Scientific Data e-Infrastructures in the European Capacities Programme

Krystyna Marek, Carlos Morais Pires, Kostas Glinos

European Commission

Information Society and Media Directorate-General

BU25 - 4/64, 1049 Brussels

EMail: *Krystyna.Marek@ec.europa.eu*

e-Infrastructures and Scientific Information in the Digital Age

The recent Communication "*ICT infrastructures for e-Science*" [1] highlights the key role of ICT based infrastructures (*e-Infrastructures*) as a crucial asset underpinning European research and innovation policies. Building in Europe a high-quality and reliable repository infrastructure is a key component of a renewed strategy for *e-Science*, *e-Infrastructures*, and *innovation*.

The European Commission, in a coordinated effort with EU Member states, is investing in different domains of e-Infrastructures. The aim is to connect researchers, scholars, educators and students through high speed research networks (GÉANT), provide access to cooperative grid-computing platforms (EGEE, e-Science grids), develop supercomputing capacity for very demanding applications (DEISA, PRACE) and help the emergence of a robust platform for access and preservation of scientific information (Scientific Data e-Infrastructure). Already in the Communication "*On Scientific Information in the Digital Age: Access, Dissemination and Preservation*" [2], the European Commission has pointed out that building in Europe a dynamic information society requires providing wide access and ensuring long term preservation of scientific information. The document stresses the fact that "*The Internet […] opened new ways to use masses of data resulting from experiments and observations in the scientific process and to extract meaning from this data stored in repositories in combination with other scientific information resources. This leads to a 'continuum' of scientific information space from raw data to publications across different communities and countries*".

Several international declarations have emphasised the importance of research data for modern science. Incorporating e-Science digital repositories and their holdings into an open information ecosystem will deepen and broaden Europe's *Single Information Space*. It will support new scientific methods and paradigms, improving the efficiency of the scientific process and its impact.

The *Scientific Data e-Infrastructure* (SDI) domain was launched in 2007 under the Capacities programme of FP7[1]. It complements the investment done by the European Commission, in a coordinated effort with EU Member states, in different domains of e-Infrastructures.

e-Science

e-Science is one of the three vectors of the renewed strategy presented in the communication [1] that the European Commission addressed to the European Parliament and Council of Member States.

e-Science is generally defined *as the combination of three different developments: the sharing of computational resources, distributed access to massive datasets and the use of digital platforms for collaboration and communication* [3].

Disclaimer: "The views expressed in this presentation are those of the authors and do not necessarily reflect the views of the European Commission"

[1] Seventh Framework Programme for research and technological development.

By its very nature, e-Science is broad in scope, covering the complete spectrum of modern research and education from the arts to the traditional physical science, from the theoretical to the experimental, from the commercial to the academic, amongst numerous other dimensions.

e-Infrastructures

As defined in the communication [1], an e-Infrastructure is "*an environment where research resources (hardware, software and content) can be readily shared and accessed wherever this is necessary to promote better and more effective research. Such an environment integrates networks, grids and middleware, computational resources, experimental workbenches, data repositories, tools and instruments, and the operational support for global virtual research collaboration*."

The concept of an e-Infrastructure evolves as Science itself is evolving and bringing new requirements. Changing the ways of doing science, by using fast networks and computational technologies induced the emergency of a data-centric science which is based wholly or to a great extent on exploiting existing information (usually, very large datasets of raw experimental data or curated data or even scholarly papers with research results).

A wider definition of e-Infrastructure including the Scientific Data aspects:

An e-Infrastructure for e-Science digital repositories is taken to be the technical and administrative framework and facilities underlying e-Science digital repositories. Until recently, the concept of e-Infrastructure has usually been defined minimally, to include networks, authentication and authorisation mechanisms, middleware, computational resources (in particular high-performance computers), and those which enable collaborative working, including Grid technologies. We adopt a wider interpretation and include technologies of various kinds for creating, collecting, annotating, manipulating, storing, finding and re-using information and services such as those to provide user support, and training, preservation. Further, we include information resources and associated tools such as vocabularies, ontologies, rights management and privacy protection systems, and curation. Several of these resources depend upon manual human input.

