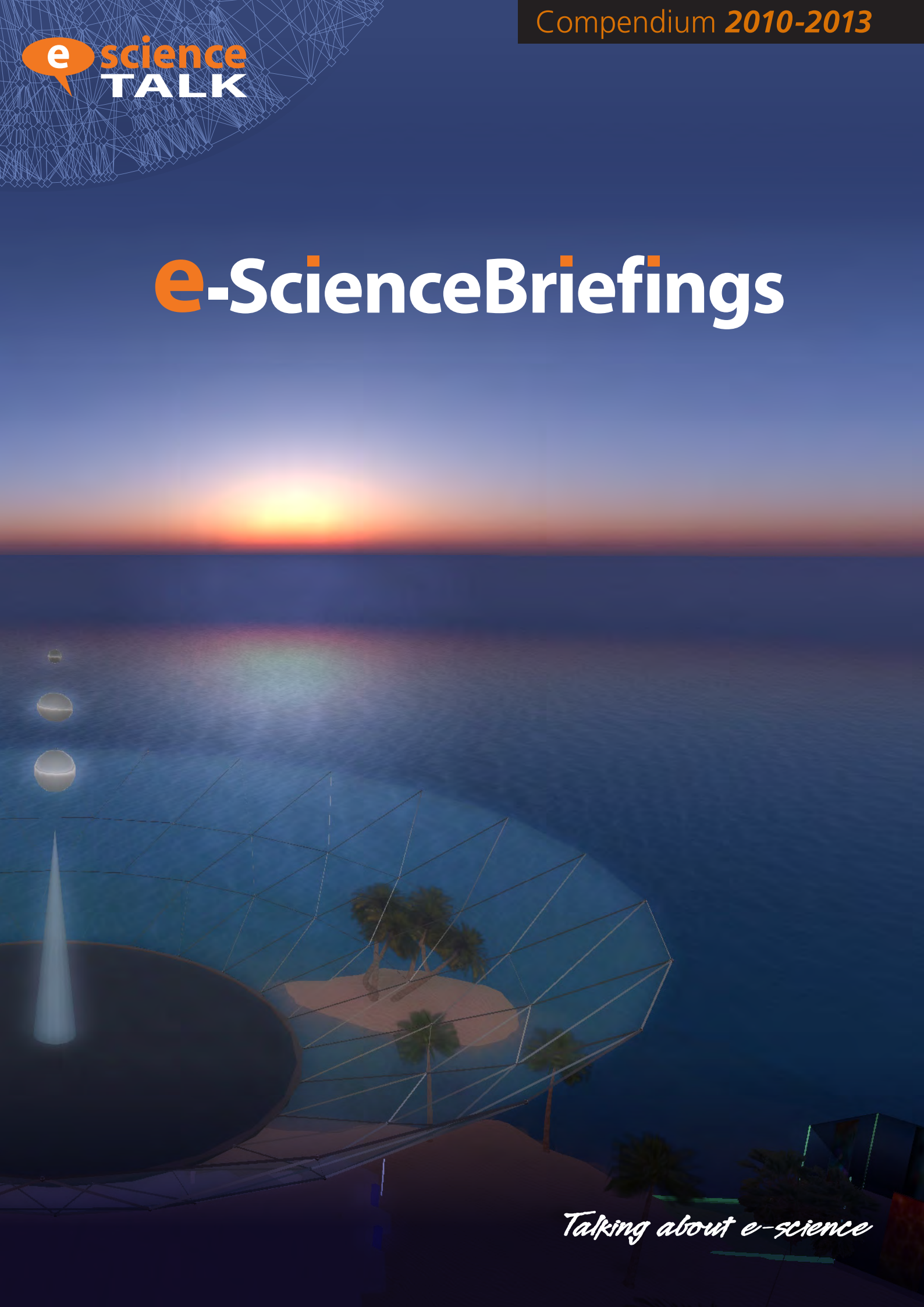
Compendium **2010-2013**

science TALK

e-ScienceBriefings



Talking about e-science

TABLE OF CONTENTS

FOREWORD

On July 4 2012, grid computing reached the headline news, as scientists working on the Large Hadron Collider (LHC) at CERN announced they had discovered a new fundamental particle: the Higgs Boson. The Worldwide LHC Computing Grid was credited as being critical to this global achievement, a message that quickly made its way into the world's press.

The opportunities and challenges associated with e-infrastructures are now very much aligned with those of mainstream research, as more and more research disciplines employ intensive computational methods to cope with the data deluge. Digital science has grown well beyond its origins in the high energy physics domain. Now, the astronomical, life-, and environmental-sciences communities have established a firm foothold in the world of e-infrastructures, and increasingly find themselves working alongside researchers from the social sciences and humanities.

'Big Data' is not just about big volume. It's also about variety, including numerical and text-based data, audio, video and images. Data is also being generated, and being processed and understood, at an ever greater rate. When time-critical events such as natural disasters unfold, scientists are now able to process data in real time and compare events with computer simulations, which, in turn, are more complex and realistic than ever. Here, research is not just impacting on policy, but on real-world actions that affect lives and economies, an example of the three pillars of science, industry and society at work. Through EUDAT, and active involvement in the global Research Data Alliance, efforts are being made to store and categorise vast quantities of data for posterity, providing the exciting possibility that improved analytical methods in the future might extract new discoveries from data recorded today.

Within the private sector, Big Data is driving innovation as surely as it is in research. Cloud technologies have been widely adopted by industry just as grids such as the European Grid Infrastructure have become part of the day-to-day reality of academic research in many fields. But perhaps it is not such a clear-cut distinction - researchers also see the potential of cloud computing for science. The strengthening of public-private partnerships and pre commercial procurement within the next framework programme, Horizon2020, will mean that in future, academia and industry will work more closely together than ever.

As Big Data grows in volume, universal access to publically funded research through projects such as OpenAIRE provides a key to unlocking its potential for the European Research Area. Businesses and governments now increasingly see the value in opening up data for current and future use. As access to data, computation and storage becomes more integrated through EUDAT, EGI and PRACE, providing universal, federated access to these services is now high on the agenda.

One of the most powerful developments in e-Science is the growing citizen science movement, with volunteer desktop grids and online portals allowing the average European citizen to engage with and contribute to research. Just as open access unlocks research findings to the public, so citizen cyberscience allows them to contribute to those findings and develop a sense of ownership. In addition, it will contribute to better acceptance of emerging technologies and therefore to faster innovation. By engaging the public in science and encouraging individuals into science and ICT careers, Europe will be better able to build its human networks and facilitate the free movement of people and data across national borders.

In these pages, you can read about the coordinated approach being taken to e-Science research in the European Research Area, ultimately providing world- class, integrated resources and e-infrastructures that can be shared by all. This is very much a focus of Horizon 2020, the next funding framework for EU research, and this compendium includes a forward look at how this exciting programme is developing.

Together, these e-science briefings show that e-Science in Europe is reaching a real maturity and delivering tangible results, which are in turn promoted and disseminated by projects such as e-ScienceTalk. We can look forward to a bright future on the horizon for European science and society.

Together, these e-science briefings show that e-Science in Europe is reaching a real maturity and delivering tangible results, which are in turn promoted and disseminated by projects such as e-ScienceTalk. We can look forward to a bright future on the horizon for European science and society.

Thierry Van der Pyl,
Director European Commission
DG Communications Networks,
Content and Technology Directorate C – Excellence in Science

**The views expressed are those of the author and do not necessarily represent the official view of the European Commission on the subject.*

Mapping the e-Infrastructure Landscape

Today the World Wide Web provides information for people across the globe but, as yet, no single networked system provides a similar service for researchers to help them access, share, store and process large amounts of data. With this in mind three reports – a "Blue Paper" from the e-Infrastructure Reflection group (e-IRG), a High Level Expert Group report on Scientific Data and the Distributed Computing Infrastructure Collaborative Roadmap – have recently detailed ways in which Europe's e-Infrastructures can work together to ensure a more integrated service. This briefing details the findings of these reports and the actions we can take to provide a harmonised landscape of services for our researchers.

A global vision

Building a coordinated e-Infrastructure landscape is a little like building a railway before matching track sizes were agreed – or even standard time zones! e-Infrastructure providers are currently laying the 'tracks' that will allow researchers to access whatever data or computing power they need, simply and quickly. However, in order for this to work, each provider needs to ensure that their tracks are built to a set of standards, so that the trains that run over them have unimpeded access to the entire network (different e-Infrastructure providers) and don't stop short.

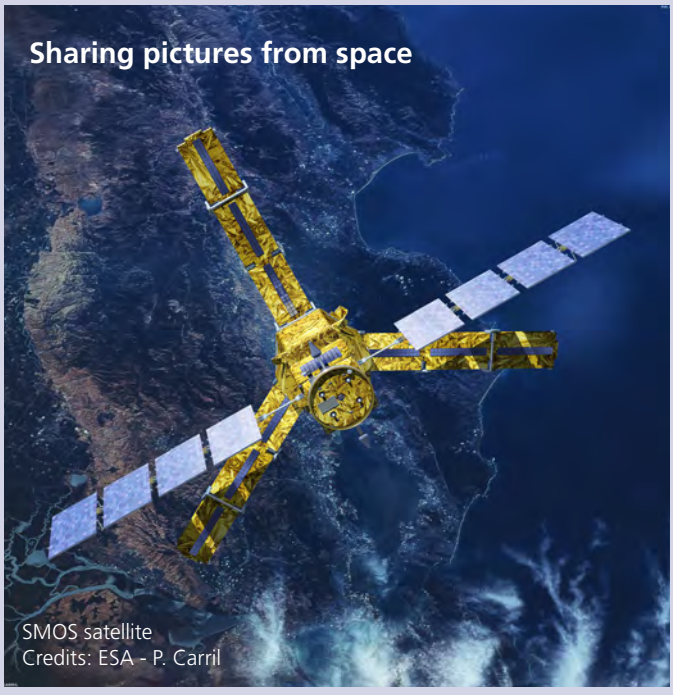
In practice, developing such an integrated e-Infrastructure is much more complicated than making sure that one size fits all – no scientific discipline makes the same demands on the infrastructure as the next. But providers expect that by working together, and pinpointing the areas of common need, they can present researchers in every field with an integrated e-Infrastructure service that can grow and evolve over time to suit their needs.

Neelie Kroes,Vice-President of the European Commission responsible for the Digital Agenda : *"Science has always been based on exchange of information and intense interactions between researchers. Today, thanks to the availability of global communication networks, we profit from truly global and massive scientific collaborations. To this end, the EU's Digital Agenda for Europe has called for the development of research infrastructures and e-Infrastructures, including for scientific data."*

What the experts say

In Autumn 2010 three different reports were released which set out ways in which Europe can achieve its vision of a united e-Infrastructure vision:

The e-IRG Blue Paper: The e-IRG is an inter-governmental policy body comprised of national delegates from more than 30 European countries; it works to define and recommend best practices for pan-European e-Infrastructures.



European satellites collect a wealth of information about Earth from space. The data produced can be used by researchers to gain a better picture of our planet and increase our understanding of key issues such as climate change. But in order for this to happen satellite operators, space agencies and data providers need to work together so that data is accessible and exploitable to those who need it.

Traditionally in Europe there has been poor cooperation in this field - in the past there was no common approach for long-term preservation and access to space data. Now, things are changing.

The European Space Agency, recognising the need for cooperation and sharing has launched a Long Term Data Preservation (LTDP) programme addressing the preservation of all Earth Observation data from Europe and Canada. The availability of this data in the long term supports several applications such as climate change monitoring. The application of a set of LTDP guidelines, approved in Europe in 2009, will help to ensure that data are properly preserved and made accessible to users.

Working with ESFRI

The ESFRI (European Strategy Forum on Research Infrastructures) projects are large research infrastructures that span social and biomedical sciences, earth and physical sciences, energy, infrastructures and analytical facilities. For example ELIXIR, an ESFRI project, will build a distributed but interlinked collection of biological data resources and literature for life science researchers.

The ESFRI projects are likely to be key users of Europe's e-Infrastructure. As such it is vital that providers work together with these projects to understand their needs. Only through close collaboration can we provide them with a useable and useful service. These issues are assessed in the e-IRG Blue Paper. Another report by the EEF (European e-Infrastructure Forum)examined the requirements of the ESFRI-projects and outlined the services and resources that the e-Infrastructure community can offer (see GridBriefing no 13, Future needs of the ESFRI projects).

In October 2009, ESFRI asked the e-IRG to examine ways in which ESFRI projects and their users can engage and exploit common e-Infrastructure services. The resulting Blue Paper, released in September 2010, stressed the importance of bringing communities together in order to improve their mutual understanding and collaboration.

The DCI Collaborative Roadmap: In an effort to explore the collaborative opportunities available within the community, six Distributed Computing Infrastructure projects funded by the EU have set out ways in which they can work together over the next two years. The projects – EGI-InSPIRE, European Middleware Initiative (EMI), Initiative for Globus in Europe (IGE), European Desktop Grid Initiative (EDGI), StratusLab and VENUS-C – cover areas from managing a European Grid Infrastructure through to middleware engineering and cloud computing. This DCI Collaborative Roadmap is a starting point to strengthen potential collaborative opportunities between the funded projects.

High Level Expert Group report on Scientific Data: The High Level Expert Group on Scientific Data was asked by the European Commission (EC) to develop a vision of scientific data e-Infrastructures in 2030. The group's resulting report was presented to the Commission in October 2010.

Together these reports encompass every feature of the e-Infrastructure landscape – from fast networks providing the foundations of connectivity through to shared resources and the researchers using them. More detailed findings on the issues these three reports cover are given in the following section.

Connecting people together

Research networks provide researchers with high-quality internet services, allowing them to connect easily and quickly to collaborators. For example, the GÉANT network, facilitated by the EC through the GN3 project, works with Europe's national research networks to connect 40 million users in over 8,000 institutions across 40 countries.

High quality networks provide the keys to cutting-edge Research Infrastuctures (RIs) and as such must stay accessible and easy-to-use in the face of changing research needs. To aid this, the e-IRG recommends that new RIs participate in networking coordination bodies to define, test and use new networking services.

As research resources become more interconnected, aligning authentication and authorisation across infrastructures is also vital. Future pan-European e-Infrastructure and ESFRI projects are encouraged to define their access control policies and mechanisms from the start, in accordance with the standards and best practices adopted by the research community.

The development and spread of remote instrumentation techniques and technologies will also open new opportunities for scientific communities. Sharing expensive scientific equipment such as radio telescopes and synchrotrons through remote use can cut costs and open up the facilities to more researchers. An increasing reliance on remote instrumentation has been identified as a particular issue for the environmental science ESFRI projects. To ensure that these tools can be used and accessed across all of Europe's e-Infrastructure, the e-IRG recommends developing standard interfaces for these technologies.

Steven Newhouse, EGI - *"The EGI-InSPIRE project is establishing mechanisms to bring new innovative technologies into the European Grid Infrastructure in order to support the innovative research taking place within the ESFRI and other research communities. The results from the other DCI projects and activities taking place within the National Grid Initiatives will provide a platform for the development of innovative software services and tools to help support distributed data analysis."*

Riding the data wave

How to tackle the growing amounts of data being produced worldwide is one of our biggest challenges. Data initiatives must work out how to store, access and preserve data for researchers in the decades to come. In response to this, the EC asked a High-Level Group on Scientific Data to look at the challenges and benefits resulting from the rise in data. Their findings are detailed in the recent report 'Riding the Wave: How Europe can gain from the rising tide of scientific data'.

In the report the Group detailed the potential benefits of developing an e-Infrastructure for data. It can allow different domains to collaborate and enable the use, re-use and combination of data while still maintaining the data's integrity and ownership. But to enable this to happen they recommend the following:

- The EU to should develop a framework for a collaborative data infrastructure.
- Additional funds should be found to develop such an infrastructure.
- Incentives and rewards for data sharing should be introduced, as well as ways to measure data value.
- Researchers should be trained so they recognise the importance of sharing data.
- Create incentives for green technologies within data infrastructures.
- Establish a high-level international group to plan for data infrastructure.

Monica Marinucci, Oracle - *"Technology advancements, networking capabilities and tighter integration of ICT in the scientific process are opening up new frontiers for science and enabling scientists to explore new ways of doing research. It is crucial for the future of science and economy to set up shared research infrastructures and to consolidate basic research services that will allow scientists and their institutions to do better and innovative research in an efficient, scalable and collaborative manner. Industry should support this process by providing cost effective, reliable and open solutions for the whole research lifecycle. Innovation and technology produced then translates into technology progress, jobs and wealth for the society at large."*

Up in the clouds

The VENUS-C project is set to explore how cloud computing can be used in European scientific settings. It will initially look at applications ranging from biomedicine, civil protection and emergencies, civil engineering and data for science, expanding its scope over time.

