ESFRI project requirements

for

Pan-European e-infrastructure resources and facilities

An EEF-publication

Version 1.0 April 9th 2010

Table of contents

Preface

The European Strategy Forum on Research Infrastructures (ESFRI) has issued a list of major Pan-European facilities and services and subsequent updates thereto, which are unique to Europe due to their sheer size or cost involved in their establishment. The e-IRG 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 have 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 document and the actions 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 after in order to avoid doubling of efforts. Because the main goal behind this EEF effort is to have the best facilities offered to the scientific communities that make use of the facilities in the ESFRI-projects, tailored to their need where possible or required, but avoid double efforts and investments by the ESFRI-projects to serve their own communities at the expense of the project budgets. This should 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 European E-infrastructures Forum (EEF) represents 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 effort is to tailor our services to the ESFRI-project's 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.

Place, date

Summary of findings

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 matrix a common series of themes were emerging from the expressed requirements and that it would be worthwhile investigate these subjects further. As a consequence EEF developed a questionnaire which it has used to collect further 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 results of a number of other events involving e-infrastructure and ESFRI project representatives, notably the NEERI09[1], the European association of national Research Facilities laboratories (ERF) workshop[2], a series of workshops organised by the European Commission for the biological and medical (BMS)[3], social sciences & humanities (SSH)[4] and environmental sciences (ENV)[5] as well as the e-infrastructure concentration meeting[6].

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

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 found:
CLARIN, ESS, DARIAH, SHARE and CESSDA.

CLARIN[7] 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.

---

[1] NEERI09, Helsinki on 1-2 October 2009 http://www.csc.fi/english/pages/neeri09
[2] Future Access to European Research Infrastructures: Benefits to Academia, Industry and Society, Lund, 27th of October 2009, http://www.europeanresearchfacilities.eu/
[3] Workshop on ICT and e-infrastructure needs for European Research Infrastructures in the field of Life Sciences (BMS), Brussels, 16 December 2009
[4] workshop on Common Needs and Common Solutions for the ESFRI research infrastructures for the Social Sciences and Humanities (SSH), Brussels, 20th January 2010
[5] Workshop on common ICT and e-infrastructure needs for the ESFRI Research Infrastructures in the field of Environmental Sciences (ENV), Brussels, 18th March 2010
[6] 7th Concertation Meeting, held in Brussels on 12-15 October 2009
[7] http://www.mpi.nl/clarin/

ESS[8] (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.

DARIAH[9] (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.

SHARE[10] (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.

CESSDA 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 virtually 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 SSH the 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 seen as being of lesser importance.

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

[8] http://www.europeansocialsurvey.org
[9] http://www.dariah.eu
[10] http://www.share-project.org

ELIXIR[11]: a secure, evolving platform for biological data collection, storage and management, consisting of an interlinked set of core and specialist resources.

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

ECRIN[12]: supports multi-national clinical research trials in Europe by connecting together nationally coordinated networks of clinical research networks and clinical trials units.

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

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

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

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

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

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

neuGRID: 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 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:

[11] http://www.ebi.ac.uk
[12] http://www.ecrin.org

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

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

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

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

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

EUSAAR-I3: is a project to provide for the integration of atmospheric aerosol properties measured at a distributed network of European ground stations.

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

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

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

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

Material and analytical facilities

EGEE/EGI:

EGEE09: http://indico.cern.ch/sessionDisplay.py?sessionId=12&confId=55893
Namely European XFEL

- One paragraph (extracted from ESFRI roadmap) explaining what disciplines material & analytical covers
- List of ESFRI-projects consulted (give one sentence to define each project):.

General description of the nature of the domain and the general types of requirements, an overview

Physical and engineering sciences

EGEE/EGI:

EGEE09: http://indico.cern.ch/sessionDisplay.py?sessionId=12&confId=55893

Namely European CTA, FAIR, SKA

- One paragraph (extracted from ESFRI roadmap) explaining what disciplines phys & eng covers
- List of ESFRI-projects consulted (give one sentence to define each project):.

