

ESFRI project requirements
for
Pan-European e-infrastructure
resources and facilities

European E-Infrastructure Forum

April 28th 2010

Table of contents

Table of contents	2
Preface	3
Methodology of the fact-finding process	4
Overview of the ESFRI projects consulted and their requirements	5
Social sciences and humanities	5
Biological and medical sciences	6
Environmental sciences	7
Energy	8
Material and analytical facilities & physical and engineering sciences	9
Overview of the Pan-European e-infrastructures as represented by the EEF	11
Networking	11
Grid-infrastructures	12
High Performance Computing	13
Data management services	14
EGEE/EGI	14
Existing European scientific data infrastructure projects	15
Conclusions from the initial requirements analysis	18
Single sign-on	18
Virtual Organisations	19
Persistent storage	19
User support	19
Training and consultancy	20
Web-service interfaces	21
Workflows	21
Global scope	21
Integration with cloud systems and volunteer desk-top systems	22
Next Steps	23
Members of the EEF	24

Preface

The European Strategy Forum on Research Infrastructures (ESFRI) has issued a list of major Pan-European facilities and services and subsequent updates¹ thereto, which are unique to Europe due to their sheer size or cost involved in their establishment. The e-IRG² has issued clear roadmaps describing present and future Pan-European e-infrastructures, that due to their scale require long term planning, and cross-border operations beyond a single administrative domain, but has a fairly short technological refreshment cycle of about three years.

The ESFRI-projects (as they are commonly referred to) do require such e-infrastructures to professionally and efficiently conduct science with or on the very facilities they are concerned with. But the establishment of these projects, in particular the early starters, is taking place rather independent of the establishment of the e-infrastructures that are being developed at pretty much the same time and proper inter-reference and certainly close interoperation is yet lacking, irrespective incidental co-operations.

This report³ and the next steps proposed address the presently known and foreseeable requirements for European scale e-infrastructural resources or facilities by the ESFRI-projects and the services and resources that the e-infrastructure community can offer to the ESFRI-projects at the European level. Also a template/checklist is developed that can be used to structure future ESFRI-project-proposals in order to avoid duplicating effort. A key objective of this European E-infrastructures Forum⁴ (EEF) initiative is to achieve a seamless interoperation of leading e-Infrastructures for the scientific communities that make use of the research infrastructures in the ESFRI roadmap, tailored to their needs where possible or required, and optimise the return on investments by the ESFRI-projects and user communities. This could also help the EC in the evaluation of future ESFRI-projects in the application phase, regarding the use of already available pan-European e-infrastructure resources and facilities.

The EEF is a forum for the Pan-European e-infrastructures providers in the areas of High Performance Computing, Networking, secure data-storage and services and the European Grid-infrastructure. The interest of the EEF in this work is to tailor our services to the ESFRI-projects' needs, to avoid parallel e-infrastructures being set up without connection to existing or planned investments and to have links established from the EEF to the ESFRI-projects to help the e-infrastructure providers and policy makers to provide the best services at the best conditions to the European flagship research facilities. The EEF has, through a series of meetings and a questionnaire, gathered a set of e-infrastructure requirements from the ESFRI projects. Some 28 ESFRI projects were consulted as part of the requirements gathering process. Based on the information received the EEF has made an initial analysis which is recorded in this report. The implications and opportunities for the European e-infrastructures have also been analysed and included in this report. The EEF sees this activity as a first iteration in an on-going dialog that is required between e-infrastructure and ESFRI representatives and foresees a number of steps that will continue this process.

Geneva, 28th April 2010

¹ European Roadmap for Research Infrastructures Roadmap 2008, ISBN 978-92-79-10117-5

² <http://www.e-irg.eu/>

³ Copies of this report are available online from <http://www.einfrastructure-forum.eu/>

⁴ <http://www.einfrastructure-forum.eu/>

Methodology of the fact-finding process

The EEF members first came together at the ICT08 event in Lyon, France, in November 2008 and in early 2009 started discussing and sharing information about contacts between the e-infrastructure and the ESFRI projects and their perceived needs. This led the EEF to formalise its work-plan and a series of sessions were organised at the EGEE09 conference in Barcelona September 2009 where 11 ESFRI projects were invited to attend and present their requirements for their use of e-infrastructures. During this meeting it was agreed that a common set of themes was emerging from the expressed requirements and that it would be worthwhile investigating these further. As a consequence EEF developed a questionnaire which it has used to collect information on the requirements from the ESFRI projects.

In addition to the presentations at EGEE09 and the responses to the questionnaire, EEF has taken into account the information presented at a number of other events involving e-infrastructure and ESFRI project representatives:

- NEERI09⁵ workshop for social sciences and humanities;
- the European association of national Research Facilities laboratories (ERF) workshop⁶;
- a series of workshops organised by the European Commission for the biological and medical (BMS)⁷, social sciences & humanities (SSH)⁸ and environmental sciences (ENV)⁹;
- the 7th e-infrastructure concentration meeting¹⁰;
- the findings of the Sixth European Conference on Research Infrastructures (ECRI2010¹¹) conference provided material on the potential relationship between e-infrastructures and Research Infrastructures;
- a report¹² from the OSIRIS project suggested major challenges concerning ICT Research Infrastructures;
- the EGEE User Forum¹³ offered the occasion to interact with a further 9 ESFRI projects and present to them the initial findings of this study and gather their initial feedback.

Based on the information received the EEF has made an initial analysis which is recorded in this report. The implications and opportunities for the European e-infrastructures has also been analysed and included in this report. The EEF sees this activity as a first iteration in a prolonged dialog that is required between e-infrastructure and ESFRI representatives and foresees a number of steps that will continue this process.