Aquamaps, from the D4Science project, is one such application which hopes to benefit from VENUS-C. Aquamaps uses environmental information, such as sea water salinity and temperature, along with data on fish environments to map the biodiversity of our oceans. This gives the fishery community a way to predict the existence of fish anywhere in the world at any given time. They can also gauge the impact of changing climatic factors as well as of pollution, natural and man-made disasters. By moving these calculations onto the cloud, VENUS-C hopes to provide a quicker way to map species, and ultimately provide scientists and decision makers with more information, faster.

VENUS-C will work with other DCIs in a number of activities. For instance, security and accounting have been identified as two potential fields of cooperation with EMI.

Andrea Manieri, VENUS-C *"Cooperating with other European Distributed Computing Infrastructures offers an important opportunity to share knowledge on an evolving landscape. For VENUS-C, it is also an opportunity to evaluate how we can capitalise on European investments and expertise from a cloud computing perspective. By standing on the shoulders of giants, understanding the driving forces, costs and impact, VENUS-C will also play a part in defining a cloud computing strategy for European scientific communities."*

Neil Geddes, e-IRG delegate for the UK and chair of the Blue Paper editorial board - *"The Blue Paper represents one step on the road towards better exploitation of the opportunities presented by advanced ICT and computing for the European research communities. It acknowledges the importance attached to this area by both ESFRI and the e-IRG and illustrates the growing collaboration between these two groups. Both the e-IRG and ESFRI are prepared to transform the common findings stemming from the Blue Paper into actions supporting the ESFRI Roadmap projects in building the European Research Area."*

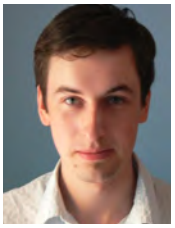
Supercomputers for super research

The Partnership for Advanced Computing in Europe (PRACE) provides Europe with world-class high performance computing (HPC) systems and service. Twenty European states are members of the PRACE association which is managed as a single European legal entity. It represents the top of the European HPC ecosystem and works closely with other European e-Infrastructure projects such as DEISA, EGI and HPC-Europa2.

credit: FZJ

PRACE's target is to offer state-of-the-art super-computing systems (Tier-0) to the European scientific communities. In the medium term, each system will provide researchers with several petaflops of computing power and by 2019 the aim is to reach exascale computing. PRACE is well integrated into the European HPC ecosystem and will include services presently offered by DEISA as of 2011.

At the moment PRACE has two Tier-0 machines (JUGENE in Germany and Curie in France).



David O'Callaghan, Trinity College Dublin - *"We need to work together to share requirements and results between projects and in particular to benefit from the strengths and specialities of national and international initiatives. StratusLab will work with EGI and the other distributed computing infrastructure projects to provide users with the mix of cloud, virtualisation and grid technologies they need to do research effectively."*

Powering research

Whether using a single PC, large-scale distributed data processing, or the world's fastest supercomputer, modern researchers invariably require access to computing power. In Europe, two international computing infrastructures – the European Grid Infrastructure (EGI) and the Partnership for Advanced Computing in Europe (PRACE) – are managing the development of grids of high throughput computing (HTC) and high performance computing (HPC) services, respectively.

For HPC to support researchers far into the future, the e-IRG recommends closer collaboration with users to better understand researchers' requirements of HPC and the opportunities it can deliver to them. But another challenge is on its way; in the next decade we're expected to reach exascale computing. This will provide computing a thousand times more powerful than now. But for these machines, simply scaling up existing software is not enough - it will have to be redesigned to work at such fast speeds.

In the case of distributed computing – whether delivered by grids, clouds, volunteer computing or shared data centres - stronger collaborations are needed between grid and cloud infrastructure users and resource providers. Six Distributed Computing Infrastructure (DCI) projects have detailed ways in which they expect to work together over the next five years in the DCI Collaborative Roadmap. Key areas for collaboration include providing each other with dissemination and training, and working together to define standards for interoperability.

Give the people what they want

Users and their needs are key when developing any e-Infrastructure. As a general, but important, message of the e-IRG Blue Paper, the e-IRG recommends that RI, e-Infrastructure and user requirements should evolve in tandem and collaboration between RI and e-Infrastructures at all levels should be actively supported. The annual EGI User Forum, for example, provides an opportunity for users to talk to e-Infrastructure providers.

Researchers are often grouped into Virtual Research Communities (VRCs), or Virtual Organisations (VOs) which rely on ICT to allow a group of geographically dispersed researchers to work together. The e-IRG recommends that VRC developments should proceed gradually, starting with domain-specific shared access to distributed resources, and expanding to integrate different research activities.

For more information:

DCI Collaborative Roadmap: ***https://documents.egi.eu/document/172***

e-IRG Blue Paper: ***www.e-irg.eu/images/stories/eirg_bluepaper2010_final.pdf***

High Level Expert Group report on Scientific Data: ***http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204***

e-IRG: ***www.e-irg.eu***

GÉANT: ***www.geant.net***

PRACE: ***www.prace-ri.eu***

VENUS-C: ***www.venus-c.eu***

EGI: ***www.egi.eu***

iSGTW: ***www.isgtw.org***

e-ScienceTalk : ***www.e-sciencetalk.org***

e-ScienceTalk is co-funded by the EC under FP7

Supercomputing: Empowering research

Computer simulations have evolved as an essential part of scientific research, complementing theory and experiment. Scientists and engineers use simulations when problems are complex, or experiments are too dangerous, expensive, or impossible to carry out. Today supercomputing has simulated rat brains, tested engineering structures and modelled global warming. Investment in these technologies can lead to real economic benefits while ensuring researchers remain internationally competitive.

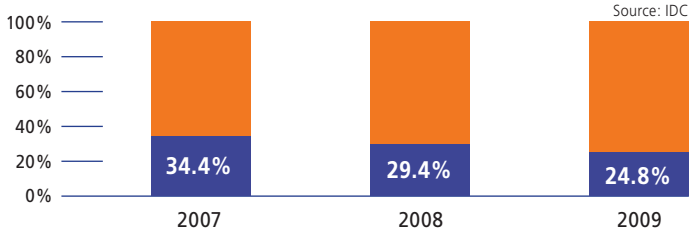
What is a supercomputer?

Unlike distributed grids, where thousands of networked computers are linked together across many geographical locations, a supercomputer consists of thousands of integrated computers in a single room. These supercomputers are at the frontline of current processing capability, particularly speed of calculation.

A supercomputer's power is measured by how many floating point operations it can complete per second (flop/s). The world's fastest machines are listed in the biannual Top500 list. In November 2010, the Chinese Tianhe-1A system at the National Supercomputer Center in Tianjin topped the list, with a record peak performance of 2.57 petaflop/s - the equivalent of everybody on the planet doing 367,000 calculations a second.

Supercomputers are used for highly complex problems in which many intermediary results are dependent upon one another. Therefore supercomputers are often used for highly calculation-intensive tasks such as those found in astrophysics, climate modelling and biomedicine. The use of supercomputers is commonly referred to as 'high performance computing' or HPC.

European percentage of worldwide HPC revenue



A study in the US has shown that that investment in HPC increases research competitiveness at academic institutions. In Europe there is a wealth of HPC-related experience and talent. However a 2010 IDC report for the European Commission found that recently Europe has fallen behind other regions. In order to fully benefit from HPC the report recommends that we need to both increase its HPC investments and find ways to apply HPC in a more productive and innovative manner.

Unravelling the mystery of Dark Energy

Friedrich Röpke's research group at the Max-Planck-Gesellschaft, Garching, Germany is using PRACE's JUGENE Tier-0 supercomputer to study type Ia supernovae explosions. They hope their work will contribute to a better understanding of mysterious dark energy in the universe.

Type Ia supernovae are exploding stars that are among the brightest objects in the universe. They are an important phenomenon, as scientists can use them to infer cosmic distances, by looking at their apparent brightness.

Credits: NASA/ESA/JHU/R.Sankrit & W.Blair

Distance determinations derived using type Ia supernovae show that the universe is expanding at an accelerated rate. However the reason for this effect is still unclear. Researchers attribute it to mysterious 'Dark Energy' which they believe forms the main constituent of the universe today.

Type Ia supernovae distance determination can contribute to a better understanding of this mysterious energy form. The necessary distance measurements, however, require high precision and great observational effort. Therefore we need to match this with a sound theoretical understanding of the supernova explosions.

Röpke's group aims to shed light on the physics of the explosion mechanism. But the explosion physics are complex and can only be studied in sophisticated numerical simulations which require the huge amounts of computational power provided by PRACE.

Thomas Eickermann, PRACE - *"European researchers and engineers need access to world-class supercomputer resources and services to remain internationally competitive. Therefore, Europe needs to continuously invest in this area to provide a leading-edge HPC infrastructure. Strategically, Europe also needs to strive for independent access to supercomputing, as it already has for aviation and space."*

Grids of supercomputers

In Europe, PRACE (Partnership for Advanced Computing in Europe) has created a high-end supercomputing infrastructure. Together with grid facilities, data resources, software and experimental facilities, the PRACE Research Infrastructure will meet the needs of scientific and industrial domains in Europe. Currently, PRACE has two world-class supercomputer (Tier-0) systems: JUGENE, at FZJ, Jülich, Germany and CURIE at CEA, Bruyères-le-Châtel, France. A third Tier-0 system, SuperMUC at Leibniz Supercomputing Centre, Munich, Germany is underway.

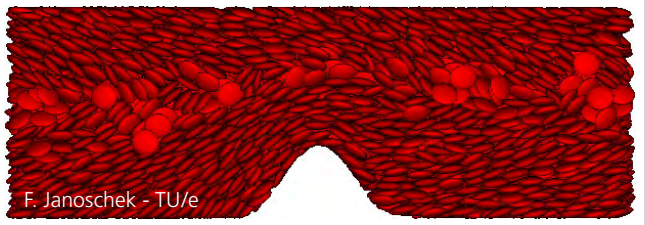
PRACE collaborates closely with DEISA, the Distributed European Infrastructure for Supercomputing Applications, which currently provides access to eleven European national supercomputing centres (Tier-1 systems). User access and coordinated operation of PRACE centres employs software and services that have been pioneered and deployed by the DEISA project. When DEISA ends in 2011, PRACE plans to integrate current DEISA services into its own infrastructure. PRACE also has close links with EGI, the European Grid Infrastructure, to ensure grid access and interoperability today and in the future.

Simulating blood flow

Federico Toschi and his team at the Eindhoven University of Technology are using PRACE supercomputers to learn more about blood flow in the body.

Studying this phenomenon is not easy as it requires blood flow simulations which have to take into account interactions between large numbers of cells in a variety of geometries. Previous models simulate blood as either a homogeneous fluid or as a relatively small number of red blood cells, but the latter are not able to reach relevant time or length scales.

The team in the Netherlands has developed a model which allows them to perform simulations of millions of blood cells over more realistic times and distances. They have been awarded 39 million compute hours on PRACE resources to run their model, and ultimately learn more about fatal and serious blood clots.



In the US, The National Science Foundation's TeraGrid integrates high-performance computers, data resources and tools, with high-end experimental facilities around the country. Currently, TeraGrid resources include more than a petaflop of computing capability and more than 30 petabytes of online and archival data storage, with rapid access and retrieval over high-performance networks. Researchers can also access more than 100 discipline-specific databases.

John Towns, TeraGrid Forum Chair - *"The next phase of the US National Science Foundation's (NSF) investment in high-end computing will be launched in 2011: the eXtreme Digital, or XD, program which will succeed TeraGrid. Many of the existing resources will continue, and new ones will be added."*

Supercomputing challenges

Initiatives such as PRACE, DEISA and TeraGrid help to expand scientific knowledge and in turn deliver social and economic benefits. However, the use of supercomputers raises a number of issues:

- **Cost:** Supercomputers are expensive. The Chinese Tianhe-1A system for example is estimated to have cost 600 million yuan (68.7 million euro). But in such a fast moving field, these machines can rapidly become out-of-date. The number one computer in the Top500 list is rarely number one the following year. However such investment is necessary if we are to benefit from HPC science and industry research.
- **Access:** Supercomputers are powerful scientific instruments that can be used by researchers across the world. In the US, for example, supercomputers at the National Center for Supercomputing Applications (NCSA) are used remotely by more than 1,500 scientists and engineers. However, in practice, research time on these machines is limited. Supercomputing providers need to find fair ways to allocate time on their machines. PRACE, invites researchers to submit proposals which are subject to both technical and scientific peer review. TeraGrid also allocates resources via a quarterly review process, with successful proposals receiving allocations of high-performance computing, scientific visualisation, data storage, and advanced user support for 12 months at a time.

Elizabeth Cherry, Rochester Institute of Technology - *"Myself and colleagues from Cornell University and the Max Planck Institutes in Germany used TeraGrid resources to develop an effective and potentially less painful defibrillation method for treatment of atrial fibrillation, a heart disorder that affects millions of people. Cardiac modelling demands supercomputing because it requires a great range of scales - both spatially, from a single cell to the full-size heart, and over time, from microseconds to minutes. Our supercomputer simulations have shown the feasibility of our low-energy approach."*

- **Efficiency of use:** The TOP500 lists sites according to the Linpack benchmark - a 'horsepower number' which correlates to the machine's ability to perform calculations. However, harnessing that power to solve real world problems efficiently is still a major engineering challenge, requiring a change in algorithmic and programme design to fully utilise future multi- and many-core processors. PRACE experts continuously work on scientific software packages to make best use of these new computing engines as they are deployed in Europe.

Daniel Ahlin, PDC, KTH - *"I expect HPC power efficiency to be an extremely important field in the next few years since the environmental and power cost problems it tries to mitigate are critical issues, not just for HPC, but for society as a whole. A good, easy-to-reach and high-yield example is heat-reuse which has just begun to be explored more widely."*

- **Energy consumption:** As supercomputers become more powerful their energy consumption becomes a real issue. Energy costs make up a substantial proportion of supercomputing costs – cooling processors is a particular problem. PRACE is engaged with several green IT prototypes, including its new Tier-0 system, SuperMUC. SuperMUC will use water to cool its processors and memory, resulting in a 40% energy reduction compared to an air-cooled machine. The warm water can also be redirected to the heating systems of nearby buildings. As the trend towards more powerful computers continues the Green 500 list, which ranks computers according to their energy efficiency, is expected to become more important than the Top500 list.

- **Training:** In order to achieve the best results from these instruments, researchers need to be trained to use supercomputers. Most supercomputing providers offer this, enabling researchers to make the most out of these machines. TeraGrid offers an abundance of HPC training opportunities and PRACE has also developed an extensive training programme, including seasonal schools throughout Europe.

Pekka Manninen, PRACE training coordinator CSC - *"Using Tier-0 systems efficiently is highly non-trivial; in order to benefit from the competitive edge inherent in the infrastructure, our users must be skilled enough to use it properly. The PRACE training scheme builds a permanent network for training in the field of Tier-0 computational science. In the First Implementation Project (PRACE-1IP), the most visible actions are a top quality face-to-face training event curriculum and the establishment of an online training portal together with various education outreach activities."*

Reaching the exascale

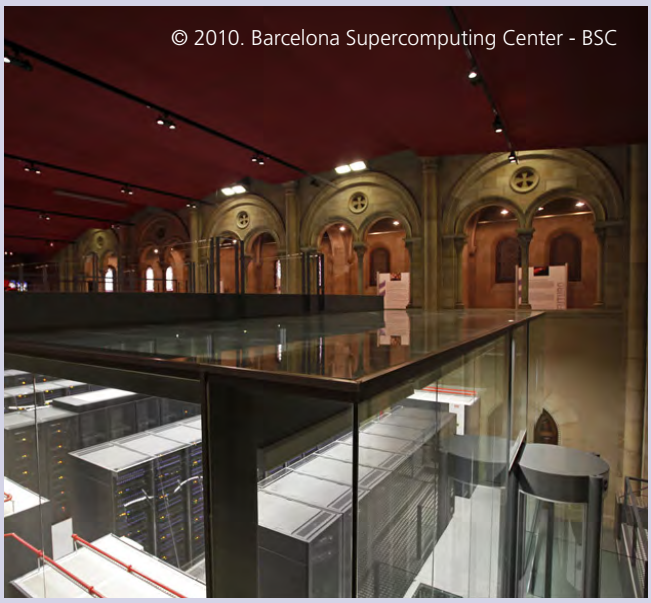
These are not the only challenges. Petascale technologies have allowed us to simulate protein folding and model climate data. However the complexity of these scientific models is such that to truly understand them we require exascale computing – a thousand times more powerful than the current top of the line petascale computers.

Rosa M. Badia, Barcelona Supercomputing Center - *"Supercomputers are essential tools for today's research in physics, weather forecasting, molecular modeling and many other disciplines. However, the complexity of programming such computers prevents efficient use of them. Our research in programming models enable other scientists to boost their computations in MareNostrum (the supercomputer at the Barcelona Supercomputing Center) and promote the advancement of their research."*

The most beautiful supercomputer in the world?