General description of the nature of the domain and the general types of requirements, an overview

Energy

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

The areas covered are:
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 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 tokamaks: JET in the UK, ASDEX Upgrade in Germany, Tore Supra in France, MAST in the UK, 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 since two years, has been interfaced in a pilot project both to EGEE grid and DEISA supercomputer resources. 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 Jan 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).

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

Networking

DANTE/TERENA:

General description, projects and resources and services

Data storage and services

Bob:

- List facilities currently available project such as D4ScienceII, GENESI-DR, DRIVER, OPENAIRE. (http://www.grdi2020.eu/index.php?page=synergies)
- Summary of data services currently offer by EGEE and DEISA (extracted from online material and IPG notes)
- Mention PARADE/EUDAT white paper
- Explain the situation that we have an empty seat for this area in EEF

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 funded by the European Commission to implement, deploy and maintain a distributed computing infrastructure to support researchers in many scientific domains.

In 2009 this infrastructure supports world class science in over 50 countries, consisting of about 290 sites, encompassing more than 144,000 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.

In its third and final phase, the EGEE project has set its eyes on the future of grid computing. It is absorbed in preparations for transitioning to a sustainably funded model—the European Grid Initiative, a federated infrastructure coordinated by EGI.eu, and based on National Grid Initiatives, that will ensure grid operations on a national level. The new EGI will provide a sustainable environment for providing grid-based computing services through a stable collaborative European and national co-funding scheme. It will enable easy sharing of resources - such as computation, storage and data—across borders to ensure the technological

interoperability of global grids and contribute to the realisation of the European Research Area.
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. Once the LHC is fully online, the experiments will produce an expected 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.
The EGEE project will come to a close at the end of April 2010. A new organizational model, implemented by the European Grid Initiative [3] (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.
Creating and maintaining Virtual Organisations: An individual can only use resources if they are a member of a virtual organisation. VOs must contribute resources equivalent to their average usage. A resource allocation group acts as a broker between resource centres and the VOs.

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

Training: Since getting on the grid is not necessarily easy, EGEE offers many forms of training. Schools are held for end-users, application developers, site managers and even the trainers themselves.

Application porting: Many grid users begin with applications they already use in their daily work, which need to be 'gridified' or ported to run on the grid. Experts from the Application

Porting Support group work closely with application owners to understand their requirements and to identify suitable approaches and tools for the porting process.
Software: Middleware—commonly thought of as the 'glue' holding the grid together—is responsible for both basic and advanced functions. It ensures the security of the infrastructure, manages monitoring and accounting systems, and access to computing and storage resources. At a higher-level it also manages job execution, data catalogues and data replication.

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.

Operational monitoring: EGEE supports many monitoring and accounting tools. Such information contributes to the overall health of the infrastructure by reflecting its performance and identifying room for improvement.

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.

Conclusions from the initial analysis

NETWORKS

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 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. The portfolio of services that is available across the GÉANT 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. The issues come in respect of the following areas.

Geographic Scope

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.

Complex Requirements

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

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, GÉANT is prepared to analyse and diagnose performance problems.

Identity Management

This is something that is available as part of overall network capability. Its precise role in supporting project needs requires further development.

Bob:

Clear "issues", lacking facilities or resources, challenges of administrative nature, legal aspects

Explain findings made based on the analysis performed (based on Bob's notes from EEF meeting of 29th March)

- **identity management (with single sign-on)**

All the ESFRI projects consulted identified consistent identity management and single sign-on as a basic requirement

All the e-infrastructures in EEF have existing Authentication and Authorization Infrastructures (AAI) which are similar but not identical and are separately managed. 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 identify can be established once and accepted by all the e-infrastructures.

Dai said integrated systems for identify management at the network layer across Europe is a goal for GEANT3 but is considered difficult. The GEANT contact for these aspects is Josh Howlett.

From the SSH event held in Helsinki (NEERI09) it was stated that CLARIN had been working networking groups to explore AAI. At this event Diego Lopez (RedIRIS) presented the state of federation technologies and requirements. Diego explained the plans for GEANT and NRENS to offer Messaging, trust and Identity services to the eScience community.
He highlighted the role of ECAM (US equivalent is MACE) and REFEDS as well as TACAR (www.tacar.org) which is the basis for EUGridPMA, eduRoam and IGTF. He mentioned the Terena Certificates for Servers and how the world will not have a single structure but rather it will be federations of federations which are independently managed but have compatible AA infrastructures.
He also outlined SCHAC (SCHema for ACademia) for building a service repository and discovery service.
Prototypes bringing all of these elements together with single sign on are being built by the DAMe project and GEMbus.