⁵ NEERI09, Helsinki on 1-2 October 2009 <http://www.csc.fi/neeri09>

⁶ Future Access to European Research Infrastructures: Benefits to Academia, Industry and Society, Lund, 27th of October 2009, <http://www.europeanresearchfacilities.eu/>

⁷ Workshop on ICT and e-infrastructure needs for European Research Infrastructures in the field of Life Sciences (BMS), Brussels, 16 December 2009

⁸ workshop on Common Needs and Common Solutions for the ESFRI research infrastructures for the Social Sciences and Humanities (SSH), Brussels, 20th January 2010

⁹ Workshop on common ICT and e-infrastructure needs for the ESFRI Research Infrastructures in the field of Environmental Sciences (ENV), Brussels, 18th March 2010

¹⁰ 7th Concertation Meeting, held in Brussels on 12-15 October 2009

¹¹ Sixth European Conference on Research Infrastructures, ECRI2010 (<http://www.ecri2010.es/en>), Barcelona, March 23rd-24th, 2010

¹² FP7-ICT-248295/IMCS/R/CO/D2.1

¹³ 5th EGEE User Forum held in collaboration with EGI and NDGF in Uppsala, Sweden, April 12-15, 2010 <http://egee-uf5.eu-egee.org/>

Overview of the ESFRI projects consulted and their requirements

The EEF has tried to establish contact with as many ESFRI projects as possible from all the sectors. A total of 28 ESFRI projects were consulted during the preparation of this report. The EEF was most successful in its interaction with the BMS, SSH and ENV sector ESFRI projects. The interaction with these sectors has been facilitated by the actions of the EC to organise themed workshops. A possible explanation for the enthusiasm of these projects to discuss their e-infrastructure issues could be due to the distributed nature of their proposed Research Infrastructures and hence the novel computing models that will be required to address their needs. Below each sector is examined and their common requirements highlighted.

Social sciences and humanities

The Social Sciences and Humanities contribute actively to and are necessary instruments for our profound understanding of the cultural, social, political and economic life in Europe as well as for the process of European cohesion and bringing about changes. In practice these disciplines make significant contributions to important areas like strengthening employment, modernising our social welfare and education systems, and securing economic reform and social cohesion as part of a knowledge- based economy.

In Social Sciences and Humanities five ESFRI projects are listed:

CLARIN¹⁴ is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available for the whole European Humanities (and Social Sciences) community.

ESS¹⁵ (The European Social Survey) is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations.

DARIAH¹⁶ (Digital Research Infrastructure for the Arts and Humanities) aims to enhance and support digitally-enabled research across the humanities and arts, as well as to develop and maintain an infrastructure in support of ICT-based research practices and to share expertise and tools for the creation, curation, preservation, access and dissemination of data.

SHARE¹⁷ (Survey of Health, Aging and Retirement in Europe) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks.

¹⁴ <http://www.mpi.nl/clarin/>

¹⁵ <http://www.europeansocialsurvey.org>

¹⁶ <http://www.dariah.eu>

¹⁷ <http://www.share-project.org>

CESSDA¹⁸ is an umbrella organisation for social science data archives across Europe. The CESSDA Catalogue enables users to locate datasets, as well as questions or variables within datasets, stored at CESSDA archives throughout Europe.

Within the Social Sciences and Humanities community several areas of commonality have been identified. Data archiving and curation is a common need for several of the ESFRI projects. To enable this they identify a requirement for a flexible repository system and a system to provide Persistent IDentifiers (PIDs) which together will provide the basis for data storage/archiving and management. The sensitive nature of the data to be stored leads to a need for a fine-grained Authentication and Authorization system, it also is imperative that such a system provides Single Sign On (SSO) functionality. The ability to use grid/cloud compute facilities for the processing of the stored data is also foreseen in some projects. Finally education and training covering the e-infrastructures and associated technologies was clearly requested by the SSH community.

Biological and medical sciences

The biological and medical sciences (BMS) projects within ESFRI cover a range of disciplines with a general focus on health and drug development. There is also some work in the area of Marine biology. Developments in the field and the application of ICT are leading to huge increases in the amounts of data available. These in turn require access to well structured databases, which should be broadly accessible. Recognizing that the value of the data being collected far exceeds the costs of storing and accessing it, the development of distributed infrastructure to store, curate and provide access globally is a key part of the planning.

In order to organize user-friendly data access, a major investment in computer infrastructure and storage is envisaged, along with the development of appropriate standards and ontologies. Developments in imaging are likely to give rise to significant increases in data volumes, which will place new demands on computing, networking and storage. The main demands on e-infrastructures are seen in terms of storage, grids, networks and general computing, with high performance computing, essentially large scale parallel computing, seen as being of lesser importance.

The following BMS projects have been consulted as part of the requirements gathering process:

ELIXIR¹⁹: a secure, evolving platform for biological data collection, storage and management, consisting of an interlinked set of core and specialist resources.

BBMRI²⁰: a distributed pan-European infrastructure of bio-banks, incorporating biomolecular research tools and biocomputational tools.

ECRIN²¹: supports multi-national clinical research trials in Europe by connecting together nationally coordinated networks of clinical research networks and clinical trials units.

¹⁸ <http://www.cessda.org>

¹⁹ <http://www.ebi.ac.uk>

²⁰ <http://www.bbmri.eu>

²¹ <http://www.ecrin.org>

EMBRC²²: will provide an infrastructure connecting the main coastal marine laboratories in Europe, to facilitate common research and training.

ERINHA²³: involves the development and cooperation of European bio-safety level 4 laboratories in Europe.

Euro-BioImaging²⁴: is planning the construction and operation of connected facilities, providing access to imaging technologies, covering both biological and medical applications.

Infrafrontier²⁵: is organising infrastructure among 15 European laboratories to provide large-scale phenotyping and archiving of mouse models.

Instruct²⁶: is organising a distributed infrastructure of core and associated centres for integrated structural biology.

EU Openscreen²⁷: The European Infrastructure of Open Screening Platforms for Chemical Biology will be used by European researchers in order to identify compounds affecting new targets.

In addition, two further projects were also consulted:

eNMR²⁸: aims to provide the European biomolecular nuclear magnetic resonance community with a platform for access to appropriate computational methods.

neuGRID²⁹: is planned as a GRID-based facility for the neuroscience community to assist in research on degenerative brain disease.