Source: e-SciDR study

First steps towards European e-Infrastructure for Scientific Data

There has been a growing concern in the past five years both on policy and technical issues related with managing scientific information.

Some EU Member States launched initiatives[2] to support education and research institutions and communities to manage their scientific data. Several projects were supported by the Sixth European Framework programme (FP6) under thematic programmes (Health, Environment, etc.). They were focused on problems of designing, operating and maintaining databases holding discipline specific content. The area of Research Infrastructure supported, still in FP6, a few projects addressing challenges related with data in distributed computing environments (data grids).

The ICT programme has been supporting the adaptation of Libraries to new Internet paradigms of on-line communication. The focus has been on tools for information management and preservation of

[2] As part of national e-Science programmes.

digital documents in the context of the European Cultural Heritage. The *Europeana* initiative has emerged in that context[3].

All the above initiatives were the first responses to the so called '*data flood*'. They reflected the specific needs and organisational structures of research and education institutions, laboratories, data centres and libraries.

In spite of these efforts, the European data infrastructures landscape as a whole remains heterogeneous and scattered. The approaches to the most pressing problems related with storage, access, quality assurance, and preservation are not well coordinated and the integration with e-Science computational environments is far from being achieved. It is estimated that only a very limited part of European research output is managed in digital repositories.

Capacity building in the Scientific Data domain was therefore very much needed. In launching the SDI area in FP7 several issues were taken into account:

- the need to gain insight into the needs, practices and requirements of major stakeholders in research: researchers, libraries and data centers, publishers and research funders;
- the importance to coordinate with EU Member States and leverage national e-Science initiatives;
- the need to bring together dynamic scientific communities with their most pressing problems of data management;
- the objective to provide e-Infrastructures capable to fulfil European policies for widespread access and long term preservation;
- the need to exploit synergies with e-Infrastructures for networking and computing in order to offer integrated services supporting e-Science;
- the objective to place Europe in key global scientific partnerships acknowledging that science is mostly an international endeavour which requires bridging national and institutional policy and funding structures.

First two calls for proposals in FP7

Having in mind the emergence of e-Science and the need for Europe to be an active and inspiring partner in global collaborations, the European Commission involved the wider community of interested stakeholders in several preparatory workshops and launched studies to analyse the landscape of repositories in Europe and provide recommendations for the SDI area.

Issues like promoting multi-disciplinary research and widespread access to scientific information resources, sharing practices and software and exploiting synergies of investments were explicitly set as some of the desirable features for the European e-infrastructure for scientific repositories.

The heterogeneity and lack of the necessary maturity before engaging in larger scale e-Infrastructure initiatives resulted in a decision of splitting the available resources for the SDI area in two open calls for proposals.

The first call provided the opportunity for communities who had already well identified challenges to propose projects addressing their most pressing problems related with data management. The selected projects cover the areas of bioinformatics protein databases, astronomic data, earth observation data, metadata for climate modelling and space-physics. Two other projects were selected

[3] http://www.europeana.eu/portal/aboutus.html

focusing on the development of e-Infrastructure inter-linking digital repositories and supporting the emergence of virtual research environments by making ICT-tools available for data analysis, computation, simulation and visualisation.

The second call was designed to encourage communities to identify and tackle common (or boundary) problems, which cut across different scientific domains, to promote multi-disciplinary collaborations integrating with other e-infrastructures domains of networking and computing.

The FP7 Open Access Pilot: the third call for proposals

A third call for proposals was organised to address the European Open Access (OA) policy initiative.