Steven said EGI considered the subject important but there is no specific manpower allocate for the subject in the EGI-InSPIRE project (currently under negotiation) so its action will be to try and lobby the middleware providers to ensure the different identity systems are recognised and accepted.
John said beyond the current LDAP scheme which involves each institute, DEISA is currently testing shibboleth and the Short Lived Credential Service (SLCS).

Steven noted that SLCS is included in the work programme for EMI (Consolidation of security models across the three middleware stacks).

The question of authorisation (i.e. permission to consume resources) is dealt with below.

- **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. Steven said for EGEE and EGI this is implemented and relies on the Virtual Organisation Management Service (VOMS).
John said that DEISA is working on the integration with the VOMS version which supports SAML and at the moment eh mapping of communities is done only at the level of projects. Hermann added that in the future DEISA expects to be able to delegate the responsibility for creation of users accounts to the communities themselves.

The idea of virtualising the network layer is something that was not being considered at the moment by GEANT while the support for multiple VPNs is possible though and how to map the HPC-grid interpretation of VOs to network resources is unclear and requires further investigation.

- **Secure data management**

The ability to strictly control access to sensitive data was identified as a requirement for specific use cases for sub-sets of the SSH and BMS (notably medical trails) communities.

Secure data management facilities exist within EGEE grid infrastructure using extension of the gLite middleware employing encryption technology (Hydra) built on the AAI and linked to VOMS. Otherwise separate infrastructure have been deployed for particularly sensitive patient data (notably via the health-e-child project).

It was noted that EEF has a vacant seat for a European data infrastructure provider to which it would delegate responsibility for providing advanced data management facilities.

- **Persistent data**

The ability to provide long-term (measured in years) storage and accessibility was identified by the SSH and BMS sectors though terminology differs. The SSH community has the clearest views on what they require. It is assumed that ENV also has similar requirements (notably for LifeWatch) though these have not been made explicit. For EEF these points are important:

- The user communities have to understand that no-one is going to provide long-term storage facilities for free so either they have to find centres in their community that are able and prepared to do this or we have to assume the yet-to-be-defined European data infrastructure provider will do this.
- The middleware deployed by EEF has to be able access persistent data (see the discussion about PIDs below taken from my NEERI09 notes) at these sites.
- 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 grid operations monitoring scheme (EGI).

Persistent Identifiers (rather than temporary URLs) and metadata are key issues for the SSH community. There are important developments in the Persistent Identifiers (PIDs) domain with services for registering, storing and resolving identifiers based on handles (see below) being offered and the formation of consortia to run such services. If these services prove useful to projects such as CLARIN then EEF and grid middleware will have to be able to use PIDs if it wants to work with and support this community.

Information about PIDs:

(http://www.handle.net) which has been designed to provide reliable, scalable handle resolution system for persistent identifiers
The procedure for introducing a new set of identifiers and servers is described here: http://www.handle.net/start.html
From the quick facts on the handle website:

Some quick facts about the Handle System.

- There are over 1,000 handle services running today, located in 51 countries, on 6 continents; more than 750 of them are at universities and libraries.
- Handle services are being run by user federations, national libraries, national laboratories, universities, computing centers, libraries (national and local) government agencies, contractors, corporations, and research groups.
- The number of prefixes, which allow users to assign handles, is growing and passed 200,000 in January 2009. The total number of handles under those prefixes is not precisely known (since users do not have to declare them) but certainly exceeds 600 million.
- The International DOI Foundation's implementation of handles, the DOI® System, has over 40 million registered handles.
- There are four top-level Global Handle Registry® servers that receive (on average) 68 million resolution requests per month. Proxy servers known to CNRI, passing requests to the system from the web, receive (on average) 50 million resolution requests per month.
- The Handle System infrastructure is supported by prefix registration and service fees. The majority of those fees come from single prefix holders, while the largest single contributor is the International DOI Foundation.
- Among the objects we know of that are identified by handles are journal articles, technical reports, books, theses and dissertations, government documents, metadata, distributed learning content, and data sets. Handles are being used in digital watermarking applications, GRID applications, repositories, and more.