When the Spanish and Catalan government needed a location for one of the fastest computers in Europe as well as an innovative research center, they came up with a novel solution.

Tasked with building a new supercomputer in just four months, the newly created Barcelona Supercomputing Center had a tough challenge on their hands. Surprisingly the best place was a chapel, Torre Girona, located at the North Campus of the university in Barcelona, which had enough interior space to fit the new computer, named MareNostrum.



Today, after an increase in capacity in November 2006, MareNostrum is the 118th most powerful computer in the world, as named in November 2010's Top500 list. It aids research in areas such as computer sciences, life sciences, earth sciences and computer applications in science and engineering.

Moving towards exascale computing will not only exacerbate issues of cost, access and training; it will raise entirely new problems. Whereas in the past there was scope to scale up, we are now reaching fundamental physical limits of technology. Current machines cannot simply be made bigger to achieve exascale processing – entirely new technologies are required. Exascale computers will also need an energy capacity which cannot be delivered by any computer centre today.

Engineers already acknowledge that exascale computers will have to rely on parallelism, along with compatible algorithms and software. Because memory bandwidth and capacity do not follow Moore's Law, developers will have to become better at managing the constraints of less memory per core and investigate novel communication-avoiding algorithms.

Sebastian von Alfthan, CSC – IT Center for Science - *"Exascale computing brings major challenges that need to be met on the hardware, software and programming model level; but also opportunities for groundbreaking research at an unprecedented scale. Addressing these challenges provides Europe a chance to take a greater role in shaping future HPC technologies. "*

An expected 100 million cores will be needed to deliver an exaflop but at such large scale, frequent hardware failures will be inevitable. Application developers must embrace this and factor it into algorithms and software. At the same time physically interconnecting 100 million cores will not be trivial. Network reliability, intrinsic latencies and message size will all need to be considered when designing the next generation of supercomputers. And even when solved, it is highly unlikely that one architecture will be suitable for all applications.

With so many hurdles in the way, exascale computing will need dedicated time, effort and money. Experts believe that we could see the first exascale systems by 2018 but warn that they will require an estimated $1 billion of investment. However this investment could lead to real economic benefits.

Glossary

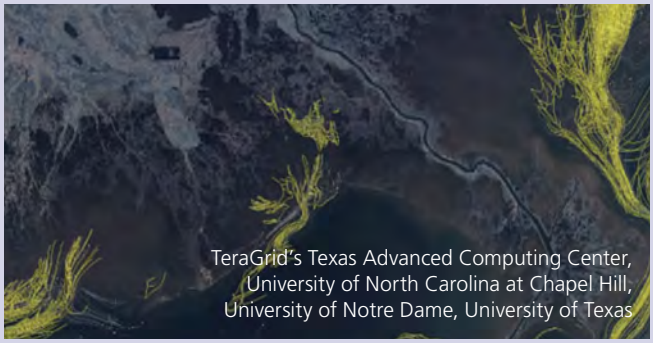
Parellelism – a type of computing which carries out many small tasks concurrently.
Memory bandwidth – the rate at which data can be read to or from memory by a CPU.
Moore's Law – states that the number of transistors on a chip will double about every two years.
Core – Part of a processor, which reads and executes instructions independently of other cores on the same chip.
Latency - the delay experienced in transmitting signals from one part of a computer to another.

Scan this QR code into your smart phone for more on this e-ScienceBriefing

Cleaning up the spill

When the blowout of the BP Macondo well destroyed the Deepwater Horizon oil rig in the Gulf of Mexico on April 20, 2010, it sparked a massive environmental disaster. TeraGrid resources were ready to assist in the mitigation process.

Clint Dawson, head of the Computational Hydraulics Group at The University of Texas at Austin's Institute for Computational Engineering and Sciences (ICES) received an emergency three-million-hour allocation from the NSF to computationally model the spill. Using the Ranger supercomputer at the Texas Advanced Computing Center, Dawson's group created high-resolution forecasts showing how the spill might migrate over a 72-hour period and affect large parts of the coastline. These models helped response teams position booms and prepare for the possibility of a natural disaster – a hurricane - which could compound one of the nation's worst man-made disasters.



TeraGrid's Texas Advanced Computing Center, University of North Carolina at Chapel Hill, University of Notre Dame, University of Texas

Separately, the Louisiana Optical Network Initiative, another TeraGrid partner, provided the National Oceanographic and Atmospheric Administration's surveillance team with critical high-bandwidth network connections to quickly transfer large amounts of vital data between the spill site and Washington D.C.

For more information:

Barcelona Supercomputing Center: ***www.bsc.es***
DEISA: ***www.deisa.eu***
HPC in Europe (with link to IDC report): ***www.hpcuserforum.com/EU***
PRACE: ***www.prace-ri.eu***
Study into HPC and US research competitiveness: ***www.jiti.net/v10/jiti.v10n2.087-098.pdf***
TeraGrid: ***www.teragrid.org***
The Green500: ***www.green500.org***
TOP500: ***www.top500.org***
EGI: ***www.egi.eu***
iSGTW: ***www.isgtw.org***
e-ScienceTalk : ***www.e-sciencetalk.org***

e-ScienceTalk is co-funded by the EC under FP7

Cloud computing: What's on the horizon?

Cloud computing is making the revolutionary collaboration models of today's European e-Infrastructure more broadly accessible and applicable. Clouds can allow businesses, governments and educational institutes to access services and data on demand, and pay for what they need at the point of use. For those who need large amounts of computing power for short periods of time, the cloud appears to be the perfect solution, as well as a useful complement to existing e-Infrastructures.

A European vision

Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda - *"Normally I prefer clearly defined concepts. But when it comes to cloud computing I have understood that we cannot wait for a universally agreed definition. We have to act."*

Clouds for storage and computing are changing the way that businesses, governments and academia deal with computing services. As such, Europe is keen to be at the forefront of developing this technology. According to Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda, Europe should not just be 'cloud friendly' but 'cloud active'. At an address given to the World Economic Forum at Davos in January 2011, she outlined plans for an EU-wide cloud computing strategy.

The many faces of cloud computing

Cloud computing is still an ill-defined term and can be used to describe any of the following:

- **Infrastructure as a Service** (IaaS) - buying access to raw computing hardware over the internet, such as servers or storage. Also known as utility computing. IaaS often employs virtualisation in which users can create their own 'virtual computer', with specified applications, software and operating machine to deploy in the cloud. Advantages are that users don't need to worry about what software is being run where but disadvantages include that this is difficult to do.
- **Platform as a Service** (PaaS) - developing applications using web-based tools so they run on systems software and hardware provided by another company. The Google App Engine is an example of PaaS.
- **Software as a Service** (SaaS) - using a complete application running on someone else's system. Web based email is an example of this.

This will aim to look at the legal framework for cloud computing, resolve technical issues and establish a more open and competitive market for IT services.

A report from the EC entitled 'The Future of Cloud Computing' recommends Europe should exploit the available expertise and results from areas such as grid and other e-infrastructures to help realise the next generation of services on cloud systems. The EC intends to hold a series of discussions with both cloud providers and users with the aim of developing a plan for future actions by 2012.

Clouds, grids and bioinformatics

StratusLab is applying cloud technologies to grid sites. By taking existing grid sites and running them on top of a cloud, StratusLab is simplifying the management of grids.

Grid sites at academic institutes can benefit from StratusLab, but these are by no means their only customers. The cloud distribution platform developed by the project is open for use by anybody who wants to outsource IT services. The StratusLab toolkit can provide increased flexibility thanks to virtualisation. Users often need applications with specific software requirements for their experiments.

StratusLab's ability to pre-define very specific and well suited virtual machines and to make them available in central repositories ready for deployment on academic clouds, would satisfy even the most demanding user. For example, StratusLab has already had interest from the French bioinformatics network ReNaBi.

Clouds for science

Clouds are increasingly being used by businesses, governments and scientists in areas from astronomy to zoology. Clouds can give scientists added flexibility for both storing and addressing data which can be easily accessed regardless of time or location. They can prove particularly useful for small research groups who require additional IT services.

Guy Coates, Wellcome Trust Sanger Institute - *"The life sciences community is reeling under the 'data deluge' generated by high throughput DNA sequencing. Sequencing technologies and data-production rates are rapidly evolving, placing IT organisations under increasing strain. The large data volumes produced require sophisticated data-tracking and large scale storage technologies, whilst analysis of the data needs access to HPC resources. On-demand cloud based IaaS and SaaS services from both academic and commercial providers can meet these data management and HPC requirements, and can react quickly to the rapidly changing needs of next-generation sequencing."*

A number of European Commission funded projects are demonstrating how academic and research communities can best profit from cloud services. For example the VENUS-C project brings together industrial partners and scientific user communities to develop, test and deploy an industry-quality cloud computing service for Europe.

The RESERVOIR project is also enabling massive scale deployment and management of complex IT services. The OpenNebula open source toolkit, a key outcome of this project, is being used as an open reference stack for cloud computing by several new European projects including StratusLab, BonFIRE and 4Caast.

Complementary technologies

Clouds are just the latest in a range of computing technologies such as grids and supercomputing which are available to researchers.

While clouds are sometimes seen as the successor to grids, the two technologies can complement each other as they offer users different benefits. For a full discussion of the differences between grids and clouds see the GridBriefing 'Grids and clouds: The new computing' (January 2009).

Clouds can provide significant added value to the existing range of computing resources. External parties can gain access to data grids and supercomputers through secure and paid access services via the cloud model. Integrating clouds with other forms of computing is an opportunity for development of both technologies. For example StratusLab aims to develop and deploy cloud technologies with the aim of simplifying and optimising the use and operation of distributed computing infrastructures such as the European Grid Infrastructure (EGI).

Charles Loomis, Linear Accelerator Laboratory - *"Cloud is a very selfish technology. Individuals acquire resources, use them for their work and get rid of them when they're finished. Grids are more about collaboration – sharing data and experiences. One of the challenges for cloud is how to enable federation, for example, how to move around data in an academic environment."*

Moving components onto third-party clouds can be useful for a number of reasons including added resilience against events such as local power loss at sites. In the UK, GridPP, the grid for particle physics, has demonstrated that some grid components that are usually run locally can be successfully moved onto third-party clouds.

Building with clouds

The VENUS-C project is providing a platform for scientific applications, striving towards interoperable services which avoid vendor lock-in. VENUS-C supports a host of research including drug discovery, marine biodiversity and even predicting the risk of wildfires.

One of VENUS-C's most promising developments is a platform that can perform static and dynamic analysis of building structures using the cloud. This will involve running multiple simulations of buildings under a variety of conditions – be it different earthquake loads or building materials – leading to a reduction in construction costs and response time.

By working with VENUS-C, users have the opportunity to be the first to experiment with a cloud platform at industry-level quality. They gain direct access to developers, infrastructure experts and the computing platform until June 2013. They can test and improve their applications and business models and be prepared for a sustainable exploitation after the project ends. VENUS-C is an important opportunity to learn, test, share knowledge and gain visibility for research in the cloud.

Working together

Standards will play an increasingly important role in the drive towards an interoperable system of e-Infrastructures to ensure Europe retains its world-class position and plays a leading role in tackling global challenges.

The SIENA initiative is focused on accelerating and coordinating the evolution of interoperable distributed computing infrastructures by fostering an open and constructive dialogue with industry, standards bodies and major stakeholders. SIENA will build consensus on the best practices for cloud and grid technologies which are driving the development of standards. With the advent of cloud computing in both commercial and scientific settings, interoperability is becoming key to success, while the credit-squeezed economy puts additional pressures on eGovernment and research communities to ensure value for money from the public purse.

Silvana Muscella, Technical Director of SIENA - *"In collaboration with the European Commission, SIENA is working with major stakeholder communities on a roadmap focusing on interoperability and standards. The roadmap will define scenarios, identify trends and investigate the innovation and impact sparked by cloud and grid computing, delivering insights into how standards and the policy framework are shaping current and future development and deployment in Europe and globally. The development of the roadmap is timely in support of the aims of the Digital Agenda for Europe, which reinforces the need for effective interoperability between IT products and services to build a truly digital society by 2020."*

Addressing user concerns

The fast pace of cloud development has raised a number of technical and commercial questions, such as security, access and availability of cloud services. Unlike collaborative infrastructures such as grids most cloud providers have unique and proprietary application programming interfaces (APIs). However to gain the best deals, users would like to be able to move their data around from provider to provider. Provider lock-in is a major concern.

Christine Morin, Project Coordinator Contrail - *"Contrail aims at removing some constraints that are currently present in cloud systems. It will federate clouds allowing companies and research organisations to easily switch from cloud providers. Contrail will develop a complete, secure cloud computing stack, including Platform-as-a-Service and Infrastructure-as-a-Service."*

The Contrail and mOSAIC projects are both exploring ways of making choosing a particular cloud easier. The mOSAIC project, especially, hopes to tackle the issue of provider lock-in. It will develop a platform to demonstrate how users can determine and access the cloud provider that offers the best services for them.

The overall quality of service such a multi-cloud environment offers can be hard to predict. With single clouds risks can be mitigated through Service Level Agreements (SLAs) between users and providers. However, managing SLAs between multiple parties requires applying lessons already

Improving drug research

Contrail is investigating how cloud computing could help cut costs and increase the impact of drug research.

The massive explosion in the volume of data generated through genomic research, pharmacological sources and clinical trials has greatly increased the number of potential drug compounds. However finding promising candidates using only traditional computing technology is hugely challenging. The cloud makes it possible to use more computing and data storage power at the same cost.

Accessing and analysing the data through the cloud, and drawing on the unused computing resources of other companies or organisations, could potentially lower the cost of commercial electronic drug discovery services. It could also enable SMEs to play a competitive role in the pharmaceutical industry.

learned from other e-Infrastructures. For example the gSLM project analyses grid environments to come up with recipes and best practices that address this challenge.

Bob Jones, Chair of the European e-Infrastructure Forum - *"For scientific research communities, the question is not whether they will use cloud computing but rather how quickly and under what conditions."*

Legal and security minefields

By putting data into the cloud, users can lose control of where it is stored. Cloud providers operate across or in many jurisdictions - data centres span countries, and even continents. In these cases determining which country's laws apply to stored data is not straightforward.

For scientists working with sensitive data such as health records, a private or community cloud may be a better solution to commercial cloud providers as users can be sure of where data is being stored. Otherwise legal issues such as data protection, privacy and user's rights will need to be factored into cloud provisioning.

Gudmund Høst, e-IRG Chair - *"The e-IRG has had a focus on cloud computing since 2008, where the e-IRG White paper 2009 made a first comparison of clouds and grids. The e-IRG roadmap also gave recommendations on commodity and cloud computing 'recognising the importance of transparency and compatibility of the organisational and financial models as key factors in maximising the benefits of broad commodisation of computing services for all scientific users.' The e-IRG Blue paper 2010 to ESFRI touched on the issues of grid, cloud and virtualisation proposing 'collaboration among grid and cloud infrastructure providers and users to raise awareness of the range of available technologies and how to best use them'. The current e-IRG White Paper 2011 is also considering cloud computing as a natural element within the new service-oriented e-Infrastructure."*

Groups such as the e-IRG (e-Infrastructure Reflection Group) have been looking into both legal and governance issues for e-Infrastructures. Regarding security, the Cloud Security Alliance works to promote the use of best practices for providing security assurance within cloud computing. Events such as Cloudscape also provide an opportunity for experts, developers and end users to have meaningful discussions on the subject

Governing from clouds

A number of countries are employing the cloud model for use in government. The US has employed clouds for transparent government, while the EU uses the technology for public procurement. In Japan clouds are being deployed to aid collaboration across ministries. This shift to indirect governance of data through clouds leads to a number of specific challenges.

Governmental data could be subject to the laws and regulations of other countries if moved beyond national borders. National laws and regulations of the Member States of the European Union currently impose some restrictions on the movement of data outside national territory. However due to proprietary systems of cloud providers this may be hard to monitor and enforce.

The European Network and Information Security Agency (ENISA) suggests that there are a number of factors that will need to be taken into account in this situation. The ENISA report 'Security & Resilience in Governmental Clouds' proposes that governments assess whether current legal frameworks can be changed to facilitate the communication, treatment and storage of data outside their national territory, without exposing the security and privacy of citizens or national security and the economy to unacceptable risks.