Launched in August 2008 as a result of an extensive consultation with EU Member States the FP7 OA pilot follows the EU Competitiveness Council conclusions inviting the Commission *"to experiment with open access to scientific data and publications resulting from projects funded by the EU Research Framework Programmes in order to assess the appropriateness of adopting specific contractual requirements*".[4]

A key objective of the FP7 OA pilot is to establish fast and reliable access to EU-funded research results, specifically peer reviewed research articles, in order to drive innovation, advance scientific discovery and support the development of a strong knowledge-based economy.

Under the OA pilot, grant recipients in seven areas of FP7 (energy, environment, health, parts of information and communication technologies, research infrastructures, science in society, and social sciences and humanities) are required to deposit peer reviewed research articles or final manuscripts resulting from their FP7 projects into an online repository and make their best efforts to ensure open access to these articles within either six or twelve months after publication (the so-called embargo period).

The OpenAIRE project, which was selected as a result of this call for proposals, involves 38 partners from all EU Member States plus key European research laboratories. The project is based on the idea of federating national, institutional and thematic repositories that can share resources and information.

This initiative puts Europe in a world leadership position leveraging the high-quality work done in EU Member States and providing an open platform for researchers, students, educators to access scientific information.

Projects supporting the emergence of European Scientific Data e-Infrastructure

Overall, there was a very high response to all calls with oversubscription in the order of 6:1. It was possible to launch 15 projects (briefly described below) corresponding to an overall, investment of 45 million euro of EC contribution.

Many scientific disciplines responded to the calls. As can be seen from the description of the running projects (Box 1) many scientific domains (bioinformatics, astronomy, space-physics, biodiversity, earth observation, chemistry, etc) succeeded in launching relatively short projects (the average duration 2,5 years). In many cases user communities and technology developers address issues of common interest. Some projects are linked with national initiatives that could become European scale collaborations.

[4] Council Conclusions on scientific information in the digital age: access, dissemination and preservation 2832nd Competitiveness (Internal market, Industry and Research) Council meeting Brussels, 22 and 23 November 2007

Box 1 – Short description of projects in the SDI domain

OpenAIRE: The Open Access pilot was launched in August 2008 following the request of EU Member States to experiment with open access to scientific data and publications resulting from projects funded by the EU Research Framework Programmes in order to assess the appropriateness of adopting specific contractual requirements. The OpenAire project will provide an electronic infrastructure supporting mechanisms for the identification, deposit, access, and monitoring of FP7 and ERC[5] funded articles. In order to help attracting users to the platform the project is establishing a European Helpdesk System.

PARSE.Insight: There is a real risk that much of the scientific data and documentation that exists may be lost for future generations unless permanent access is secured. PARSE.Insight is supporting the activities of the Alliance for Permanent Access [of the Records of Science] and aims to highlight the longevity and vulnerability of digital research data.It concentrates on the parts of the e-Science infrastructure needed to support persistence and usability of the digital assets of European research.

DRIVER II: The heterogeneous landscape of scientific repositories in Europe would benefit from efforts looking for interoperation at technical level of many different national or institutional initiatives. DRIVER aims at delivering a pan-European infrastructure federating scientific repositories. It uses open standards and supports complex information objects independent of discipline. Digital repositories will make information - both published and experimental data - widely available and used, supporting new paradigms for research.

SEALS: The Internet became the privileged channel to make information accessible. Discovery and management tools are essential to manage extremely large data sets. The project SEALS is developing an infrastructure to allow the remote evaluation of semantic technologies used in advanced web services. This will allow users to benchmark and compare the most promising technologies and advance the state of the art. The SEALS Platform will be used in two public world-wide evaluation campaigns and the results of these evaluation campaigns will be employed in creating semantic technology roadmaps, identifying sets of efficient and compatible tools for developing large-scale semantic applications.

EuroVO-AIDA: The EURO Virtual Observatory (VO) aims at deploying in Europe an operational technical infrastructure to support shared access to astronomic data. The EuroVO-AIDA project is a main building block of this initiative, unifying the digital data collections of European astronomy. It is committed to move the astronomical European Virtual Observatory into a fully functioning operational phase with global impact.

