which to build new services. Integration should start with a common catalogue of services and a federated identity management system offering a single sign-on facility to access services across all suppliers.
- Define a lightweight governance model for the exchange to ensure it is operated in a transparent and inclusive manner for European service providers and users from both the public and private sectors.
- Establish a network of public research organisations to implement joint procurement actions with the goal of augmenting the capacity and functionality of the platform. Such procurement actions would be of interest to national, regional and European funding agencies.
- The service providers should jointly develop a communication plan to promote the existence of the exchange to audiences including research communities, policy makers and funding agencies.
- Engage funding agencies to sponsor the activities of researchers and SMEs to use and develop new services on the platform.