Chrysanthi Papoutsi, Oxford Internet Institute - *"e-Infrastructure governance becomes increasingly important in the effort towards sustainability and cost-effectiveness. Long-term experience drawn from governance in other IT related areas can be used productively to achieve strategic alignment. In the face of the opportunities and challenges presented by technologies such as cloud computing, careful consideration of the legal issues involved relies partly on governance structures and mechanisms, which enable different decision-making processes in different e-Infrastructures."*

For more information:

Contrail: ***www.contrail-project.eu***

e-IRG: ***www.e-irg.eu***

ENISA: ***www.enisa.europa.eu***

Grids and Clouds GridBriefing: ***www.e-sciencetalk.org/briefings/GridBriefing_Grids_and_clouds.pdf***

HPC in the Cloud: ***www.hpcinthecloud.com***

Mosaic: ***www.mosaic-cloud.eu***

SIENA: ***www.sienainitiative.eu***

StratusLab: ***www.stratuslab.eu***

The Future of Cloud Computing report: ***http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf***

Venus-C: ***www.venus-c.eu***

EGI: ***www.egi.eu***

iSGTW: ***www.isgtw.org***

e-ScienceTalk: ***www.e-sciencetalk.org***

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7

Asia-Pacific Special Issue

Covering countries such as Australia, Vietnam, Japan and Indonesia, the Asia-Pacific region is both geographically vast and culturally diverse. But while the Asia-Pacific is home to a number of languages, cultures and peoples, its countries face similar challenges, such as natural disasters, climate change and connectivity to the rest of the world. Global e-Infrastructures, such as networks and grids, are already helping scientists in the Asia-Pacific contribute to science on a world-wide stage, in areas such as natural disaster modelling and life sciences. By further exploiting these infrastructures, scientists from the region can collaborate, share and store data, and achieve far more than they could alone.

Funding an infrastructure

Distributed computing technologies such as grids, clouds and volunteer computing could be vital in helping Asia-Pacific researchers work together to tackle regional challenges such as mitigating natural disasters, and to contribute to global questions such as climate change. But, as with any large infrastructure, securing and coordinating funding across an entire region is an enormous challenge.

In 2008 the EUAsiaGrid project, funded by the European Commission with partners across Europe and Asia, set out to promote awareness of the Enabling Grids for E-sciencE grid infrastructure, middleware and services in Asia. Over its two-year duration, EUAsiaGrid built up a community in Asia, supporting research such as earthquake mitigation and drug development. While the EGEE project ended in 2010, its work is now being continued in the European Grid Infrastructure, coordinated by EGI.eu on behalf of its participants and part-funded by the EGI-InSPIRE project, which includes eight partners from the Asia-Pacific region.

The Asia-Pacific market is large and fragmented, with individual countries taking different approaches to grid and cloud technologies. In countries such as Vietnam, for example, researchers have accessed grid technologies through the EGEE and EUAsiaGrid projects. While emerging cloud technologies are seen as a new opportunity, securing sufficient funding for engineers and PhD students to establish a reliable distributed infrastructure in the country remains a problem.

In Singapore on the other hand, distributed technologies are well adopted, with many government departments relying on clouds for their work. However, the Singapore government does not fund grid and cloud computing through academia, choosing instead to promote these technologies through industry. Nevertheless, in the research community, a variety of biomedical researchers use distributed technologies in vaccine design, virus research and genomic projects.

Simon Lin, ASGC - *"Distributed computing is capable of combining the results of exponential growth from information and communication technologies. Academia Sinica started as a Tier 1 centre for the Worldwide LHC Computing Grid but we are now taking advantage of the progress made and applying this to important issues of the region such as earthquakes, tsunamis and floods."*

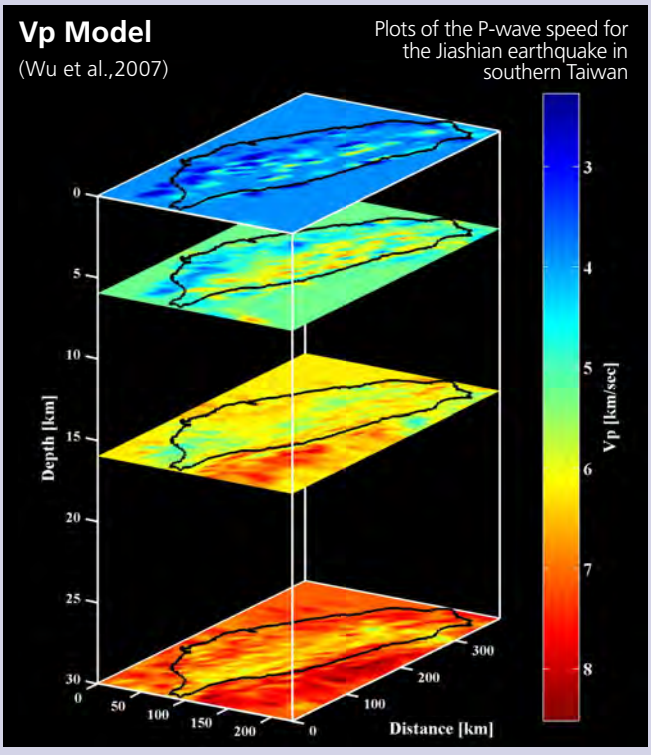
Shake, rattle and roll

For those living in countries along the so-called Ring of Fire, natural disasters such as earthquakes are a very real threat. And when disaster strikes, detailed knowledge about the event can make all the difference.

Researchers often use shake movies to visualise the motion of earthquakes after they occur. These movies can give valuable information to aid rescue efforts, education and outreach as well as helping to evaluate future risks.

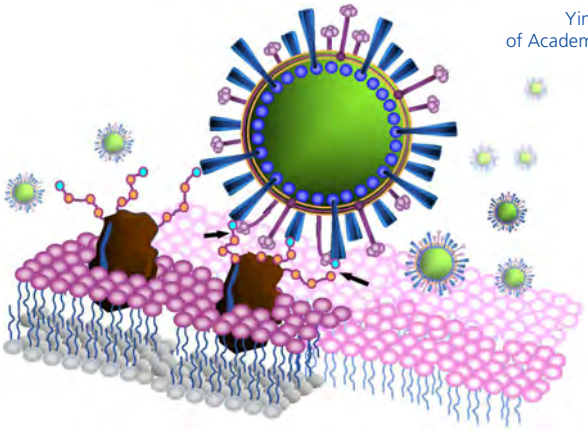
Researchers create shake movies by performing calculations on models of earthquakes as well as the Earth's structure. However the production process is computationally intensive, taking more than an hour to create a movie on a large computing cluster.

Scientists at the Institute of Earth Sciences at Academia Sinica, Taipei now plan to farm out tasks to volunteers' computers - cutting the time needed to create shake movies to not hours, but minutes.



Global science

By its very nature big science requires collaboration, and a key challenge for the Asia-Pacific area is to enable cooperation between countries in the region, as well as with the wider world. Collaborations between Asia and Europe for example, have already been used to find new drug targets for malaria, through the WISDOM project, which used the EGEE infrastructure.



Ying-Ta Wu of Academia Sinica

The Trans-Eurasia Information Network (TEIN3) provides a dedicated internet link for research and education communities in the Asia-Pacific. TEIN3 connects academics in Australia, China, India, Indonesia, Japan, Korea, Laos, Malaysia, Nepal, Pakistan, the Philippines, Singapore, Sri Lanka, Taiwan, Thailand and Vietnam. The project will soon expand to include institutes in Bangladesh, Bhutan and Cambodia.

TEIN3 also enables connectivity to the rest of the world thanks to direct links to similar initiatives such as GÉANT in Europe. This particular link enables users in the Asia-Pacific region to participate in joint projects with their European peers. To date TEIN3 has been used in the genetic sequencing of rice and has even connected dancers in Korea to music being played by an orchestra in Stockholm. The stable, reliable network provided by TEIN3 and GÉANT, also proved vital during Typhoon Emong, which hit the Philippines in 2009. In the lead up to the event, the Philippine Weather Bureau used this link to collaborate with the German Weather Bureau, DWD, in order to forecast and issue warnings about the oncoming disaster, saving thousands of lives.

Cooperation and interoperability

The CHAIN project, funded by the European Commission also aims to foster cooperation between different regions. A report from the project, due in September 2011, will include an analysis of which middleware is being run on sites across the world in order to understand what is needed to achieve interoperability between them.

Scan this QR code into your smart phone for more on this e-ScienceBriefing

In terms of technology development, the Pacific Rim Application and Grid Middleware Assembly (PRAGMA) sustains collaborations and advances the use of grid technologies. PRAGMA works with a community of investigators from leading institutions around the Pacific Rim. Its members can share technologies, test each other's code and provide useful feedback, in order to improve applications.

On the other hand the Asia Cloud Computing Association encourages stakeholders - developers, users, policy makers and researchers - to collaborate in order to accelerate adoption of cloud services. The Association aims to complement work done by other organisations, such as the Open Grid Forum, the Distributed Management Task Force, Inc. and the Cloud Security Alliance, all of which take more of a global standpoint.

Should I stay or should I go?

Migration has played an important part in Taiwanese social development since the 1600s. However, determining the motivations behind this phenomenon is far from straightforward.

The SimTaiwan project, headed by researchers from Taiwan and the UK, aims to discover the motivations behind Taiwanese migration. The team of computer and social scientists use a technique called agent based modelling to simulate the interactions that might take place in a dynamically populated and changing environment such as Taiwan.

In social based modelling, different attributes such as age, gender, health, or socio-economic status are assigned to a large number of individual 'agents'. By running the model on a computer, researchers can act out different scenarios over time.

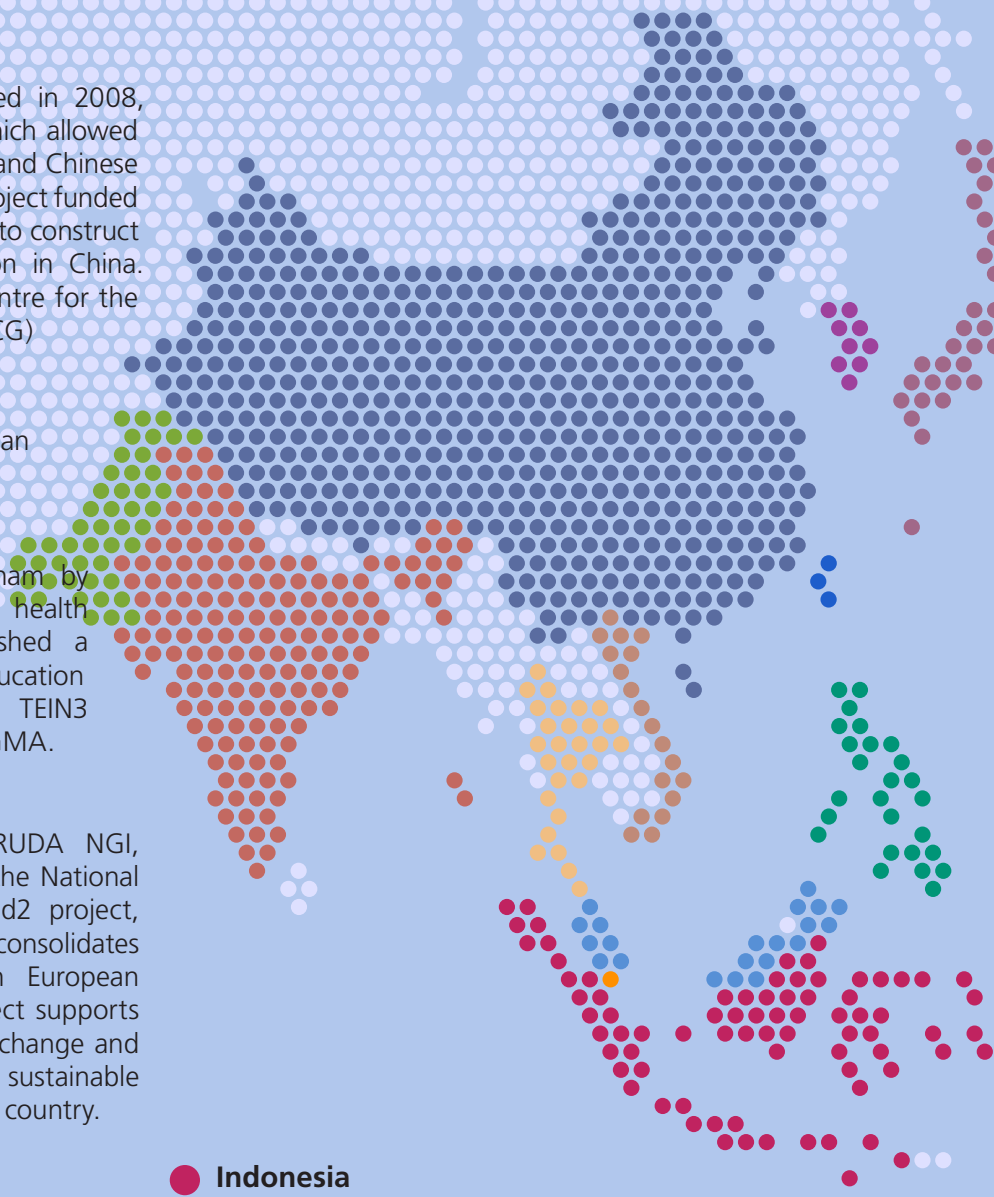
SimTaiwan uses census data to provide attributes for the 'agents'. But, before running the simulation, researchers need to test the performance and scale of the model. The SimTaiwan team have been helped by grid computing resources provided by Academia Sinica in Taipei to debug the model, as well as running stability and sensitivity tests to validate and verify simplified models.

For more information:

Academia Sinica Grid Computing Centre: ***www.twgrid.org***
Asia Cloud Computing Association: ***www.asiacloud.org***
CHAIN: ***www.chain-project.eu***
PRAGMA: ***www.pragma-grid.net***
TEIN3: ***www.tein3.net***
EGI: ***www.egi.eu***
iSGTW: ***www.isgtw.org***
e-ScienceTalk: ***www.e-sciencetalk.org***

e-ScienceTalk is co-funded by the EC under FP7

Examples of distributed computing across the Asia-Pacific



China
The EUChinaGrid project, which ended in 2008, aimed to design an e-Infrastructure which allowed full interoperability between European and Chinese e-Infrastructures. Today ChinaGrid, a project funded by Chinese Ministry of Education, aims to construct a platform for research and education in China. The country is also host to a Tier-2 centre for the Worldwide LHC Computing Grid (WLCG)

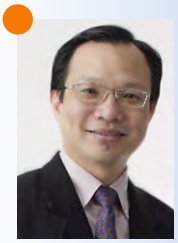
Pakistan
The National Centre for Physics in Pakistan is a Tier-2 centre for the WLCG.

Vietnam
EGEE grids were introduced to Vietnam by CNRS in France and used mainly in health research. Vietnam has now established a dedicated network for research and education (VinaREN) which is member of the TEIN3 consortium. It also participates in PRAGMA.

India
Indian e-Infrastructures include GARUDA NGI, the Indian component of WLCG and the National Knowledge Network. The EUIndiaGrid2 project, with partners in both India and Italy, consolidates and enhances cooperation between European and Indian e-Infrastructures. The project supports biology and materials science, climate change and High Energy Physics. It also promotes a sustainable approach to e-Infrastructure across the country.

Thailand
Thailand has participated in EUAsiaGrid, and is currently a member of PRAGMA and EGI-InSPIRE. The Thai National Grid Centre supports active research and is looking into cloud technologies, and the National e-Science Project aims to provide national infrastructure for e-science in Thailand.

Indonesia
inGrid is the Indonesian Grid infrastructure. Research areas include weather forecasting, natural disaster mitigation, chemistry and bioinformatics and establishing a digital library.



Tan Tin Wee, National University of Singapore - *"As an EUAsiaGrid and EGI-InSPIRE participant from Singapore, we see tremendous value in being connected to the R&D computational and storage e-Infrastructure. One exciting project which we are pioneering with the Asia Pacific Bioinformatics Network is building interoperable databases and resources for the life science research communities in Europe and Asia."*

Australia
NeCTAR - Australia's National eResearch Collaboration Tools and Resources programme – aims to enhance research collaboration through the development of shared e-research infrastructure. The project is split into four streams - virtual laboratories, research cloud, research tools and a national server program.

South Korea

KISTI, Korea Institute of Science and Technology Information, contributes one of the products of EMI and is an unfunded partner in EGI-InSPIRE. It has also has been an ALICE Tier 2 Centre for the Worldwide LHC Computing Grid since 2007. In clouds, South Korea's Communications Commission (KCC) has committed over US$500 million to the development of Korean cloud computing facilities.

Japan

The REsources liNKage for E-sclence (RENKEI) project aims to develop middleware to share resources distributed among multiple organisations, such as research laboratories, national computer centers and international grids. For clouds, GICTF, a non-profitable open technology forum, identifies the technical needs for secure cloud interworking and promotes global standardisation of inter-cloud system interfaces through collaboration between academia, government and industry. Japan also participates in the computing needs of the ATLAS particle physics experiment on the LHC accelerator at CERN, through an ATLAS Tier-2 centre based in Tokyo.