GENESI-DR: The amount of information being generated about our planet is increasing at an exponential rate.In order to apply it to the global needs relating to the state of the Earth it must be easily accessible. The GENESI-DR project is taking the lead in providing reliable, easy, long-term access to Earth Science data via the Internet. GENESI-DR is contributing to allow scientists from different Earth Science disciplines located across Europe to locate, access, combine and integrate historical and fresh Earth-related data from space, airborne and in-situ sensors archived in large distributed repositories.

IMPACT: Biology is a well-known scientific domain in which ICT is having a huge impact. In recent years it became a high-throughput science with high volumes of complex data being produced and made available to scientists all over the world. The IMPACT project works to improve protein annotation through better coordination and integration of 10 major signature databases that are used by life science researchers all over the world.

METAFOR: Climate Change is in the world's press headlines as a global policy priority for a sustainable development. It is a global social challenge that needs deep scientific insight to find the most cost-effective approaches. The main objective of METAFOR project 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. METAFOR will optimize the way climate data infrastructures are used to store knowledge, thereby adding value to primary research data and information, and providing an essential asset for the numerous stakeholders actively engaged in climate change issues (policy, research, impacts, mitigation, private sector).

NMDB: The worldwide network of standardized neutron monitors measure variations of the primary cosmic rays which are an ideal complement to space based measurements. The Neutron Monitor database (NMDB) project is developing an affordable standard registration system that will submit the measurements to the database via the internet. It will address the problem of different data formats and for the first time will allow the real-time use of cosmic ray measurements for space weather predictions. Besides creating a database and developing applications for them, a part of the project is dedicated to create a public outreach website to inform about cosmic rays and possible effects on humans, technological systems, and the environment.

[5] European Research Council (http://erc.europa.eu/).

PESI: Because the correct use of names is essential for biodiversity management, the availability of taxonomically validated standardised nomenclators (name databases) is fundamental for data infrastructures. 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 European based nomenclators and the network of European based Global Species Databases. It is a standard based, quality controlled, expert validated, open-access infrastructure for research, education, and resource management.

4D4Life: A coherent classification and species checklist of the world's plants, animals, fungi and microbes is fundamental for accessing information about biodiversity. The 4D4Life project will establish the Catalogue of Life as a state of the art e-science facility based on an enhanced service-based distributed architecture. User-driven enhancements in the presentation of distribution data and bio-data will be made. It 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.

VAMDC: A vast range of atomic and molecular (A&M) data is being produced that is of critical importance across a wide range of applications such as astrophysics, atmospheric physics, fusion, environmental sciences, combustion chemistry and in industrial applications from plasmas to lighting. The Virtual Atomic and Molecular Data Centre (VAMDC) project aims at building a secure, flexible and interoperable e-science environment-based interface to the existing A&M data through an infrastructure that is easily tuned to the requirements of a wide variety of users in academic, governmental, industrial or public communities. VAMDC will provide the European and global scientific community with access to a comprehensive federated set of Atomic and Molecular (A&M) data and application resources.

GEOSEAS: In the ocean and marine sector there is a key European initiative operating and further developing a Pan-European infrastructure for managing, indexing and providing access to ocean and marine data sets (SeaDataNet). The Geo-Seas project has taken a strategic decision to adopt the SeaDataNet interoperability principles, architecture and components wherever possible. The overall objective of the Geo-Seas project is to effect a major and significant improvement in the overview and access to marine geological and geophysical data and data-products from national geological surveys and research institutes in Europe by upgrading and interconnecting their present infrastructures.

HELIO: Heliophysics is a new research field that explores the Sun-Solar System 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 will provide the most comprehensive integrated information system in this domain; it will coordinate access to the resources needed by the community, and will provide access to services to mine and analyze the data. 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.