Environmental sciences

The ESFRI Environmental Science (ENV) projects represent a diverse set of demands in respect of e-infrastructures, including measurement and monitoring facilities, access to analytical facilities such as synchrotrons, as well as large-scale access to unique global facilities and distributed facilities on a pan-European basis.

The overall objectives are to support the sustainable management of the environment by monitoring and measuring major environmental systems. Significant investment is proposed in both fixed and mobile server systems, collecting data from land, sea and air measurements, using fixed and mobile data collection. The ICT challenges associated with the sector include data capture, particularly from sensor networks, the combining, processing and storage of

²² <http://www.embrc.eu>

²³ <http://asso.orpha.net/HBSL>

²⁴ <http://www.eurobioimaging.eu>

²⁵ <http://www.infrafrontier.eu>

²⁶ <http://www.instruct-fp7.eu>

²⁷ <http://www.eu-openscreen.de>

²⁸ <http://www.enmr.eu>

²⁹ <http://www.neugrid.eu>

large and complex data sets. There are also some significant real-time requirements in terms of collecting and processing data.

The following ENV projects have been consulted as part of the requirements gathering process:

EISCAT-3D³⁰: is a planned as a distributed network of incoherent scatter radar, capable of making measurements of the upper atmosphere.

EMSO³¹: is a European Multi Disciplinary Network of seafloor observations, providing permanent monitoring of the deep sea.

EPOS³²: is an integration of existing Plate Observation Systems into a coherent distributed research infrastructure.

EUFAR-COPAL³³: is a proposal for a heavy pay-load, long-endurance aircraft to provide a platform for airborne measurements across a range of disciplines.

EURO-ARGO³⁴: is a proposal to develop the European component of a global ocean observation system.

EUSAAR-I3³⁵: is a project to provide for the integration of atmospheric aerosol properties measured at a distributed network of European ground stations.

EARLINET-ASOS³⁶: is a cooperative activity among operations of Aerosol LIDAR systems across Europe.

IAGOS³⁷: is a project exploiting the routine measurement of atmospheric composition by installing instruments on commercial aircraft.

ICOS³⁸: is a project which plans to integrate terrestrial and atmospheric observations of greenhouse gases into a single dataset.

LifeWatch³⁹: is a network of observations and biological collections brought together in a virtual laboratory to measure biodiversity.

Energy

The ESFRI roadmap stresses the importance of economically competitive, environmentally friendly and sustainable energy resources for European development. A coherent policy for

³⁰ <http://www.eiscat3d.se>

³¹ <http://www.emso-eu.org>

³² <http://www.epos-eu.org>

³³ <http://www.eufar.net>

³⁴ <http://www.euro-argo.eu>

³⁵ <http://www.eusaar.net>

³⁶ <http://www.earlinet.org>

³⁷ <http://www.iagos.org>

³⁸ <http://www.icos-infrastructure.eu>

³⁹ <http://www.lifewatch.eu>

Research Infrastructures is needed to maintain Europe's world leadership in efficient use of energy, in promoting new and renewable forms and in the development of low carbon emission technologies. A Strategic Energy Technology Plan (SET Plan) has been adopted to meet by 2010 the challenging goals of greenhouse gas reduction by 20%, to triple renewable energy consumption up to 20% and to increase the share of appropriate biofuels. The Research Infrastructures in the energy sector listed in the 2008 update of the ESFRI roadmap were all invited to contribute to this plan. The areas covered addressed include carbon dioxide capture and storage, nuclear fission and fusion, wind energy, solar energy, biofuels, ocean/marine energy, hydrogen, and smart energy grids.

Example: Fusion research community, ITM of EFDA and EUFORIA FP7 project

The Integrated Tokamak Modelling group (ITM) of EFDA and the EUFORIA project are building tools for predictive simulations for the ITER and DEMO experiments and contribute to the understanding of results obtained on present tokamak type fusion devices, and in future on ITER and DEMO.

Existing codes applications and modules are coupled into a work-flow using standardized interfaces. Depending on the requirements, the work-flow is executed on local resources, on a compute grid or on remote High Performance Computers. Standardized ways are needed for code coupling and execution, for monitoring and for data management.

Data sources are experiments (existing tokomaks include JET and MAST in the UK, ASDEX upgrade in Germany, Tore Supra in France, etc.) and computer simulations. Data sizes vary from MBytes to TBytes. Compute resource requirements range from small Linux clusters to supercomputers. Good network connectivity will be required for data transfers between experimental facilities and the sources of simulation data, for visualization of remote data, and for bulk transfers of accumulated data.

The EUFORIA gateway computer, operational for the last two years, has been interfaced to both the EGEE grid and DEISA supercomputer resources as part of a pilot project. For the European Fusion Research Community a 100 TFlop/s computer (HPC-FF at FZJ, Juelich) was put into operation in 2009, and a PetaFlop/s computer shall start operation at IFERC (Rokkasho, Japan) in January 2012. ITER operation is currently scheduled to start around 2020. Data lifetimes should cover present simulation results over the lifetime of ITER and DEMO (40 years).

Material and analytical facilities & physical and engineering sciences

The projects of the Material and analytical facilities have been grouped with those of Physical and engineering sciences due the closely related nature of the Research Infrastructures concerned. The ESFRI roadmap highlights that the development of new materials contributes to all areas of human activity from energy generation and storage through to medical implants and computer components. For the physical and engineering sciences the roadmap notes that the facilities have become much larger, technically more complicated and that they drive the development of new technologies and new ways of working. The following projects have been consulted as part of the requirements gathering process:

European XFEL⁴⁰: construction of the Hard X-ray Free Electron Laser in Hamburg has already started.

CTA⁴¹: the Cherenkov Telescope Array for ground-based high-energy Gamma-ray astronomy.

FAIR⁴²: the Facility for Antiproton and Ion Research which is being built in Darmstadt.