Taiwan

EUAsiaGrid has been a major driver for grid development in Taiwan. All resources are integrated and managed by the Academia Sinica Grid Computing Centre. Research supported by the e-infrastructure includes earthquake hazard maps, shake movies and understanding changing climate and environmental changes.

Philippines

PSciGrid – the country's e-science grid – aims to establish a national e-science grid infrastructure to enable collaborative science in the areas of earth science and life science. The Philippines is a contributor to EUAsiaGrid and a member of PRAGMA.

Malaysia

MYREN, the Malaysian education and research network is connected to TEIN3, and is planning to merge grid and cloud services for its user communities. MYREN 2 is set to offer MYRENCLOUD services which will also be open to bodies such as community centres and schools.

Marco Paganoni, INFN - *"The entire world is facing issues like climate change, an increasing population, economic problems as well as competitive disaster mitigation. I believe in all these issues, as well as biomedical, biochemical research and particle physics we need cooperation from everyone to profit from the continuously increasing amounts of data. Projects like CHAIN are looking into this and trying to foster this type of collaboration worldwide."*

Preserving Malaysia's cultural heritage

Malaysia is well known for its diversity. With both a multi-ethnic and multi-religious population, it is a country with a rich culture and heritage.

However, Malaysia is at risk of losing this. Cultural artefacts, folklores, performances and rites are gradually being forgotten thanks to the adoption of foreign values and cultures.

In an effort to preserve its heritage, Malaysian researchers are turning to e-technologies to create digital versions of cultural objects and traditions. For example, dances are being mapped and stored as 3D coordinates while objects are being digitised and annotated for future generations.

Malaysian researchers hope grid technology could hold the key to the long-term preservation and processing of this highly heterogeneous material, as well as storing it in geographically distributed digital archives. By building up an e-culture and heritage digital library, Malaysia's cultural diversity can be preserved for the years to come.

Desktop grids: Connecting everyone to science

Today's personal computers are powerful but, most of the time, a large proportion of their computational power is left unused. A desktop grid takes this unused capacity, no matter what its location, and puts it to work solving scientific problems. With over 1 billion desktop computers in use, desktop grids can offer a low cost, readily available computing resource for scientists while allowing citizens across the world to contribute to scientific research. Together with grids and supercomputers, desktop grids can be a useful complement to the e-Infrastructure landscape.

What is a desktop grid?

Desktop grids fall into two categories - local and volunteer. While local desktop grids are comprised mainly of a set of computers at one location, a business or institute for example, the resources in a volunteer desktop grid are provided by citizens all over the world.

Today researchers are using desktop grids to simulate protein folding *(Folding@home)*, find ways to provide clean water *(IBM's World Community Grid)*, and model climate change *(Climateprediction.net)*. Scientific problems that can be split up into small tasks and sent to different computers for computation are perfect for solving with desktop grids. These projects farm out tasks to computers located across the globe which send the results back to scientists once they are complete.

Desktop grids can open up computing to researchers that otherwise would be unable to access such large amounts of computing power. For example the Citizen Cyberscience Centre has been set up to help scientists in developing countries access the power of internet-based volunteer networks. Initiatives such as Africa@Home and Asia@Home have also encouraged more researchers in these regions to use desktop grids.

David Anderson, BOINC director – *"BOINC is being used by over 50 volunteer computing projects, doing research in everything from quantum mechanics to cosmology. About 430,000 PCs from all over the world, many of them equipped with GPUs, participate in these projects. Together they supply about 6 PetaFLOPS of computing power."*

A large majority of volunteer computing projects are based on open source software called BOINC. BOINC allows scientists to plug their own projects into the software, so volunteers can easily download and run applications on their computer. The BOINC client, used by the volunteers, can be configured to run only when the PC is not in use, often as a screensaver, or to run at the lowest priority while the PC is in use. Other desktop grid middlewares include XtremWeb, developed by INRIA/CNRS, which is mainly used to manage computations on desktop computers within an organisation.

Bringing physics home

© CERN

LHC@home 2.0 aims to bring the world's largest particle accelerator into your home. The platform – an extension of the already successful LHC@home – allows volunteers to connect to CERN-based research projects simply by donating their extra computing power. The project Test4Theory, for example, simulates high-energy particle collisions which scientists can compare to real-life collisions, such as those occurring in the Large Hadron Collider (LHC).

"My dream is to be able to establish a 'virtual LHC', which would require being able to generate 40 million events per second, as much as the real LHC, running at full steam," says Peter Skands, the lead scientist behind Test4Theory. "We estimate that it would take somewhere between 10 000 and 100 000 connected computers to achieve this, a combined amount of computing power that we have only faintly begun to imagine, since we started working with LHC@home 2.0. With the enthusiasm we have seen in the public so far, there definitely appears to be awesome possibilities for what we can do with this platform."

Nicole Vasapolli, BOINC volunteer - *"I'm a meteorologist so was originally interested in donating my computing time for climate research as it was related to my work. But, as time went on, I began to take part in many other exciting life science projects, to assist scientists in solving pressing problems. I've found that BOINC is free, easy to install and it's an entertaining way to be part of scientific progress."*

Netting malaria

Anyone going on holiday to a malaria-affected country will often head to their doctor for a course of malaria tablets. But for those who live in countries at risk, taking preventative medicines is impractical.

So what do citizens of these areas do? They use mosquito nets or insect repellent, treat their houses with insecticide, or get rapid treatment in the event of becoming ill.

While none of these methods are perfect, each can make a big impact given the widespread nature of the disease. Malaria is preventable and treatable but three quarters of a million people die from it every year, putting healthcare services in affected countries under enormous strain.

Thomas Omondi / UK Department for International Development

Healthcare providers and governments need to find ways to determine the most effective combination of treatments for their area. They can use mathematical models to simulate the effectiveness of different combinations of malaria control, and work out what the best solution is for a given situation.

In 2003 researchers at the Swiss Tropical and Public Health Institute started running malaria models to answer these questions. Starting with just 50 of their own PCs, they soon opened up the project via BOINC and malariacontrol.net was born. Today 50,000 people contribute computing time to the project, through over 70,000 PCs. In total, malariacontrol.net and its volunteers have notched up over 10,000 CPU years helping healthcare professionals fight malaria.

Enhancing other e-infrastructures

Desktop grids are just one of a number of ways in which researchers can access computing capacity. They can provide a useful complement to the other facilities in the e-infrastructure landscape. While supercomputers are able to solve a wide variety of complex computational problems they are expensive and are limited to a relatively small number of researchers. Cluster-based grids can provide a cheaper solution for more researchers, but for a more limited set of applications. Assuming the computers making up a desktop grid are already paid for, they can open up computational research to more scientists at an even lower cost.

Mikhail Posypkin, Institute for Systems Analysis of Russian Academy of Sciences *"Desktop grids offer a cost-effective alternative to supercomputers or service grids. Unlike supercomputers or service grids large desktop grids are almost free: all you need is a meaningful distributed application. Presently desktop grids can contribute their resources to existing grid infrastructures thus producing a really powerful combined distributed computing infrastructure."*

The EDGeS (Enabling Desktop Grids for e-Science) project, which has now finished, worked to connect desktop grids to the wider EGEE (Enabling Grids for E-sciencE) European infrastructure. Using the gLite middleware the project defined common policies to integrate desktop grids into existing EGEE service grids. Today EDGI (the European Desktop Grid Initiative) is continuing this work by connecting desktop grids into the European Grid Infrastructure (EGI). As well as focusing on gLite, EDGI aims to build bridges to the UNICORE and ARC middlewares.

Desktop grids have high capacity but not a guaranteed quality of service as the available computing power depends on which computers are not being used. EDGI hopes to include a cloud in the infrastructure, which can be used as and when necessary. This will allow the service to be used by applications which need to be completed within a specific deadline.

By integrating desktop grids with other e-infrastructures, researchers can run applications across different types of computing resources, matching parts of the computational problem to the most suitable execution environments. For example, some parts of an application can be run on a desktop grid, and others on a high-end supercomputer. Collaborations between a number of infrastructure providers, such as those established through the International Desktop Grid Federation (IDGF) are already in place to enable these applications. On the international scale the DEGISCO project aims to export and share desktop grid knowledge outside of the EU, while policy bodies like e-IRG are preparing the setup of the legal and political frameworks.

Vicky Huang, ASGC – *"Asia is a geographically large region, with diverse and scattered resources (technologies and facilities) coupled with the general problem of insufficient investment from government in academic hardware supply. As such, the concept of desktop grids and volunteer computing is very suitable and useful to popularise in the Asia-Pacific region through initiatives such as DEGISCO."*

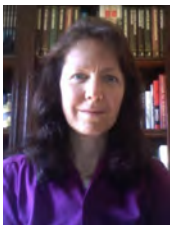
Desktop computing challenges

Desktop grids can provide a variety of different benefits, however their use raises a number of challenges. A Desktop Grids for eScience Road Map produced by the DEGISCO project in July 2011 took a closer look at some of the following issues:

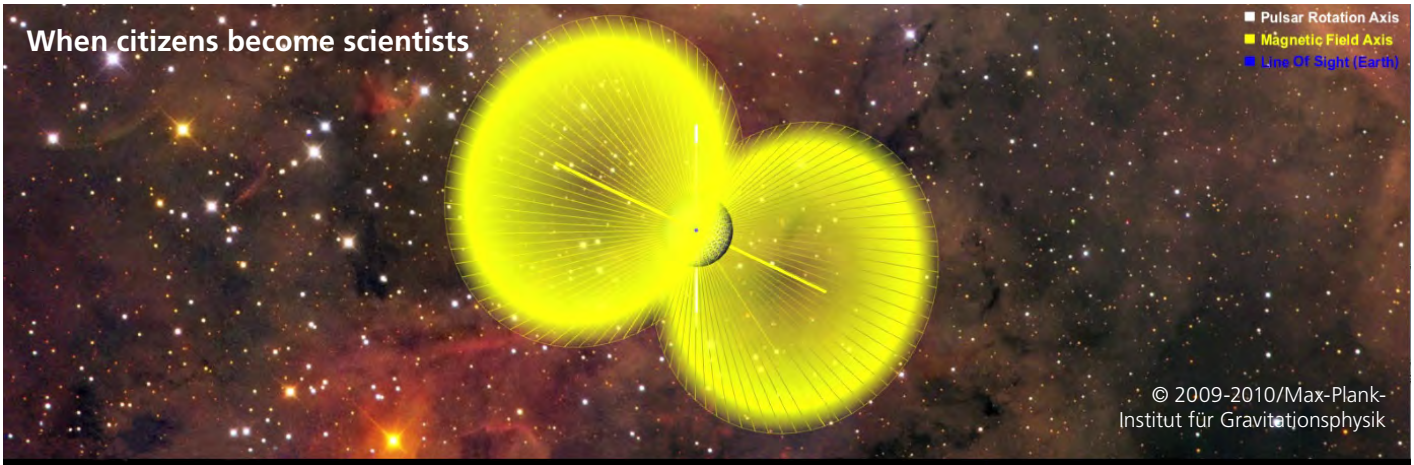
- **Supporting a desktop grid:** Aside from having to develop applications that can run across a number of heterogeneous systems, the distributed nature of a desktop grid poses unique problems. As volunteers provide the resources, it is difficult to test and fix applications.
- **Making it green:** Desktop grids are often touted as a 'green solution' as they use computing resources already in existence. However, in reality, determining whether a desktop grid is green, or not, is complex. How volunteers choose to donate their computing time plays a big part in this – adding on a CPU load to a machine running at a low capacity doesn't cost much energy, but using a computer that would otherwise be switched off does. Even the country a machine is running in can make a real difference. Connecting a computer in a hot country such as Dubai to a desktop grid is likely to use more energy, as the machine needs to be kept cool.

Morgan Duarte, BOINC volunteer - *"I'm sharing my computing resources with BOINC to help solve tomorrow's challenges and be more involved in scientific progress and our future. I believe it is an efficient way to use our continuously increasing computers' power without affecting my own personal use. For me, volunteering my computer for science is very rewarding."*

- **Local policies:** Desktop grids are subject to the local ICT policies at the institute or organisation that is hosting the donated computer. For example, if a company chooses to switch off computers at night, this can affect the availability of the desktop grid.
- **Evolving hardware:** Today increasing numbers of people are accessing the internet through new technologies such as mobile phones instead of PCs. In the future this evolving situation could have consequences for the desktop grid concept as it currently stands.



Leslie Versweyveld, AlmereGrid & IDGF *"The special feature desktop grids have to offer is that they are already part of the e-Infrastructures landscape. We are actually sitting on a huge source of computational power that is largely left unused in numbers of universities, research institutes, companies, home offices and households. It is already there, we only need to tap into it, technically gain access to it and transform this enormous resource of computational power to fuel e-science research in all possible areas. Basically, it is a mere question of ecological recycling."*



When Einstein@Home discovered a new pulsar its discovery wasn't credited to astronomers, but to its volunteers - Daniel Gebhardt, from Mainz, Germany, and husband-and-wife team Chris and Helen Colvin of Ames, Iowa.

Pulsars are highly magnetised, rotating neutron stars that emit a beam of electromagnetic radiation. Einstein@Home, a BOINC project, was originally set up to search for gravitational waves in data from the US LIGO Observatory. However in March 2009, the project also began to use its volunteers' computers to search for signals from radio pulsars in observations from the Arecibo Observatory in Puerto Rico.

The new pulsar, discovered in October 2010 and named PSR J2007+2722, was the first deep-space discovery by Einstein@Home. Since then a further seven pulsars have been discovered by Einstein@Home volunteers, showing how donating your computer can make a real difference.

Managing a desktop grid

Unlike supercomputers or cluster-based grids, desktop grids have an extra component that needs to be managed – their volunteers. Using volunteers to donate computing time forms the basis of all volunteer desktop grids, and can create positive links between citizens and science.

Francois Grey, Citizen Cyberscience Centre
"In my view, the most revolutionary aspect of volunteer computing is the public participation. Far from being passive, many participants turn volunteer computing into a serious hobby. Some contribute to debugging the software, others help newcomers in the forums, still others set up teams and events to encourage more participation. I predict that ultimately, this will lead to public involvement in setting the agenda for the research that is carried out using public resources. Just as has already happened for journalism on the web, the distinction between amateur and professional will start to blur."

The first step - recruiting volunteers - needn't be a difficult one. When the project LHC@home began, its creators thought it would attract no interest. However one thousand people downloaded the application in the first 24 hours with no publicity effort at all. Often volunteers are interested in the area of science they are contributing towards such as searching for new drugs or ways to generate clean water. AlmereGrid has taken recruitment one step further, by setting up a 'city grid' intended to reach out to volunteers that may not be traditionally interested in donating computing time. AlmereGrid has partnered with local and national companies across Almere in the Netherlands to disseminate information on volunteer computing and get more people interested in the topic.

While projects do not need to pay volunteers to use computing resources, they do need to keep volunteers informed. To ensure volunteers' interest is sustained over a project's lifetime they should be provided with feedback and information on how the project's research is progressing.

Glossary

CPU: Central Processing Unit; a microprocessor (a processor on an integrated circuit) inside a computer that can execute computer programs.
GPU: Graphics Processing Unit; a device that renders graphics for a computer. GPUs have a highly parallel structure that makes them more effective than general-purpose CPUs for some complex processing tasks.
Quality of service: the ability to guarantee a certain level of performance.

Fundraising through science

The Charity Engine has ambitions to be a worldwide computer. Launching in summer 2011, Charity Engine will provide volunteers' computing time to a collection of hand-chosen projects and raise money for charities at the same time. By joining Charity Engine, its volunteers will also have the chance to win a cash prize of up to a million dollars, every few weeks.

Charity Engine raises funds for its associated charities, as well for its prize draws, by selling volunteers' computing time in bulk to science and industry. Its volunteers are not asked to support any particular science project they simply agree to let Charity Engine send ethical work to their PCs.

"Our volunteers are joining to make computer-generated charity donations and prize draw entries, they might not actually care about the science," says Mark McAndrew, founder of Charity Engine. "But that's fine, because all that idle, wasted computing power will make Charity Engine the ultimate supercomputer - and we love the science."

For more information:

AlmereGrid: ***www.almeregrid.nl***
BOINC: ***http://boinc.berkeley.edu***
Charity Engine: ***www.charity-engine.org***
Citizen Cyberscience Centre: ***www.citizencyberscience.net***
DEGISCO: ***www.degisco.eu***
Desktop Grids for eScience - A Road map: ***http://bit.ly/DEGISCOroadmap***
EDGI: ***www.edgi-project.eu***
Einstein@Home: ***http://einstein.phys.uwm.edu***
e-IRG: ***www.e-irg.eu***
IDGF: ***www.desktopgridfederation.org***
LHC@home: ***http://lhcathome.web.cern.ch***
Malariacontrol.net: ***www.malariacontrol.net***
XtremWeb: ***www.xtremweb.net***
EGI: ***www.egi.eu***
iSGTW: ***www.isgtw.org***
e-ScienceTalk: ***www.e-sciencetalk.org***

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7

Research networks: global connectivity

As science becomes increasingly global and collaborative, researchers' dependence on fast and reliable data and communication links continues to grow. Research and Education (R&E) networks are designed to meet these demands, providing high-speed and reliable internet links to support applications and experiments crucial to research.