D4Science/D4Science-II: D4Science leverages on existing European e-Infrastructure to establish an integrated platform serving multidisciplinary research. Environmental Monitoring and Fisheries and Aquaculture Resources Management were selected to provide concrete use-cases for D4Science platform. The experience gained with these large communities will be fed into future extensions of the e-Infrastructure to improve their capabilities and respond to the needs of other scientific communities. This is being done by: disseminating the project outcomes, training, exploiting and collecting feedback for the D4Science e-Infrastructure through the implementation of the Virtual Research Environments.

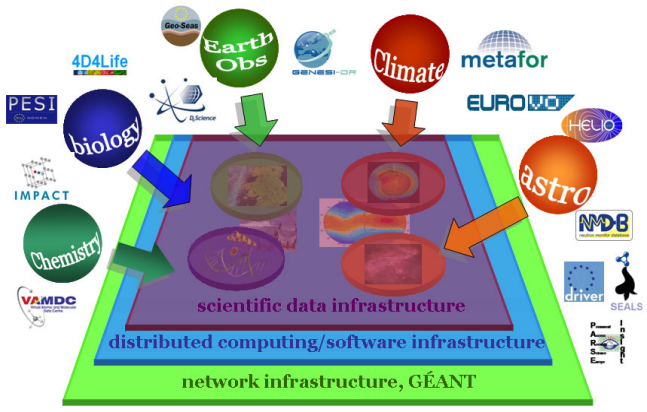

Figure 1 – Landscape of e-infrastructure projects in the SDI area

Figure 1 shows the current landscape of e-infrastructure projects in the SDI area. It highlights the fact that data e-infrastructures use resources and functionalities made available by computing and network infrastructures.

The European Virtual Observatory

Astronomers are turning to the e-Infrastructures paradigm to cope with the progress of astronomical discoveries. The new astronomical data resource forms a Virtual Observatory (VO) which helps them to explore the digital Universe. Alike a real observatory which consists of telescopes, each with a collection of unique astronomical instruments, the VO consists of a collection of data centres each with unique collections of astronomical data, software systems, and processing capabilities. These developments enable making new science.

This is what we have learnt with the EuroVO-AIDA project (Astronomical Infrastructure for Data Access) which is the most recent development of a “Virtual Observatory” (VO) for European astronomers. The project is developing a federated repository of astronomical data conforming to globally-agreed access protocol standards. It represents the transition from an engineering/test phase to a fully-operational public release for the European VO infrastructure.

Conclusions

e-Infrastructures are a key enabler for education and science in the 21st century. The emergence of big and complex-data science is here to stay. It will open completely new ways to dig knowledge out of big amount of information that is becoming available. There are many issues to tackle and as stressed in the communication [1] "*Science will experience major changes in the way it is performed. Researchers will be facing unprecedented levels of complexity in tackling scientific challenges with a global societal impact. Bringing together knowledge from different fields of science will be essential. e-Infrastructures provide the underlying platforms for computationally intensive applications that enable collaboration combining knowledge from different fields of science*".

Projects and initiatives like EuroVO-AIDA are shaping the European Scientific Data e-Infrastructure. There is a need to reinforce, combine and coordinate the efforts at national and EU level to ensure that the most efficient and effective use of the scarce resources while providing a first-class service to the research and education communities.

References

[1] *ICT Infrastructures for e-Science*: Communication from the Commission to the European Parliament and the Council – COM(2009) 108 final

[2] *On Scientific Information in the Digital Age: Access, Dissemination and Preservation*: Communication from the Commission to the European Parliament and the Council – COM(2007) 56 final

[3] P. Wouters - *What is the matter with e-Science*? *Thinking aloud about information in knowledge creation* (http://depot.knaw.nl/3749/)

[4] *eSciDR study: study on the landscape of repositories in Europe which provided recommendations for the SDI area*.