SKA⁴³: the Square Kilometre Array for radio-astronomy to be built in the Southern hemisphere.

The requirements of such Research Infrastructures are similar to those from the energy sector, with a definite need for high performance modelling, simulation and processing facilities and reliable high-speed network connections for data acquisition as well as its distribution to a large number of researchers around the world. Their data also needs archiving facilities and open access for the user communities where they stress the importance of a single sign-on service.

⁴⁰ <http://www.xfel.eu/>

⁴¹ <http://www.cta-observatory.org>

⁴² <http://www.gsi.de/fair>

⁴³ <http://www.skatelescope.org>

Overview of the Pan-European e-infrastructures as represented by the EEF

This section gives an overview of the existing pan-European e-infrastructures and their plans for the future. This ecosystem of e-infrastructures has been developed over many years as a joint under-taking between national funding bodies and the European Commission. The e-infrastructure comprises high-speed networking, high-capacity grid systems and specialised high-performance computing centres and servers a significant proportion of Europe's research community.

Networking

GÉANT is an advanced pan-European backbone network that interconnects National Research and Education Networks (NRENs) across Europe and provides worldwide connectivity. With an estimated 40 million research and education users in 40 countries across the continent, the “network of networks” created by GÉANT and the NRENs – known as the GÉANT Service Area – offers unrivalled geographical coverage. The GÉANT Service Area has European links totalling more than 50,000 km in length, while its interconnections with networks in other world regions extend its coverage across the globe. GÉANT's vast geographical reach, along with its high performance, high bandwidth and high data-transfer speeds all enable its European users to share huge quantities of data and collaborate effectively both with each other and with their peers throughout the world, regardless of distance or location. Links to networks in other world regions include extensive connectivity to North America as well as to TEIN (Asia-Pacific), ALICE (Latin America), EUMEDCONNECT (Mediterranean), ORIENT (China) and UbuntuNet Alliance (Southern African). GÉANT is also working towards connecting to Central Asia (CAREN) and South Eastern Africa.

GÉANT's innovative technology opens up new connectivity service possibilities. For example, it is possible to reserve paths and capacity across the network that appear to the user as a dedicated private facility: a “virtual” private network (VPN). These specialised “point-to-point” connections provide guaranteed bandwidth and performance for specific user communities without the cost and difficulty of building and managing an actual private network. A portfolio of services has been developed to better describe the technical capabilities available to support the needs of projects.

As well as connectivity services, GÉANT offers network performance services and end-user applications aimed at optimising the network performance, providing remote-access options and, to improve user access and the ability of users to exploit the network for their own requirements. These additional services all serve to further facilitate global collaboration between researchers, educators and innovators.

GÉANT's key objective is to deliver real value and benefit to society by enabling research communities across Europe (and the world) to transform the way they collaborate on ground-breaking research. It achieves this through:

- Operating and expanding the European backbone network, interconnecting NRENs through high-bandwidth links.
- Developing and supporting the GÉANT Service Area through a portfolio of advanced multi-domain connectivity and network support service options, and a range of end-user application services to ensure seamless network performance.
- Pursuing initiatives targeted at closing the “digital divide” of research and education networking in Europe and investigating emerging technologies that will help shape the future Internet.

The GÉANT network is the core element of the GÉANT project, now in its third term and co-funded by the European Commission under the EU’s Seventh Research and Development Framework Programme. The project partners are 32 European National Research and Education Networks, DANTE, the organisation that manages the GÉANT network and project on behalf of the partners, and TERENA, the Association of National Research and Education Networks in Europe.

Grid-infrastructures

Europe’s largest computing grid for publicly funded research is Enabling Grids for E-science (EGEE). The project provides an e-Research platform for high-throughput data analysis to the European research community and their international collaborators, representing over 17,000 users across 160 projects. With a heritage stretching back over nearly a decade, EGEE-III (and its preceding projects EGEE-II, EGEE and the European Data Grid, EDG) is co-funded by the European Commission to implement, deploy and maintain a distributed computing infrastructure to support researchers in many scientific domains.

Grid technology is a system for distributed storage and processing of data, providing location-independent access to computing resources. Through a ‘grid’, internationally distributed users have access to a fully virtualised system of processing and storage elements that allow single-step access to large-scale resources on demand. The components of a grid are both physical and virtual. They are a service built on top of high-capacity internet connections—for instance, EGEE uses the GÉANT network. The network connects computing nodes (collections of processing cores) from sites or ‘resource centres’. A software stack, known as ‘middleware’ sits between the hardware and the software (i.e. users applications) integrating the systems. The people who use grids are organised in ‘Virtual Organisations’, research collaborations that are often geographically distributed, but connected technologically.

While the technology was unheard of 10 years ago, grid computing is now entrusted with managing the data for the Large Hadron Collider, located in Geneva at CERN—the world’s largest scientific experiment (or more accurately, collection of experiments) built to investigate the fundamental building blocks of matter. The LHC is now fully online and the experiments will produce up to 15 petabytes of data per year (roughly 3 million DVDs or 20,000 years of music in MP3 format). This data must be securely accessed and processed by sites all over the world.

Publicly-funded grid computing projects, such as Enabling Grids for E-science in Europe and Open Science Grid in the United States, originally sought to respond to the data access and processing requirements of high energy physicists. Today however, due to the success of grid

computing as a framework for collaborative work, these projects support research in a range of disciplines: from astronomy to finance, and humanities to epidemiology.

In 2010 this infrastructure supports world class science in over 50 countries, consisting of about 300 sites, encompassing more than 150,000 processors, 25 petabytes of disk storage and 40 petabytes of long-term tape storage—enough to store 400 million four-drawer filing cabinets full of text. This infrastructure is available continuously, 24-hours a day, and supports over 330,000 “jobs” (or executed computer programs) a day. The research network connecting together these sites and the distributed user community sustains transfer speeds of over 900MB/s each day—sending the equivalent of two entire CDs of data every second. EGEE supports a number of operational monitoring and accounting tools. Such information contributes to the overall health of the infrastructure by reflecting its performance and identifying room for improvement to ensure a high-quality of service to the end users.