In the next decade, the demand for computationally driven data collection and information-sharing will escalate dramatically. GÉANT and other R&E networks will inevitably play a central role in enabling interconnectivity and collaboration across Europe and the world.

Enabling research and innovation

Networking is an essential part of the e-infrastructure connecting people around the world to global ICT services. Without reliable access to scientific instruments, data, collaborators, and other resources many international research experiments would not be possible.

Within Europe, the dedicated pan-European R&E network, GÉANT, transfers huge quantities of data (over 1,000 terabytes per day) for fields as diverse as radio-astronomy and drug research. In the past moving such large data-sets may have taken days or would not have been possible, but now with high-bandwidth technology, transmission can take seconds.

Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda – *"The power and scope of GÉANT ensure Europe remains a central hub for research and education, offering the best infrastructure to the brightest minds in the world. Rich with these successes, GÉANT must now position itself to face the challenges of the next decade such as the upcoming 'data deluge', connectivity at world scale, and providing a seamless service to all EU scientists to build an online European Research Area."*

Knowledge without borders

The GÉANT network is fundamental to the European Commission's vision of providing equal opportunities and access for European researchers irrespective of their location within Europe.

In October 2011, a report entitled 'Knowledge without Borders: GÉANT 2020' provided an action plan to serve the needs of the community and help maintain and strengthen Europe's research agenda. Among its recommendations were the provision of a more user-based service culture, and a continued commitment to increasing collaborations with other continents as well as testing emerging internet technologies.

Helping radio-astronomers see further back in time

Reliable and robust links also allow researchers to share data in real-time. Astronomers are using networks to connect multiple radio telescopes across Europe and beyond. Using a technique called e-VLBI, or real-time, electronic Very Long Baseline Interferometry, astronomers can inspect their results almost immediately. This technique relies on GÉANT and other networks to connect telescopes to a central data processor (a supercomputer), which correlates the data from the telescopes synchronously.

Exploiting e-infrastructures such as the GÉANT network, data can be streamed from each telescope and correlated in real-time. This updated technique yields results in a matter of hours, rather than the weeks it takes with the traditional technique of recording data to disk and physically shipping them for processing. The fast turnaround provides astronomers with a better tool for studying supernovae, gamma-ray bursts and other so-called transient activity that might otherwise be missed.

© j-VLVI - Cambridge

The 'backbone of the internet'

Ultra-fast networks help to minimise the delays that build up as data is transmitted over the internet. The actual physical infrastructure (the network cables) no longer relies on copper cables but state-of-the-art optical fibres, which provide much more bandwidth and a reliable 'backbone' linking the major 'nodes' allowing researchers to collect, distribute and analyse data securely.

R&E networking in Europe is organised in a hierarchical fashion, connecting research and education community users. The network connection between two end users will be provided by a chain of several networks, each connected to the next. This chain will typically start with a campus network then may include a regional network before connecting to a national (NREN) network. Then to the pan-European backbone GÉANT, from there to another NREN and so on back down the chain to the user at the other end.

Together, GÉANT and the National Research & Education Networks (NREN) partners interconnect more than 40 million researchers and students at more than 8,000 institutions across 40 countries. Key routes on GÉANT already run at 40 Gb/s (gigabits per second), with planned upgrades to 100 Gb/s scheduled for 2012 to ensure the network remains ahead of user demand for bandwidth.

Kostas Glinos. Head of Unit "GÉANT & e-Infrastructure" in DG INFSO – the Directorate General for Information Society and Media. - *GÉANT needs to continue being ahead of the market in terms of the connectivity and services it provides to researchers; and it needs to organise itself to respond flexibly and efficiently to the needs of scientific communities for moving around extreme data volumes. GÉANT will help make Europe a hub of global e-Science.*

A global campus

e-infrastructures provide the ability for researchers to access a pool of resources (e.g. scientific instruments, data and collaborators) from anywhere ensuring equal opportunity for all researchers wherever they are located. It helps to bridge the 'Digital Divide' and ensure inclusivity. Seamless global connectivity allows virtual communities of researchers to co-operate and collaborate across continents as if they were on the same campus. As part of a larger consortium, individual NRENs can also benefit from long-term economy of scale.

Supporting 'Big Science' and everyday research

Research increasingly depends on large-scale databanks and massive processing power to help solve complex scientific or engineering problems. Any network performance issues can significantly impact scientists' ability to perform their research.

Users from a diverse number of academic disciplines rely on R&E networks including scholars in the arts and humanities. Biologists at the European Bioinformatics Institute (EBI) have utilised networks to share, store, manage and interpret bioinformatics data.

R&E networks have provided the foundation transport 'layer' for Grid infrastructures such as the Worldwide Large Hadron Collider Grid (WLCG). The 22 Petabytes of data generated from collisions at CERN is transferred and shared for analysis to 11 separate major computing centres dispersed around the world by high-speed optical fibre networking links.

Networks have also made an important contribution to speeding up the reconstruction of physical infrastructure after natural disasters. High-resolution satellite images sent for analysis for rescue teams via GÉANT and the Asia-Pacific TEIN3 network have helped plan rescues in the aftermath of earthquakes in China.

Torsten Reimer, Programme Manager (Digital Infrastructure), Joint Information Systems Committee (JISC) UK - *Today research is increasingly collaborative – across institutions but also countries and even continents – and it relies on ever growing amounts of data. In some research areas we are only beginning to understand the potential of this change, but collaborative access to data and digital infrastructure are now at the heart of research. Building and connecting research networks across and beyond Europe is critical to enable the potential of the digital transformation of research.*

Worldwide networking

In addition to its pan-European reach, the GÉANT network has extensive links to networks in other world regions including North America, Latin America, North Africa and the Middle East South, Africa and Kenya, the South Caucasus, Central Asia and the Asia-Pacific Region. Work is also on-going to connect to the Caribbean and to improve links to and within Southern and Eastern Africa.

Consequently GÉANT's extensive geographical reach provides Europe's NRENs with a gateway to NRENs worldwide, enabling European researchers to share huge quantities of data and collaborate effectively with their peers throughout the world. GÉANT is operated by DANTE (Delivery of Advanced Technology to Europe) on behalf of Europe's NRENs, who co-fund the project with the European Commission.

DANTE works closely with TERENA (The Trans-European Research and Education Networking Association), a collaborative forum that has supported and shaped the development of the internet for the last 25 years.

Fulvio Galeazzi, Project Manager, DECIDE - *The high speed Pan-European network GÉANT and other national research networks are focused on supplying connectivity and a growing portfolio of advanced services, allowing researchers to derive maximum benefit from a simple and secure access to a high capacity network. Dedicated network services for specific applications or projects, network performance monitoring tools, secure roaming services and authentication and authorization services for accessing shared resources (data and image archives, libraries, e-learning systems, etc.) are some of the innovative services available to the researchers.*

Sharing experiences

Experience and knowledge gained from R&E networking in Europe can help to advance e-infrastructure and innovation across other global regions. Advice, case studies as well as best practices in areas such as technical support, are assisting networking partners in other regions.

For developing countries, establishing an R&E network provides a framework for delivering on the United Nations anti-poverty Millennium Development Goals (health, climate, agriculture, education and the environment). It can also be one of the building blocks for creating an effective education system.

Researchers in Sub-Saharan Africa from early 2012 will be connected to international networks via the AfricaConnect project. It is expected that many research areas will advance through the high-speed connectivity and supplementary services provided by networking.

In the remote parts of Africa, researchers can benefit from distance learning and live videoconferencing, enhancing skills and knowledge in the local research community, thus unlocking Africa's intellectual potential.

In South Africa, e-Health and telemedicine, astronomy and physics are already actively exploiting the high-performance network infrastructure.

Domenico Vicinanza, Project Support Officer, DANTE - *From a network management perspective, R&E networks provide new standards of clarity and control. You can monitor use of resources in real time and rely on network repair, maintenance and development activities being managed centrally, with 24-hour central support. R&E networks offer tomorrow's network today."*

Peter Clarke, Professor of Physics at the University of Edinburgh UK - *R&E networking is vital to the Large Hadron Collider (LHC) operations. We transmit many Petabytes of data each year to be reconstructed and analysed in computing centres around the world. Without our NRENs and GÉANT we wouldn't be able to produce the results you see from the LHC.*

Guiding technological innovation

R&E networks offer opportunities for experimentation and are established pioneers in the use of advanced network applications and emerging internet technologies. By facilitating the development – from idea, to prototype, to the commercial internet – many technologies and applications find their way from research networks to the commercial world.

Assisting early diagnosis of Alzheimer's

Rapid, easy and secure access to networks is also important in healthcare. Clinicians often require access to large medical reference databases in order to compare patient imaging data for making an informed diagnosis, which is especially important for the early diagnosis of Alzheimer's.

Images courtesy of Ivan Porro

The DECIDE (Diagnostic Enhancement of Confidence by an International Distributed Environment) project uses high-speed research links to provide doctors with an easy-to-use online application for the analysis of neurological data (i.e. brain scans). The network and processing power to carry out such analysis is effectively beyond the budgets and computing power of most hospitals. R&E networks provide connectivity to hospitals and national research networks, allowing doctors to access and upload biomedical images irrespective of location, in order to collaborate and better understand the disease process.

Richard Hughes-Jones, Technical Customer Support Manager, DANTE - *"To improve the way we deal with disease, disasters and other natural challenges, we need to understand more about our world - how it works and how it's changing. If we're going to make life better for people, we have to learn to share our knowledge and our skills. The answer lies in working together effectively. R&E networking is important because it provides a platform that enables better cooperation, collaboration and integration within and between geographically dispersed research and education communities."*

User-focussed and flexible service

In addition to high-speed internet access, users benefit from a number of services provided by international R&E networks from large file transfers, computer modelling and simulations, application sharing and a whole host of visualisation tools.

However, research communities differ in their requirements, so flexibility and scalability are increasingly being built into services. For example, LHC physicists may need increased access to large volumes of data for relatively short periods of time.

GÉANT's perfSONAR MDM is a multi-domain monitoring tool that makes it easier to simplify troubleshooting and access performance problems occurring between sites connected through several networks.

Bandwidth-on-demand is expected to be valuable to users who may need to transport high volumes of data over the network in relatively short time periods. It allows users to reserve end-to-end data transport capacity when they need it, between end points participating in the service.

Future challenges

Big challenges lie ahead for R&E networking; not only will the networking consortiums have to meet the needs of supporting large scale computing but there are a number of organisational and technical hurdles to overcome.

- **Increasing capacity** by moving from 10-Gb/s to 40-Gb/s and 100-Gb/s line speeds.
- **Providing 'greener' networks** by carrying out environmental impact studies to formulate best practices across the infrastructure.
- **Safeguarding and addressing security (privacy and anonymity) issues.** As capacities increase and global connectivity advances, it will be increasingly important to develop an integrated security framework in order to safeguard against cyber-attacks. GÉANT employs an automated system - the National Security Handling and Response Process (NSHaRP). The system not only informs affected users of threats but also provides support for dealing with security incidents.
- **Moving towards interoperability** - The Open Grid Forum (OGF) is leading the global standardisation effort and interoperability between the different technologies used in distributed computing systems around the world. Their Network Services Interface (NSI) protocol will provide an interface between network domains in order to provide interoperability in a heterogeneous multi-domain environment.
- **Ensuring governance is transparent and inclusive** - Streamlining the governance arrangements to reflect the European and international dimensions, and allowing users more of a role in the development of governance activities.
- **Cutting the costs of data roaming.** Expensive data roaming within the commercial mobile networks is a big obstacle to the mobility of scientists.

Switching over: IPv4 to IPv6

The phenomenal global growth of the internet has led to a shortage of internet addresses – the numerical label assigned to each device. IPv6 is the new version of the internet address protocol that has been developed to supplement (and eventually replace) IPv4, the version that underpins the internet today. The switch to IPv6 has been validated and certified prior to wider release by GÉANT and many European NRENs.

Bringing to life ancient instruments

Reconstructing the sounds of ancient musical instruments has become a reality for archaeologists through the ASTRA (Ancient instruments Sound/Timbre Reconstruction Application) project which has been facilitated by high-speed transatlantic internet links. A technique called physical modelling synthesis, was used to reconstruct two South American instruments – a Chilean drum and a Peruvian flute – which had not been played for over a thousand-years.

Video - excerpt of a concert - www.astraproject.org/download.html

Archaeological data (e.g. fragments from excavations, written descriptions, pictures of the two instruments) were sent through the ALICE2 transatlantic link between Europe and Latin America. Several gigabytes of data were exchanged in almost real-time by two teams of researchers in the two continents. To speed up the procedure and achieve the necessary processing power, the reconstruction processes were run simultaneously on hundreds of computers throughout Europe and the lower Mediterranean (using the European Grid Infrastructure, GILDA and EUMEDGRID). The sounds were transferred back to Santiago in Chile, to be played for the first time at a public performance of an opera.

For more information:

TERENA: ***www.terena.org***
DANTE: ***www.dante.net***
GÉANT: ***www.geant.net***
TEIN3: ***www.tein3.net***
RedCLARA: ***www.redclara.net***
SURFNet ***www.surfnet.nl/en/***
CAREN: ***http://caren.dante.net***
Internet2: ***www.internet2.edu***
AfricaConnect: ***www.africaconnect.eu***
DECIDE: ***http://www.eu-decide.eu***
ASTRA: ***www.astraproject.org***
Knowledge without Borders: ***http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/geg-report.pdf***
GÉANT Real time Monitor (RTM) ***http://rtm.hep.ph.ic.ac.uk/net_webstart.php***
EGI: ***www.egi.eu***
iSGTW: ***www.isgtw.org***
e-ScienceTalk: ***www.e-sciencetalk.org***

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733

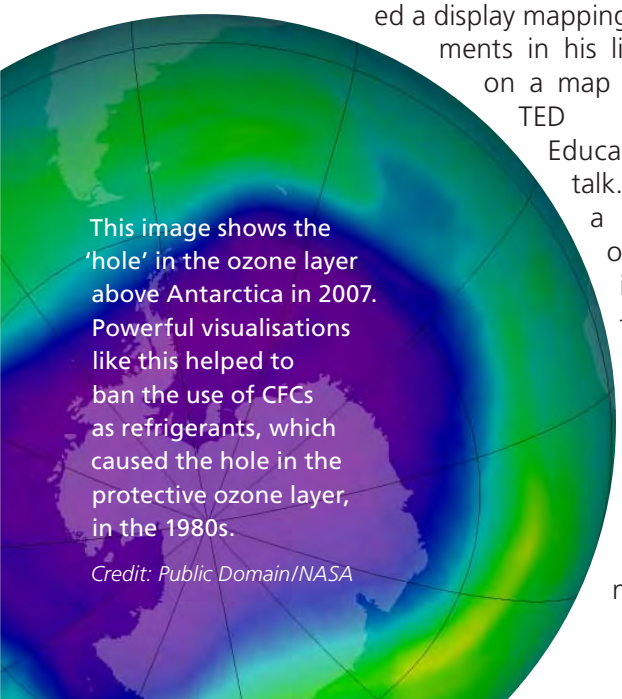
Visualisation

Information graphics, graphical information

The open data revolution, driven by a growing number of conscientious researchers and enlightened academic publishers, is making more data available to scientists and the public at large than ever before. Alone and without context, this mass of data can be a daunting deluge of numbers. Visualisation helps us not only to understand the conclusions of research, but transmits ideas across disciplines and cultural boundaries, creating a collaborative infrastructure that actually improves the quality of science being done. Visualisation may also help drive home the social and economic consequences of research, for example in understanding global environmental change.

Visualising data allows us to understand systems on a wide range of scales, from the global to the local. MAPPER – Multiscale Applications on European E-infrastructures – is a framework allowing scientists to seamlessly integrate simulations of natural phenomena, where different factors are important at different scales. "It's important that scientists should be able to zoom in and out of datasets in a coherent way," says Alfons Hoekstra, Director of MAPPER. "There are some important questions that can be explored through scientific visualisation."