Note there is also a European service using the same handle technology but operated by a different consortium of partners.

Provenance of data was also mentioned by the ENV projects and is known to be of importance for specific BMS use cases though this was not covered in the EEF questionnaire.

- **Global scope**

Although not explicit in the EEF question, it became apparent that all ESFRI sectors identified the need to collaborate with parties beyond Europe's borders.
For BMS (ELIXIR) the USA (DDBJ) and Japan (NCBI) are key partners. What is important for network layer is to understand the end points involved in such international collaboration. From a grid and HPC point of view, there are already a number of interoperation points addressed via the Infrastructure Policy Group

(http://forge.ogf.org/sf/wiki/do/viewPage/projects.ipg/wiki/HomePage) where EGEE/EGI, DEISA, TeraGrid, OSG and NAREGI meet. Further middleware points are discussed in the Production Grid Infrastructures (PGI) working group (http://www.ogf.org/gf/group_info/view.php?group=pgi-wg) of the Open Grid Forum

- **Training and education**

All ESFRI projects have expressed the need for training, education or external expertise in their use of e-infrastructures.
Dominico described the training GEANT provides concerning network performance analysis and improvements. GEANT has an E-Learning portal http://cbt.geant.net/courses.html including self paced training about perfSONAR (multi-domain network monitoring tool) http://cbt.geant2.net/repository/perfsonar_ui/perfsonar_online_training_0.2/player.html and has other self-paced education/training courses: http://www.geant2.net/server/show/ConWebDoc.2757

John and Hermann said DEISA has done similar training for the fusion and Virtual Physical Human (VPH) communities: http://www.deisa.eu/usersupport/training

Steven said EGEE has an extensive training programme (http://www.eu-egee.org/index.php?id=227) and material (http://training.eu-egee.org/index.php?id=234) which has been used at a wide range of events (http://www.egee.nesc.ac.uk/schedreg/index.cfm). EGEE has also taken part is the grid Winter School (http://www.iceage-eu.org/iwsgc10/index.cfm) which is an annual virtual training event.
He added that EGI will also maintain the repository of grid training material while the organisation of events will be the responsible of the national grid initiatives (NGIs).
Steven said there would not be a International Summer School for Grid Computing (ISSGC) this year and added that it is worth considering a summer e-infrastructure school in the future.
Thomas said PRACE has a training programme foreseen in its programme of work with funded partners.

It was agreed that EEF members will suggest that ESFRI that they co-organise training events specifically tailored to the needs of their user communities. EEF members will contribute by providing trainers and material.

Dai also added that GEANT advertises the services it has via a portfolio (http://www.dante.net/upload/pdf/GN3-10-040_DN4-2-1_GEANT_Service_Portfolio.pdf). The presentation of services via a portfolio could be extended to all e-infrastructure services.

- **Standards**

This was a rather vague notion that is understood by the ESFRI project in various ways. 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.

The consequence of these findings is that, where appropriate, e-infrastructures should offer web-service interfaces for their services and allow the user communities to build on these to produce their own customised web-services.

The issue of standards in the data management area were also popular though the specifics vary between projects and research sectors. It was agreed that EEF will leave this area of standardisation to the ESFRI projects themselves and a future European data infrastructure provider.

- **Workflows**

The need for workflows was identified by all ESFRI research sectors. For a definition of a workflow see here: http://en.wikipedia.org/wiki/Workflow
and more specifically in grid computing context see here:
http://www.gridworkflow.org/snips/gridworkflow/space/Grid+Workflow

The intention behind this question was that we have experienced examples (notably in the fusion and life sciences domains) where such work-flows span DEISA and EGEE resources (and consequently make use of the networking layer) and we wanted to understand how widespread was this requirement.
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.
There are many workflow tools or frameworks employed by the different ESFRI projects and this diversity is certain to remain.

It was agreed that EEF will not offer a specific workflow service to ESFRI projects but will support their existence by working to ensure the interoperation of the under-lying infrastructures.