The EGEE project will come to a close at the end of April 2010. A new organizational model, implemented by the European Grid Initiative (EGI), will take over, and ensure the sustainability of the European grid computing infrastructure. EGI brings together National Grid Initiatives from more than 20 countries in Europe. The same tools and services will be available to users of the infrastructure, but under the management of the EGI. To ensure that its infrastructure is usable and practical, EGEE offers support in many forms to its community. These supporting services will continue under EGI.

High Performance Computing

Driven by the dramatic progress in information and communication technology Computational Science and Engineering has evolved into a key instrument for research and development, now known as the third methodology and considered to be of equal importance to theory and experiment. In many application areas such as climate research, earth science, nanotechnology, computational chemistry, high-energy physics, nuclear fusion, and life sciences computation is the essential method for achieving high-quality results. To remain internationally competitive, European scientists and engineers must be provided with access to supercomputer systems of the highest performance class provided on a European level, and embedded into an ecosystem of national and regional HPC services to respond both to capability and capacity computing needs.

PRACE, the Partnership for Advanced Computing in Europe, is an ESFRI-listed Research Infrastructure that is now in its implementation phase. It will consist of a limited number of world-class Tier-0 centres in a single infrastructure that forms the European layer of the HPC ecosystem. Access to the infrastructure will be granted through a single European Peer Review system based on scientific merit, starting in summer 2010 with the first Tier-0 system installed at the German GCS site Juelich.

DEISA is a consortium of the most powerful supercomputer centres in Europe, operating supercomputers in a distributed but integrated HPC infrastructure. Started with EU FP6 support in 2004 and continued with EU FP7 support as DEISA2 in 2008, DEISA provides access and user support to this Infrastructure through DECI, the DEISA Extreme Computing Initiative, and through Virtual Science Community support.

PRACE and DEISA are closely cooperating with the goal to merge their efforts under a single umbrella. HPC-Europa is another FP7 project that provides transnational access to national HPC systems through a single Peer Review system.

User support in the form of application enabling and peta-scaling is of key importance for the effective use of the high-end systems and is a service that is provided along with granting access to the resources of the HPC Infrastructures. This service is not only provided on a per-user project basis, but also community-oriented.

Data management services

The importance of research data for modern science is growing daily, and new initiatives have been required to cope with the resulting “data flood”. Incorporating e-Science digital repositories and their holdings into an open information ecosystem will help support new scientific methods and paradigms, improving both the efficiency of the scientific process and its impact. This section outlines the data management services currently available via the e-infrastructure and the set of existing projects to provide data related services to specific user communities.

EGEE/EGI

A number of basic data manager services are provided by the EGEE/EGI grid. A Storage Element (SE) provides uniform access to data storage resources. The Storage Element may control simple disk servers, large disk arrays or tape-based Mass Storage Systems (MSS). Most EGEE sites provide at least one SE. Storage Elements can support different data access protocols and interfaces. Most storage resources are managed by a Storage Resource Manager (SRM), a middleware service for which there are a range of implementations providing capabilities like transparent file migration from disk to tape, file pinning, space reservation, etc.

The primary unit for Grid data management, as in traditional computing, is the file. In a Grid environment, files can have replicas at many different sites. Because all replicas must be consistent, Grid files cannot be modified after creation, only read and deleted. Ideally, users do not need to know where a file is located, as they use logical names for the files that the Data Management services use to locate and access. A file can be unambiguously identified by its Grid Unique IDentifier (GUID); this is assigned the first time the file is registered in the Grid, and is based on the UUID standard to guarantee its uniqueness. The mappings between logical and physical file names are kept in a service called a File Catalogue, while the files themselves are stored in Storage Elements.

The File Transfer Service (FTS) is a middleware service that takes care of accepting, scheduling and performing file transfers between SEs and it is very convenient for large-scale massive data transfers. More data management tools allow a user to move data in and out of the Grid, replicate files between Storage Elements, interact with the File Catalogue and more.

AMGA is a gLite metadata service which provides database access for Grid applications. AMGA is implemented as a thin layer between an application program and the underlying database, providing a Grid style authentication mechanism. AMGA uses a protocol to

transmit information which reduces latency in Wide Area Networks and implements a data streaming mechanism to retrieve information at high speed. AMGA also provides functionality to make applications interoperable with different database back-ends. AMGA's support for replication of relational data can be used to build scalable and reliable applications.

Hydra is an encrypted file storage service. It encrypts files and stores them on normal SEs. The sensitive information is the encryption key, which is split and distributed on several different places (grid sites), called the Hydra Keystores (Hydra Servers or services). This secure service provides the basis for a number of community-specific data management services, such as the Medical Data Manager (MDM) which is an interface for DICOM compliant storage. It provides access to medical data sources without interfering with clinical practice, ensures transparency so that accessing medical data does not require any specific user intervention, and offers a high data protection level to preserve patients' privacy.

Existing European scientific data infrastructure projects

While there is no singular pan-European scientific data infrastructure serving a wide range of disciplines, there are a number of projects that do provide elements of the data infrastructure for specific user communities. The GRDI2020⁴⁴ project has collected the list of such projects which is reproduced below:

Project	Description
EURO-VO	The European Virtual Observatory (EURO-VO) project aims at deploying an operational VO in Europe. The Virtual Observatory is an international astronomical community-based initiative. It aims to allow global electronic access to the available astronomical data archives of space and ground-based observatories, sky survey databases. It also aims to enable data analysis techniques through a coordinating entity that will provide common standards, wide-network bandwidth, and state-of-the-art analysis tools.
GENESI-DR	GENESI-DR, (Ground European Network for Earth Science Interoperations - Digital Repositories), has the challenge of establishing open Earth Science Digital Repository access for European and world-wide science users. GENESI-DR shall operate, validate and optimise the integrated access and use available digital data repositories to demonstrate how Europe can best respond to the emerging global needs relating to the state of the Earth, a demand that is unsatisfied so far.
Geo-Seas	Geo-Seas is implementing an e-infrastructure of 26 marine geological and geophysical data centres, located in 17 European maritime countries. Users will be able to identify, locate and access pan-European, harmonised and federated marine geological and geophysical datasets and derived data products held by the data centres through a single common data portal.
HELIO	Heliophysics is a new research field that explores the Sun-Solar System