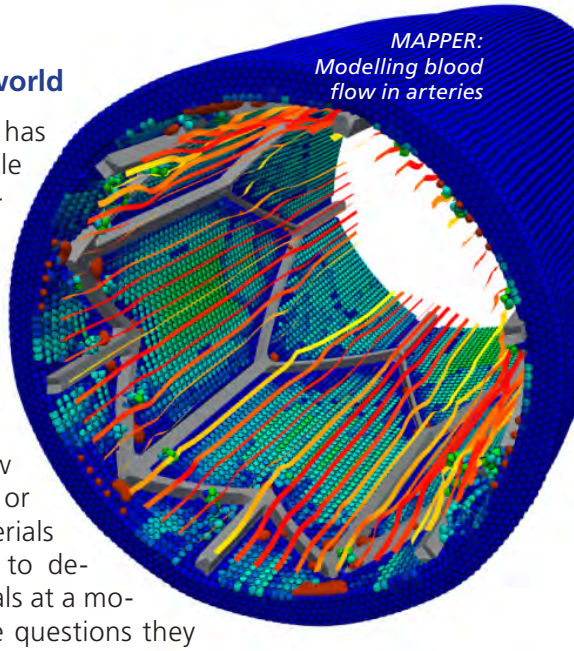
Visualising information can also connect people with scientific data more instinctively: "Visualisation is about placing data in a human context," said Jer Thorpe, data artist-in-residence at the New York Times, as he presented a display mapping out key moments in his life as points on a map for a recent TED (Technology-Education-Design) talk. As one of a new breed of infographics designers – whether graphic artists interested in data, or statisticians interested in communicating data visually – Thorpe plays a key role in helping the public to understand complex issues in science. Just as the free press has made political decision-making an open process in democracies, so these information journalists are making scientific data truly available to everyone, explaining research findings clearly and openly.



This image shows the 'hole' in the ozone layer above Antarctica in 2007. Powerful visualisations like this helped to ban the use of CFCs as refrigerants, which caused the hole in the protective ozone layer, in the 1980s.

Credit: Public Domain/NASA

An eye on the biomolecular world



MAPPER: Modelling blood flow in arteries

Visualisation has an important role to play in making e-Science more attractive to molecular scientists. Whether they are life scientists searching for a new drug candidate, or chemists or materials scientists trying to design new materials at a molecular scale, the questions they are faced with come down to 3D shapes. The capability to render a 3D model of a molecule has been around for decades, but predicting complex 3D structures from numerical data, and calculating how that molecular structure interacts with other molecules, is computationally-intensive, and has begun to benefit from advanced e-Science technologies. Emna Harigua, a PhD student who works at the Institut Pasteur in both Paris and Tunis, studies computational modelling of a single protein in a parasite that causes leishmaniasis, a disease that can cause damage to the spleen and liver.

Using the open source software Dock, Harigua has been able to see which potential drugs should interact most strongly with her protein. She has been able to screen 85,000 drug candidates down to less than 100 that could potentially treat the disease. Her work won her the 2012 L'Oreal International Fellowship for Women in Science.

Emna Harigua, PhD candidate- Institute Pasteur, Paris & Tunis – *"Visualisation in molecular biology is very important, making e-science accessible to scientists with a variety of backgrounds, not just the highly technically literate. It enables scientists to gain a better insight into their work. In my case, visualising potential drug candidates interacting with the protein I study makes things so real. It helps me cast a critical eye over the results."*

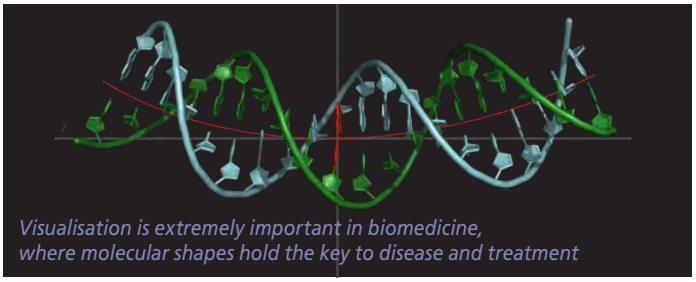
Alexandre Bonvin, WeNMR – *"Visualisation isn't just an endpoint – it is integral to how we do our research. The computer automates a lot of the process, but you still need a human eye to judge what is happening. To design a drug, you look at how the chemical groups involved in biomolecular interactions come together to rationalise your next step. Without visualisation, we would be blind."*

WeNMR: Life through the eyes of a protein

Alexandre Bonvin is Professor of Computational Structural Biology at the Bijvoet Center in Utrecht, and Coordinator for WeNMR, a project bringing collaborative, distributed computing technologies to structural biology. Bonvin uses nuclear magnetic resonance (NMR), a variant of the magnetic resonance imaging (MRI) that doctors use to scan hospital patients, to determine the distances between atoms in proteins, producing a NMR 'fingerprint'.

Proteins are large, flexible molecules that can adopt a wide variety of conformations – different shapes in 3D space – which can give rise to a vast number of subtle variations of this fingerprint. The large number of possible conformations makes figuring out what the 3D structure of the protein is very computationally-intensive, especially when only limited experimental information is available. "In order to calculate the structure, we need to do not one calculation but many tens of thousands of calculations depending on the problem," explains Bonvin. Applying grid technologies speeds up the process of turning these large tables of numbers into 3D models, enabling scientists to quickly understand what the proteins look like at minuscule scales. This is important because many medicines are based on how small drug-like molecules interact with protein-based receptors.

Henry Hocking – a postdoctoral researcher at The University of Utrecht – has been using the WeNMR infrastructure to investigate cone snail venom, a cocktail of small peptides that causes paralysis in the snail's prey. "The peptides bind to nerve cells and stop pain signals from reaching the brain," explains Hocking, "and we believe we can use the 3D models we obtain to design a powerful local anaesthetic."



Visualisation is extremely important in biomedicine, where molecular shapes hold the key to disease and treatment

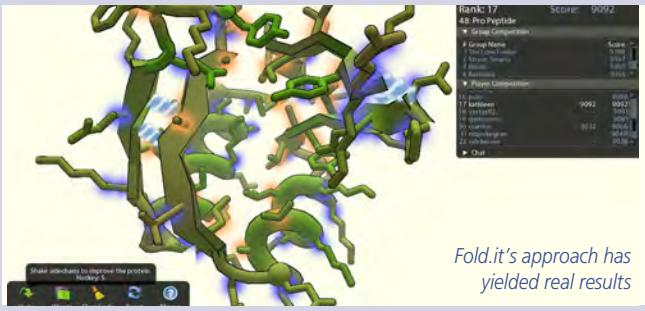
A visible change for chemists

The new wave of molecular sciences has much to gain from adopting e-Science methods in research, where chemistry meshes with materials science to produce molecular electronics, nanotechnology, novel solar power, communications, and display systems. The complexity of such systems means that modelling how they work can help to cut research times and costs dramatically. However, the interdisciplinary nature of such research means that many researchers are not fully familiar with the often physics-born technologies.

SOMA2 – a web browser-based workflow environment for graphical molecular modelling – has been developed at CSC, the Finnish IT Center for Science. "No Unix technical skills are required to access the powerful computational tools," explains Tapani Kinnunen, who has developed the gateway. "It's an intuitive and versatile visual environment that eliminates unnecessary repetitive work." The application, which has been funded by the EGI-InSPIRE project, runs on almost every web browser, and was launched in March 2012. MoSGrid, an application developed at Ludwig-Maximillians University in Garching, Munich, also aims to attract chemists, who can submit jobs to the grid – a network of computers linked together to share their processing power – without getting bogged

A democratic vision for science

Since the launch of SETI@home, volunteer computing has given us all the chance to contribute to science by donating spare computer cycles. However, in much the same way as joining a political debate is more empowering than listening to one, being able to actively participate in research projects by providing observations or interpreting visual information clearly separates what is known as 'citizen science' from volunteer computing. Two of the most popular citizen science projects, fold.it and Galaxy Zoo, allow those taking part to interact with, manipulate and assess visual data. Fold.it is presented as a web browser-based game, where players score points for finding optimal conformations of protein folding. In 2011, Fold.it players helped to find the structure of M-PMV, a virus that causes AIDS in monkeys



Fold.it's approach has yielded real results

Galaxy Zoo is another highly successful citizen science project, this time allowing users the opportunity to assess deep space objects from the Sloane Digital Sky Survey. Because the information is purely visual, participants need no training. "We thought about giving people tutorials and so on," said Chris Lintott, the academic behind Galaxy Zoo in a recent interview, "but quickly saw it would be more effective – and fun – to have people get going straight away, and use the sheer volume of observers to ensure accuracy. " The project has turned some accepted theories on their heads, even identifying new types of astronomical phenomena such as Hanny's Voorwerp.

Tim Adams, The Observer New Review, 18.03.12

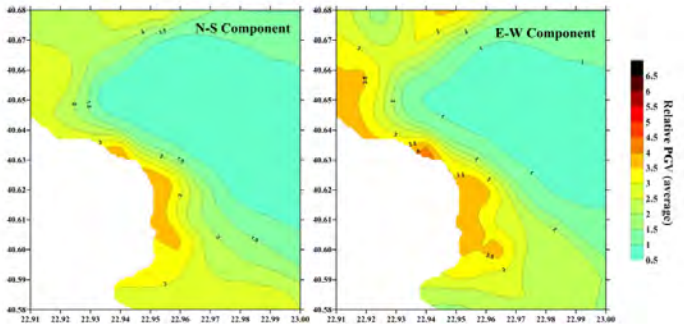
down in the technical details. "A user can submit their job – whether it's quantum chemistry or a molecular dynamics simulation – without having to know about the middleware," explains Sonia Helles-Pawla, of MoSGrid "and then just get their results back. For a chemist, that's saving them a lot of time." Of course, the more quickly and accurately results can be obtained, the quicker the chemist can visualise their structures using 3D models.

Mapping the Physical World

Physical processes in the real world such as water cycles, earthquakes and weather are incredibly complex, involving ecosystems with large numbers of interacting components. While DRIHMs and MAPPER are providing tools to understand hydrometeorological processes better, important work is also underway to understand seismic phenomena. Danseis, led by Hans Thybo at the University of Copenhagen, is visualising magma plumes beneath the Earth's crust using supercomputers. "It is believed there are around 30 plumes scattered around the planet," explains Thybo. "However, it isn't known for sure that they actually exist. Being able to model and visualise where they are on the globe can help scientists to better understand geological features and seismic events."

Hellasgrid: Seismic studies

Researchers based in the Geophysical Laboratory at the University of Thessaloniki are modelling the propagation of seismic waves as they might occur in a number of different scenarios, treating the Earth as a viscous, elastic resonator and taking account of local soil structures. Using Hellasgrid and the European Grid Infrastructure (EGI), the energy map they produced closely matches with a real map showing areas of intensive damage during the 4 July 1978 earthquake.

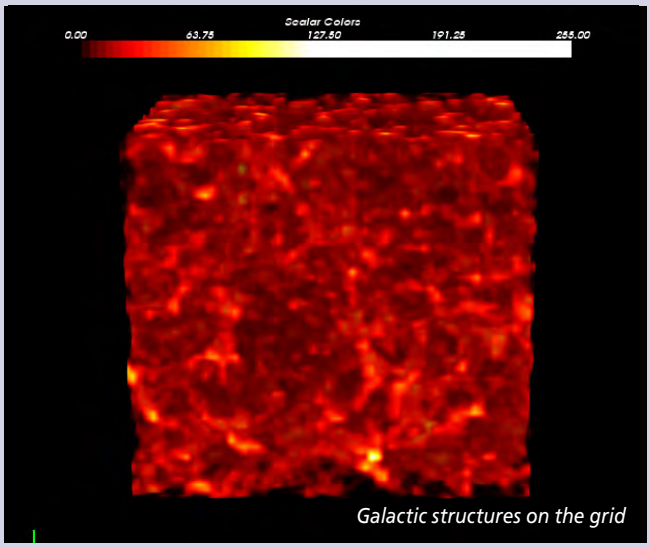


Modelling seismic waves in Thessaloniki

"It is quite a good correlation," explains Andreas Skarlatoudis, who developed the computational model for his PhD work. "It shows the model works." Accurate models like this could be used to predict which locations are likely to see high levels of damage in other earthquake-prone regions, and either strengthen buildings or avoid building certain structures, such as nuclear power plants, there entirely. Visualisations like this can help us to prepare for the worst, and can even inform how we react to data as a society, something that numbers alone often fail to achieve.

VisIVo: See the Universe on your iPad

"Being able to perform large scale simulations of the cosmos has traditionally been reliant on the availability of supercomputers," says Alessandro Costa, a technologist at the Astrophysical Observatory of Catania (INAF), "because most computational cosmology applications were developed for them. But supercomputers are often not readily available, and cannot be dedicated to a single project. This reduces the access time scientists have to be able to carry out their research."



Galactic structures on the grid

The solution, executing calculations in parallel on many CPUs, has not always been easy to achieve. "Running these applications in parallel on a cluster has also been problematic, because it requires very low latency to be efficient," explains Costa. Latency is the time it takes to move data around in between different machines in a cluster – essentially, wasted time. "However, in recent years grid infrastructures have begun to offer latencies comparable with supercomputers, making grid suitable for cosmological simulations."

Costa is a developer working on VisIVo, a visualisation and analysis application for accessing observational and theoretical astrophysical data. It provides a customised application-specific gateway to SCI-BUS (Scientific gateway Based User Support), an FP7-funded project that provides seamless access to major European distributed computing infrastructures, including clusters, supercomputers, grids and clouds. For VisIVo, the software running 'behind the scenes' on the grid is FLY, a program running on gLite middleware that can calculate the forces between hundreds of millions of stars to understand the effects of gravity between them. The VisIVo application itself runs on Apple iPads and iPhones, placing easy access to the grid (which does the calculation-intensive work) within the grasp of many researchers. "This is a fast way to access SCI-BUS," concludes Costa, "and of course this technology could be used by other projects."

What is middleware? : A number of general purpose software libraries must be installed on a grid so that all the machines in the gird 'speak the same language'.

Nicola Rebora, DRIHMS, hydrometerology study – *"Visual records from citizen scientists, such as videos, are extremely important in helping us to develop the hydrological models we use for DRIHMs. Indeed, we start with the requirements of the citizen scientists, and provide them with computational power, which is a complementary approach to volunteer computing. Their data can be exploited, empowering them to understand flash floods in their own area."*

Hellasgrid: air quality

In the Department of Meteorology and Climatology and the Laboratory of Atmospheric physics, researchers are using Hellasgrid and EGI to model the effects of climate change on air quality. Focusing on ozone, a gas that at low levels can react with other pollutants to create a toxic blanket of smog in cities, lead researcher Eleni Katragkou found that levels could increase drastically by the end of the century. "Until the 2040s, the increase in ozone will be less than one part per billion (ppb)," she explains, "But by the 2090s, it could increase by 6 ppb, especially in south-west Europe."

Human Mapping

Maps have a special place in data visualisation, and are probably our oldest method to represent information in an abstract form – one that transcends cultural and linguistic boundaries. London physician John Snow's 1854 map of the Soho cholera epidemic, which identified a focal point around which cases clustered – a contaminated water pump – spurred the local council to take action, promptly shutting the pump off and stemming the cholera cases. In this sense, maps are empowering. But they also have the potential to be democratising, both in the sense of making data understandable, and also helping minority groups to find solutions to problems that concern them.

Louise Francis, Mapping for Change, University College London – *"Mapping for Change works with communities across the UK to enable them to collect their own data. We then use mapping technologies to visualise that data. We've helped a community in Deptford, South East London, record air quality around a scrap metal yard close to where they lived, using nitrogen dioxide badges. Collecting data like this can empower citizens with their campaigning activities."*

Mapping our environment is the first step to understanding how to solve the problems we face, from disease to deforestation

Mapping for Change: Extreme Citizen Science

Muki Haklay, Louise Francis and Claire Ellul of University College London have been empowering communities in the UK and Africa by persuading the public to record their environment using visual means. Using simple measuring tools and GPS devices, projects such as Mapping for Change have brought real differences to communities in inner-city areas of London. As Professor of Geographical Information Science, Haklay wants to go further, transcending linguistic and literacy boundaries.

At the Citizen Cyberscience summit in London in February 2012, he discussed Extreme Citizen Science (ExCites), which aims to empower people regardless of their literacy. Funded by the UK Engineering and Physical Sciences Research Council, the project puts adaptable scientific tools in the hands of people in remote regions.

One project in the Congo basin of Cameroon, enables a forest community to collect data about their own environment using a highly visual, pictorial interface. "They can then analyse it and understand different changes that are happening in their area, monitoring aspects that they care about. It's a more interesting project now than when we first started." Using the pictorial interface, the forest community have mapped out resources and sacred sites important to them, helping to coordinate conservation efforts in areas where forestry companies also have an interest. Efforts like this could prove as important as John Snow's cholera map, enabling people to identify both problems and solutions for their changing environment.

Ultimately, visualisation is about taking tables of numbers, and representing them in a form that makes sense to us as human beings.