The discussion on workflows led to the subject of resource allocation policies. While each infrastructure has a well defined process for allocating resources they are different and disconnected. This situation will not change in the foreseeable future but an added value that EEF can provide is to offer a central point (i.e. web-portal) where requests from user communities could be submitted and distributed to all the e-infrastructures concerned.

The cost-sharing of resources allocated to the ESFRI projects was discussed and again, due to the different models employed by the e-infrastructures, variations will remain. However, a consistent underlying accounting model would allow user communities to get a clear over-view of their consumption of e-infrastructure resources.

Dai suggested that we are likely to see a move towards a more pay-as-you-go model for the use of e-infrastructures and this on-demand model will be more complex to manage. The subject of resource allocation, cost-sharing and on-demand services is something that was not covered in this first iteration of requirements gathering and will need to be discussed with the ESFRI project sand stakeholders.

- **Real-time constraints**

The subject of real-time constraints in the processing of data was not addressed explicitly in the questionnaire but was eluded to in the description of some of the use cases described by specific ESFRI projects. This appears to be most relevant for ENV ESFRI projects where they are working with sensor networks. The implications are not yet known because there are large variations about the scale (how many sensors, how much data from each one and how frequently), geographical deployment (from localised to large areas spanning several countries) and mobility (some ESFRI projects are mobile laboratories such as ships or aircraft) and little information about the division of responsibility.

The EISCAT_3D project will provide state-of-the art radar facilities to study processes taking place in Earth's atmosphere, taking measurements from the upper stratosphere to the magnetosphere and beyond. With antennas at multiple sites in Norway, Finland and Sweden, this new facility will greatly extend the amount of data available to scientists, both by improving its resolution and by extending its geographic, altitudinal and temporal range.
"The EISCAT_3D project will involve up to 100,000 radar antenna elements, each sampling a noise-like backscattered signal from the upper atmosphere at microsecond resolution or greater," says Ian McCrea of the project. "This creates an enormous amount of data, which has to be pre-processed on site into smaller products. Our users want as much flexibility as possible for subsequent data analysis, but each stage of pre-processing reduces this flexibility. Our goal is to provide users with a dataset large enough to be processed very flexibly, but which they can still access easily and conveniently.

For example can we assume the ESFRI projects will be responsible for data acquisition and concentration while EEF e-infrastructures would be more involved in distributing acquired data to the user communities and engaged in its processing and storage. These subjects needs to be discussed in more detail with the ESFRI projects concerned.

Having reviewed the material and common points the meeting agreed it was important to pursue this analysis with the ESFRI projects and test the assumptions and proposed approaches via a set of pilot projects based on use cases defined by the ESFRI projects themselves. The idea being that we could prototype many of the proposals outlined above via such pilot projects that would engage EEF and the ESFRI projects.

The subject of how EEF would be able to complete any necessary work associated with such pilot projects was discussed. The e-infrastructures already have a well defined set of objectives and programme of work that has been agreed with their stakeholders and the EC. Some of the points raised above are consistent with these objectives and could be included in the programmes of work while others are new or

beyond the scope of what is possible with the existing resources. Two complementary possibilities to cover the additional work were suggested:

- Adjust the existing objectives and programmes of work to reflect the high priority of the engagement with the ESFRI projects. This would imply negotiating these changes with the stakeholders of the e-infrastructures and the EC
- Seek additional resources via a new project. This assumes that this the partners can find additional manpower that they could assign to a new project and that the EC considers the subject suitable for funding via one of their forthcoming calls

Highlight that the analysis only covers technical aspects but the issue of policies for resource allocation, governance and cost-sharing must also be addressed in order to put these into service.

State that this is just the first iteration and further investigations are needed with the ESFRI projects

Consequences for e-infrastructures – areas where we see the definite need for close inter-operation (extracted from the above and including user support, dissemination/outreach)

Next steps

- Establish joint pilot projects with ESFRI projects
- Continue with requirements analysis, interaction with ESFRI events

Also need to determine how e-infrastructure will accommodate these changes/developments – adjust existing objectives/find additional resources etc.

Members of the EEF

Bob:
extract from EEF mission statement and membership docs

References

- ESFRI roadmap (2008)
- ESFRI implementation report (2009)