⁴⁴ GRDI2020 - Towards a 10-Year Vision for Global Research Data Infrastructures (<http://www.grdi2020.eu/>)

- Connection; it requires the joint exploitation of solar, heliospheric, magnetospheric and ionospheric observations. The Heliophysics Integrated Observatory, HELIO, will deploy a distributed network of services that will address the needs of a broad community of researchers in heliophysics. HELIO is designed around a Service-oriented Architecture. HELIO will be a key component of a worldwide effort to integrate heliophysics data and will coordinate closely with international organizations to exploit synergies with complementary domains.
- IMPACT** The IMPACT (IMproving Protein Annotation through Coordination and Technology) project aims to harness existing technologies (such as web services and distributed computing) and use them to dramatically improve existing information resources. As a result, the IMPACT consortium will define and adopt new data formats to facilitate information exchange between partners, as well as enabling delivery of new data to users.
- METAFOR** The main objective of METAFOR is to develop a Common Information Model (CIM) to describe climate data and the models that produce it in a standard way, and to ensure the wide adoption of the CIM. METAFOR will address the fragmentation and gaps in availability of metadata (data describing data) as well as duplication of information collection and problems of identifying, accessing or using climate data that are currently found in existing repositories.
- OpenAIRE** OpenAIRE aims to support the implementation of Open Access in Europe by establishing the infrastructure for researchers to support them in complying with the EC OA pilot and the ERC Guidelines on Open Access. It provides the means to promote and realize the widespread adoption of the Open Access Policy.
- PARSE.Insight** PARSE.Insight is a two-year project co-funded by the European Union under the Seventh Framework Programme. It is concerned with the preservation of digital information in science, from primary data through analysis to the final publications resulting from the research.
- PESI** PESI is the next step in integrating and securing taxonomically authoritative species name registers that underpin the management of biodiversity in Europe. PESI will integrate the three main all-taxon registers in Europe, namely the European Register of Marine Species, Fauna Europaea, and Euro+Med PlantBase in coordination with EU based nomenclators and the network of EU based Global Species Databases. It is a standards based, quality controlled, expert validated, open-access infrastructure for research, education, and data and resource management.
- SEALS** The goal of the SEALS project is to provide an independent, open, scalable, extensible and sustainable infrastructure (the SEALS Platform) that allows the remote evaluation of semantic technologies thereby providing an objective comparison of the different existing semantic technologies. This will allow researchers and users to effectively compare the available technologies, helping them to select appropriate technologies and advancing the state of the art through continuous evaluation.

- VAMDC VAMDC aims at building an interoperable e-Infrastructure for the exchange of atomic and molecular data.
- D4Science-II D4Science-II, the follow-up phase of D4Science, will develop the technology to enable interoperation of data e-Infrastructures that are running autonomously, thereby creating e-Infrastructure Ecosystems that will serve a significantly expanded set of communities dealing with multidisciplinary, scientific and societal challenges. To set up a prototypical instance of such an ecosystem, D4Science-II will bring together several scientific e-Infrastructures established in areas such as biodiversity, fishery resources management and high energy physics.
- 4D4Life 4D4Life is a Scientific Data Infrastructures Project of the European Commission's e-Infrastructure Programme that aims to provide, through the "Catalogue of Life", a dynamically updated global index of validated scientific names, synonyms and common names integrated within a single taxonomic hierarchy. In its Networking Activities 4D4Life will strengthen the development of Global Species Databases that provide the core of the service, and extend the geographical reach of the programme beyond Europe by realizing a Multi-Hub Network integrating data from China, New Zealand, Australia, N. America and Brazil.

The Science Data Infrastructure Projects support the deployment of a broad European multidisciplinary scientific data infrastructure able to be easily federated with other knowledge infrastructures in other parts of the world, building upon the achievements of network and grid infrastructures and opening its benefits to other potential research areas such as e-health, e-learning and others.

In addition, a white paper⁴⁵ on a Strategy for a European Data Infrastructure was published in October 2009 by the PARADE consortium (Partnership for Accessing Data in Europe) proposing a European strategy for data related services and outlines a persistent, multidisciplinary European Data Infrastructure, based on the needs of user communities.

⁴⁵ <http://www.csc.fi/english/pages/parade>

Conclusions from the initial requirements analysis

This section outlines the conclusions that can be drawn from the initial requirements analysis and the challenges and opportunities for e-infrastructures. This analysis addresses the technical and functional aspects but the policy aspects, such as resource allocation, governance, and cost-sharing must also be addressed in order to put technical solutions into production service and ensure their long-term sustainability.

Since many of the ESFRI projects have stated that they have a clear need to make use of several of the existing European e-infrastructures, improving the inter-operability between these structures will have a definite added-value for all the user communities. So as part of the analysis work EEF members have identified areas where the infrastructures can work in common:

- access to resources (i.e. harmonising policies for Authentication, Authorization, Accounting and Auditing);
- user support (i.e. problem handling procedures) and training;
- security incident handling (i.e. cooperating security incident response group);
- data management (i.e. seamless authorised access to data across the infrastructures for users);
- dealing with perceived performance issues.

To pursue these areas for inter-operability, the EEF members propose to harmonize their existing services thereby offering a consistent access to all e-infrastructure resources visible to users in a manner that reflects the priorities ESFRI projects' requirements. The section below outlines for each of the common ESFRI requirements, the implications and opportunities for the e-infrastructures.

Single sign-on

All the ESFRI projects consulted identified consistent identity management and single sign-on as a fundamental requirement. A unified single sign-on service has to ensure an individual's identity can be used across Network, HPC and grid services. All the e-infrastructures in EEF have existing Authentication and Authorization Infrastructures (AAI) which are similar but not identical and are separately managed. Harmonising policies for Authentication, Authorization and potentially Accounting and Auditing will simplify access and usage to the e-infrastructures. The issue for EEF to offer what is requested by the ESFRI projects is to make these existing AAI systems interoperate so that a users identity can be established once and accepted by all the e-infrastructures.

All the e-infrastructures have dedicated security structures, procedures and measures in place intended to ensure the secure operation of the infrastructures. There are already examples of the security incidence response groups in the e-infrastructures co-operating and this will be generalised to ensure an effective and timely response to security threats can exist across the whole of the e-infrastructure ecosystem.

Virtual Organisations

All the ESFRI projects consulted identified the ability to control access to resources, data and applications on a community level as being necessary for at least some subset of their user communities and foreseen use cases. The HPC and grid infrastructures currently offer support for virtual organisations to differing levels of granularity and with differing semantics so a goal would be to offer consistent support for Virtual Organisations by harmonising current features.

Persistent storage

The ability to provide long-term (measured in decades rather than years) storage and accessibility was identified by several sectors though terminology differs. Persistent Identifiers (PIDs) and metadata are key issues for the user communities. There are existing services for registering, storing and resolving digital object identifiers (DOI)⁴⁶ such as the handle system⁴⁷ being offered by several consortia. Effective access to persistent data from the European e-infrastructures implies:

- Guarantees of quality of service and access for long-term storage will be required for the centres offering persistent data. To provide access implies that such centres are connected to the network (GEANT) and to ensure suitable quality of service (i.e. availability/reliability) they should be integrated into operations monitoring schemes such as those deployed by EGEE/EGI.
- The middleware deployed by European e-infrastructures will have to be modified to support access to persistent data using PIDs at these sites.

Provenance of data allowing the origins of data to be recorded and traced and its movement between databases has also been mentioned by several ESFRI projects.

User support

To effectively use the European e-infrastructures, users should have quick responses to questions and high-quality documentation. All the large-scale European e-infrastructures of today offer specific user support facilities.

EGEE offers user support through the central Global Grid User Support (GGUS) portal⁴⁸ via a web form or e-mail, or at their Regional Operations' Centre (ROC) or their VO which will be continued within EGI. This central helpdesk keeps track of all service requests and assigns them to the appropriate Support Units.

DEISA has a similar support structure accessible via the central DEISA Helpdesk⁴⁹. PRACE will offer a broad range of support ranging from first-level helpdesk to user- and community-

⁴⁶ http://en.wikipedia.org/wiki/Digital_object_identifier

⁴⁷ <http://www.handle.net>

⁴⁸ <http://ggus.org/>

⁴⁹ <http://www.deisa.eu/usersupport/primer/access-to-the-user-support>

oriented application optimisation and scaling. GEANT's regional model relies on user support⁵⁰ provided by the NRENs as well as a centralised end-to-end co-ordination capability.

The EEf members will work to make these existing user support structures cooperate and requests can be issued to appropriate support groups across the different technologies and geographical regions.

Training and consultancy

All ESFRI projects have expressed the need for training, education or external expertise in their use of e-infrastructures. The EEf members all offer varying levels of existing support in this area.

GEANT provides network performance analysis expertise and has an E-Learning portal⁵¹. The requirements that projects have described in terms of their ICT requirements are very broad-ranging as far as "Networks" are concerned. In technology and capacity terms, there is nothing that cannot be accommodated by current and predicated technology departments. The requirements range from multiple access of databases from a diverse population of users, through to much more concentrated flows between key project locations. The portfolio of services that is available across the GEANT service area, including high performance IP-configurable Point to Point connections and, where appropriate, dedicated wavelength capacity, is capable of meeting the needs that have been articulated. For more complex requirements, a design, and the implementation of networking needs will require cooperative effort between the research network community and the project participants. Performance, particularly where demanding applications are being supported, will need monitoring and fine-tuning. This is a non-issue and the techniques of addressing it are established. It needs to be stated, that overall performance in terms of complex systems could be challenging, as it involves interactions between different systems under separate management control. Network tools to help debug such problems are available. As part of customer support, GEANT is prepared to analyse and diagnose performance problems.

DEISA provides general training for new users of the distributed HPC infrastructure, and has made two specific training courses for the fusion and Virtual Physical Human (VPH) communities⁵². PRACE is offering training in all HPC related topics. The material from workshops, summer and winter schools is made available online⁵³. EGEE has an extensive training programme⁵⁴ and material repository⁵⁵ which is used at a wide range of events⁵⁶ for end-users, application developers, site managers and even the trainers themselves.

Experience shows that training programmes have the most impact when tailored to the specific needs of the target user community. The EEf members are willing to contribute trainers and material to training programmes and events that could be organised by the ESFRI projects either individually or on a sector basis.

⁵⁰ <http://www.geant.net/Users/Support/pages/home.aspx>

⁵¹ <http://cbt.geant.net/courses.html>

⁵² <http://www.deisa.eu/usersupport/training>

⁵³ <http://www.prace-project.eu/hpc-training>

⁵⁴ <http://www.eu-egee.org/index.php?id=227>

⁵⁵ <http://library.nesc.ed.ac.uk:8080/egee>

⁵⁶ <http://www.egee.nesc.ac.uk/schedreg/index.cfm>

Web-service interfaces

While a range of standards were mentioned for all aspects of e-infrastructure usage, a common theme from all the ESFRI projects consulted was the identification of web-services as a standardised manner of packaging e-infrastructure services. Many ESFRI projects highlighted the importance of the well defined interfaces for web-services, a registration facility and the ability to discover (search for) new web-services and a consistent view to managing the life-cycle of web-services. The consequence of these findings is that, where appropriate, e-infrastructures should offer web-service interfaces for their relevant services wherever possible and allow the user communities to build on these to produce their own customised web-services.

The RESPECT program (Recommended External Software for EGEE CommuniTies) publicises and provides access to proven and useful grid software and services that work well in concert with the EGEE-produced gLite open source middleware. Third-party software packages (including commercially licensed ones) can also be integrated into the EGEE grid environment. Support for similar repositories of community specific and third-party packages that can run across the whole of European e-infrastructure will be required.

Workflows

The need for workflows⁵⁷ was identified by all ESFRI research sectors. There are many workflow tools or frameworks employed by the user communities and this diversity is certain to remain. As a result, the EEF members will work to provide an environment which can support a range of workflow tools and environments. The implication of supporting such cross infrastructure workflows seamlessly is that the AAI, virtual organisation and data management inter-operation aspects mentioned above must be in place.

Global scope

Although not explicit in the EEF questionnaire, it became apparent that all ESFRI sectors identified the need to collaborate with parties beyond Europe's borders. The European e-infrastructure Forum members can leverage their existing international contacts for the benefit of the ESFRI projects. For ELIXIR the USA (NCBI) and Japan (DDBJ) are identified key partners. The global requirements that have been articulated to date are within the activity footprint of the GÉANT global reach and relationships are in place to assist in organising network requirements beyond Europe. Availability of services should not be a problem but for specific locations access capacity to those locations, including capacity available for elements of the service portfolio, would need to be confirmed and, if appropriate, addressed.

From a grid and HPC point of view, there are already a number of interoperation points addressed via the Infrastructure Policy Group⁵⁸ which meets at the the Open Grid Forum⁵⁹

⁵⁷ <http://en.wikipedia.org/wiki/Workflow>

⁵⁸ <http://forge.ogf.org/sf/wiki/do/viewPage/projects.ipg/wiki/HomePage>

⁵⁹ <http://www.ogf.org/>

where EGEE/EGI, DEISA, TeraGrid⁶⁰, OSG⁶¹ and NAREGI⁶² meet regularly to simplify their interaction.

Integration with cloud systems and volunteer desk-top systems

The majority of the ESFRI projects consulted are unconcerned as to where resources come from, be they grids, clouds or supercomputers, and are primarily interested in easy-to-use, yet powerful and secure, data management facilities. External commercially operated clouds, which allow the user to create ‘virtual computers’ including the applications and operating systems of their choice, can be a very good for-purchase solution for a user who needs additional resources on demand. However, the tasks of some other ESFRI projects require computation only possible on sophisticated, high-performance computational resources such as supercomputers —currently not provided by clouds. Similarly, there are important policies questions to be addressed concerning large-scale data management and archive on commercial cloud services.

Each computing paradigm has its advantages and drawbacks, and, in the end, a custom fit solution for each user community will surely work the best. The work of standards bodies, such as Open Grid Forum, aims to make it easier to bring clouds, grids and supercomputer installations together by defining interfaces designed to simplify and promote their interoperability. Interoperability work undertaken so far among EEF partners (EGEE and DEISA), as well as EGEE with volunteer grids (such as Enabling Desktop Grids for e-Science, EDGeS⁶³) and cloud systems (via the RESERVOIR⁶⁴ project) has been driven by the needs of users, such as the fusion and life science communities. These distributed computing solutions will continue to complement grid computing in the future and, depending on the evolution of the needs of the research communities, will become part of the European e-infrastructure ecosystem offering.

⁶⁰ <https://www.teragrid.org/>

⁶¹ <http://www.opensciencegrid.org/>

⁶² http://www.naregi.org/index_e.html

⁶³ <http://www.edges-grid.eu>

⁶⁴ <http://www.reservoir-fp7.eu/>

Next Steps

The initial analysis of the ESFRI requirements described above have lead to a number of conclusions about what can be done in common between the ESFRI projects and the European e-infrastructures. The EEF members welcome feedback on this analysis and conclusions drawn from all the stakeholders and specifically the ESFRI projects.

The detailed requirements and the best mechanisms for addressing them requires further study involving expert intervention from the researchers of ESFRI user communities and the e-infrastructures. In discussions with projects from the BMS and SSH sectors, they expressed their willingness to work together on pilot projects that can explore key functionality satisfying agreed requirements for identified and engaged users. It is proposed that a series of such joint pilot projects are identified and planned across all the ESFRI sectors and e-infrastructures to ensure maximum coverage with minimum overlap of effort and resources. The objective would be that such pilot projects lead to working prototypes that are relevant for a range of user communities and ESFRI projects. Subsequently the results could lead to the introduction of the successful functionality and services in the production systems of European e-infrastructures.

Members of the EEF

The European Strategy Forum on Research Infrastructures Roadmap states that Research Infrastructures “often require structured information systems related to data management, enabling information and communication. These include ICT-based infrastructures such as Grid, computing, software and middleware.” and continues with “e-Infrastructures are critical to all projects in this roadmap”.

The European e-Infrastructure Forum (EEF) is a forum for the discussion of principles and practices to create synergies for distributed Infrastructures. The goal of the European e-Infrastructure Forum is the achievement of seamless interoperation of leading e-Infrastructures serving the European Research Area. The focus of the forum is the needs of the user communities that require services which can only be achieved by collaborating Infrastructures. Its current membership includes GEANT, Terena, EGEE, EGI, DEISA and PRACE. The forum recognises the importance of data access and management and is seeking to add a member specialising in service provision in this area for multiple research communities.

The following individuals have contributed to the writing of this report:

Bob Jones (Editor of the report, CERN, representative of EGEE)

Dai Davies (DANTE, representative of GEANT)

Hermann Lederer (Rechenzentrum Garching der Max-Planck-Gesellschaft, representative of DEISA)

Patrick Aerts (Netherlands Organisation for Scientific Research, representative of PRACE)

Thomas Eickermann (Forschungszentrum Juelich, representative of PRACE)

Steven Newhouse (CERN, Director EGI.eu, representative of EGI)