For more information:

- ***www.flowingdata.com***
- ***www.sci-bus.eu***
- ***www.drihm.eu***
- ***www.mapper-project.eu***
- ***www.mosgrid.de***
- ***www.hellasgrid.gr***
- ***www.csc.fi/english/pages/soma***
- ***www.communitymaps.org.uk***

EGI : www.egi.eu

Real Time Monitor: rtm.hep.ph.ic.ac.uk

iSGTW: www.isgtw.org

e-ScienceTalk: www.e-sciencetalk.org

email: info@e-sciencetalk.org

Scan this QR code into your smart phone for more on this e-ScienceBriefing

Books:

Edward Tufte – *The Visual display of quantitative information*

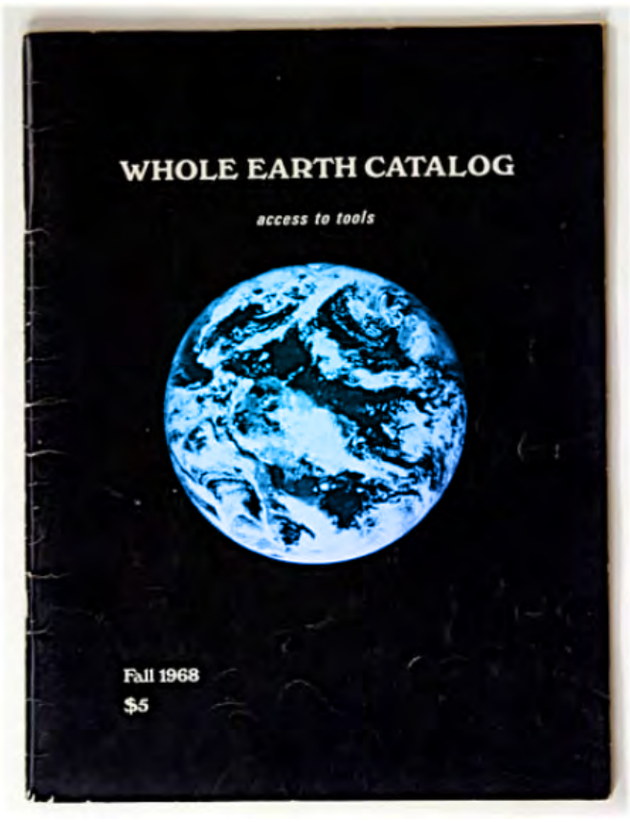
Michael Nielsen – *Reinventing Discovery: A New Era of Networked Science*

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733

Open Data, Open Science

Data: unavoidably expensive?

In the late 1960s and early 1970s, falling costs of integrated circuits meant the computer was making a transition from being a tool available to only the very few to one available to the many. As the potential for computers to be used to create, store and transmit ideas and information became apparent, technological evangelist Stewart Brand, publisher of The Whole Earth Catalog, identified a duality in the nature of digital data: "Information wants to be free. And information wants to be very expensive." [1]



The ease with which we are able to share information using computers shows how 'free' it can be. Compared to the expenses of print, the monetary cost of publishing information using the web is virtually nothing. There often remain costs associated with gaining access to scientific data on the Web, however. Sometimes, information is expensive. Scientists have long built careers by sharing their data – and staking a claim on it – through publishing it in prestigious scientific journals. Such journals often command high subscription fees even on the Web, restricting the flow of data to all but the very wealthy.

Slowly however, a quiet revolution has been gaining momentum: open access publishing, open science, and open data. The first is a change in the publishing model to one more suited to the age of the Web; the second, a change in how scientists connect with society – their major funders through taxation. Open data is even more revolutionary. Being able to share data more quickly and easily will accelerate the pace of scientific progress, and help scientists to solve the pressing problems of the 21st century: climate change, energy security and feeding the population. Open data is about sharing it freely – that means without restriction more than without monetary cost

Neelie Kroes, European Commissioner for Digital Agenda' – *"Sharing data, and having the forum to openly use and build on what is shared, are essential to science. They fuel the progress and practice of scientific discovery. That's why scientists have long sought out new tools and new ways to share their knowledge."*

When does free mean free?

The best things in life may be free, but that does not necessarily mean without cost. The word has two distinct meanings in relation to ownership, but they are often conflated. There is free, as in cost-free or gratis – you don't have to pay for it (although you may have to accept advertising). Then there is free as in libre – it comes with freedoms that allow the work to be adapted and reused. For open data and open science, libre is more important.

Software, media and data can be provided gratis, but may still be restricted by various levels of copyright, preventing or limiting a user's freedom to reproduce, reuse or adapt the work. For open data to work, the data must be 'freed' using a strongly permissive form of licence (see 'Licensing Open Data'). Software, media and data that are provided libre are also usually provided gratis – but actually they don't have to be.

1 'What the Dormouse Said', John Markoff, 2005, Penguin Books

Jenny Molloy, Coordinator, Open Science Working Group – *"Good science should be reproducible, but in many fields (not all) it is often impossible for other researchers to repeat and critically assess analyses without access to raw data. If researchers make a scientific claim they should ensure that the data is openly available to back up that assertion in a form that is reuseable by their peers for both independent analysis and inclusion in meta-analyses."*

Publishers: Opening Access

Open Access is an important step forward for open science, because it completely turns the old model of academic journal publishing on its head. Instead of charging universities and research institutes large sums to access scientific papers, Open Access publications charge authors a nominal fee (starting at around £500) per paper to publish their research, which is then made freely available over the Web. This means it is as accessible to scientists everywhere, for example in poverty-stricken regions, as well as the general public who essentially fund the research. Policy makers, governments, funding bodies and charities welcome the move because it sets the global stage for international innovation.

BMC: A Model for Open Access Publishing

BioMed Central (BMC) was set up by entrepreneur Vitek Tracz in 2000 in response to the shifting publishing landscape brought about by the advent of the Web. In the run-up to the millennium, requests for print journals were declining as scientists began to instead demand greater access to online publication repositories. The number of journals being published was meanwhile increasing, just as university libraries and research institutes struggled to afford the rising costs of subscriptions.

Tracz realised that scientists were prepared to pay to share their data – researchers generally want access to their work to be gratis, as long as they get credit for it. In that way, their main currency – their reputation – would have a wider reach. A greater global reach, arguably, than it would have if their work was restricted to only those that can afford to pay to see it. Restricting its publishing to online only has helped BMC to be profitable and has served as an example of how the Open Access model can work in a commercial environment.

Tagging open data is important to keep data searchable in the future (PDDL)

Iain Hrynaszkiewicz, Publisher, Open Science at BioMed Central – *"Publishing data and software online, whether included with or linked to journal articles, greatly increases the value and reproducibility of reported research. Publishers should embrace open data in response to scientists' needs and to drive innovation, but more efficient and reliable science is the ultimate goal. Open data is a way to help achieve that. Copyright is messy with respect to data and at BioMed Central we are working on implementing explicit public domain dedication of published data, to better facilitate data integration and reuse without legal barriers."*

Directory of Open Access Journals

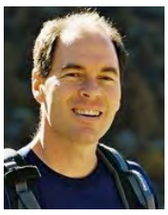
The Directory of Open Access Journals (doaj.org) now lists nearly 7000 journals that are available at zero cost. However, one barrier to truly open science is that even in Open Access, publishers can choose to impose restrictions on the use of the content they publish. Only around a fifth – including BMC, Public Library of Science (PLoS), and a number of smaller publishers – allow reuse and adaptation in the libre model. BMC (and PLoS) journals are covered by a Creative Commons Attribution licence (CC-BY), ensuring scientists get credit for their freely distributable works. Announced at the Open E-Infrastructures for Open Science, hosted by ALLEA (ALL European Academies), UK biomedical foundation, the Wellcome Trust and the World Bank have similar initiatives to open up the work that they fund.

Wouter Los, Project Leader of Lifewatch, e-infrastructure for biodiversity research– *"Understanding our environment requires large volumes of data of very different kinds. We assume that we now only capture a few percent of the data that we would like to have available. Open data are crucial, as modern interdisciplinary environmental science cannot deal with limited data sets. The same holds for society. Environmental management is dependent on sufficient and reliable open data."*

OpenAIRE

OpenAIRE
Open Access Infrastructure for Research in Europe

Understanding how central open data is to scientific advancement, 20% of the budget of the European Commission's 7th research framework (FP7) is dedicated to making the science it funds open and accessible. OpenAIRE (Open Access Infrastructure for Research in Europe) is the result: a project set up to establish and operate an electronic infrastructure for handling peer-reviewed articles, enabling researchers to deposit their final peer-reviewed manuscripts and/or post-prints either in an institutional repository or a subject-based repository. It also provides support structures for researchers wanting advice on how to make their research open, which for ERC – European Research Council – projects is often a condition of their grant.



Tim Smith, Collaboration and Information Services Group Leader at CERN – *"The strength of science has always been its open dialogue on the results and conclusions of experiments. Since data is recorded in such volume now that it cannot be communicated effectively via the results tables of scientific papers, data sets themselves need to be accessible independently to ensure the scientific hardening process. Furthermore, sharing can allow more knowledge to be derived from a data set than in the original research."*

The limitations of copyright

The legal tool of copyright, which was introduced to the world through English law in the late 17th and early 18th Centuries so that authors could be sure of a fair income for their work, has perhaps been more of a hindrance than a help in the drive towards open data. Many publishers require a 'transferral of copyright' from scientists wanting to publish with them, including those behind some of the most prestigious journals. Publishers argue that full transferral unburdens scientists from needing to assert their authorship and retain control over their own work. However, copyright can also restrict how data is re-used, and as science becomes increasingly data driven, access to the data sets can be as important as access to the paper itself. Scientists wanting to reuse the material of others, even if they cite it properly, could find themselves in breach of copyright if they do not ask the publisher's permission and pay any necessary fees, which may again affect scientists in poorer countries.

Making sense of open data in the future

As more and more data becomes available publically through the move towards open data, common shared tools are needed, so that scientists can sift through it and make sense of it in the future. One particular mechanism that meets the requirements for data organisation is the application of metadata – data that describes data. Organisations like Open Data Foundation (ODaF) are dedicated to the adoption of global metadata standards and the development of open-source solutions promoting the use of statistical data. To facilitate the sharing of data, the Europe-wide EUDAT project (eudat.eu) implements a secure means of sharing data using persistent identifiers, similar to the way written documents have been given ISBN and now digital object identifiers. Agreeing on such standards will make it easier to share data between disciplines, and to sort through mountains of data decades after it might have been generated.

Making sense of open data in the future

To solve the problem of licensing open data and protecting authorship, one solutions is the concept of "copyleft" – a play on copyright, and the practice of using copyright law to actually keep data open. Richard Stallman, a computer scientist at MIT, created the GNU Public Licence, GPL, after finding that he was legally unable to reuse some of his own code, which he had previously given freely to a corporate developer. The GPL ensures that any software released under it may be used, adapted, changed and freely distributed by its users, but that any copy or derivative work is covered by the same licence. In effect, it uses the legal system of licensing to prevent prospective developers imposing a commercial licence on GPL software-derived work. GPL can be good for some types of software, but is not always appropriate for creative works such as scientific publications.

Of all copyleft licences, perhaps the most well-known are those from Creative Commons (CC), whose Sharealike (CC-SA) licence is perhaps the closest to GPL. A common misconception is that Creative Commons is equivalent to public domain – and that a user can effectively do what they like with it. In fact, CC licences are precisely worded legal documents that use terms from the legal concept of copyright. Open data requires licences that are unrestrictive. For CC, the least restrictive and most appropriate for open data is CC0, which effectively releases a work into the public domain. However, scientists can still be sure of receiving credit for their work because the cultural norm of citation exists separately to the notion of copyright. Further information can be found at *pantonprinciples.org*

Open Standards for raw data

Open standards and raw data are fundamental to a functioning and long-lasting open science movement, and it is imperative that standards are agreed upon and adhered to. Data must not only be legally accessible, but also technologically accessible 20 years down the line despite changing software trends. This fate has already befallen some of the earliest digital archives, such as the BBC Domesday Project – an attempt to produce a digital historical record of life in the UK in 1986. The software, stored on laserdisc in the LV-ROM format, would only run on an expanded Acorn BBC Master computer with a specially produced laserdisc media drive. A few years after production, the computers were obsolete and the data was inaccessible. For data to have the best chance of surviving long into the future, it should be in its most 'raw' format. Open standard formats should be free from proprietary ownership, and simple to 'future-proof'. They could include UTF-8, for text and numerical files (.txt files, in other words), PNG for pixel-based images, SVG for vector images (e.g. technical drawings, scalable logos etc.), and Ogg Vorbis and Theora for audio and video. [3]

Virginie Simon, founder & CEO of MyScienceWork – *"MyScienceWork is dedicated to open science. Our platform enables researchers and engineers from all disciplines to communicate, share, and discover. Scientists can use our innovative search engine to access tens of millions of professional articles. By facilitating the accessibility of knowledge and reinforcing scientific communities, we promote accessibility and visibility of research. I think this is only the beginning of the transformation of how scientific data is organized and shared, and a whole new era of open science."*

ScienceSoft: Open Software for Open Science

Much of the most-used technical and scientific software is open source. From statistical software R, used by scientists across a range of fields; biochemistry application DOCK; to programming languages Python, C and Ruby, and technical typesetting package LaTeX, open source software has always played a major role in scientific computation. Open source software development is a democratic, inclusive enterprise, with some of the major projects involving thousands of volunteer programmers. The quality of software developed this way matches and exceeds commercially-developed software: a reminder of the wisdom of crowds; 'that none of us is as smart as all of us'.

So many software packages are available across the various open source repositories that scientists may struggle to find the most appropriate software for the task they have in mind. Often this may be made worse by there being several variants or 'forks' of a project, some available on only certain repositories, stored among a wealth of non-scientific software. The levels of support on user forums may be similarly variable and disparate.

2 'See Tim Berners-Lee, 'Raw Data Now' at is.gd/rawdatanow

Alberto Di Meglio, ScienceSoft Project Leader – *"Open source software is based on values like transparency, collaboration and availability. These values are also at the base of Open Science. An active, vibrant community of software developers and users contributes to making global scientific research more accessible and reproducible. New scientific and societal challenges are becoming more and more complex. They cannot be addressed anymore just by clever individuals, but by open collaborations on a global scale. Open source software is a fundamental part of this transformation."*

Initiated in December 2011, Sciencesoft builds a virtual software repository and support network coordinated by the European Middleware Initiative in collaboration with the European Grid Infrastructure, StatusLab, iMarine, OpenAIRE and other e-infrastructure projects. It brings together a wealth of software expertise to help research communities to find the software they need. Users can rate software, which provides useful feedback for developers and helps funding agencies to understand the software use of research communities.

R is a scientific programming language used by scientists in all disciplines that has been developed as an entirely open source package.

For more information:

JISC Legal Open Data guide: discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf
Panton Principles on Vimeo with Iain Hrynaszkiewicz: vimeo.com/34555054
Panton Principles: pantonprinciples.org
Data Definition and tagging: opendefinition.org
Semantic web: www.w3.org/2001/sw/
ODaF : opendatafoundation.org
EUDAT: eudat.eu
CC: creativecommons.org
Science Soft: sciencesoft.org
EGI : www.egi.eu
Real Time Monitor: rtm.hep.ph.ic.ac.uk
iSGTW: www.isgtw.org
e-ScienceTalk: www.e-sciencetalk.org
email: info@e-sciencetalk.org

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733

Transferring Technology and Knowledge

In 1991, Tim Berners-Lee, a computer scientist working at CERN, gave us the World-Wide Web. This tool, designed to make navigating information easy, is perhaps the most famous example of technology transfer to come out of e-science. It has revolutionised how we communicate; the global economy; even how we live our lives. Its pervasiveness in society was even celebrated in the London 2012 Olympic Games opening ceremony.

Though predating the foundation of an official e-science funding programme in the UK by a number of years, the tremendous computational requirements of experiments at CERN tie the foundation of the Web to what has since become known as e-science. This somewhat well-worn example does not stand alone, however. From the developments in supercomputing in the 1980s, which gave us the multiple processor cores found in the advanced mobile electronics of today; to improved information storage for huge experiments, which have helped create 'the cloud', the backbone of the Web 2.0 economy – e-science impacts many areas of industry and commerce. Bioinformatics, a field largely made possible by technological advances in e-science, has revolutionised R&D in the pharmaceutical industry. Still other advances impact on earthquake prediction and oil exploration. But it's a two-way street for both ideas and people: computer games, now multibillion-dollar business, make use of the technologies and talent of the e-science world, just as games technologies feed back into e-science..

Supercomputer in the palm of your hand

Seymour Cray had a virtual monopoly on supercomputers for two decades, with his eponymous series of machines famed for their number-crunching prowess. His mantra, 'anyone can build a fast processor, the trick is to build a fast system', coined at a time when processing power was expensive, had held true for two decades. Eventually, his designs were superseded by a paradigm shift in computer architecture: multi-processor machines, which started being built in the 1980s. At the time, whether these clusters of processors counted as single supercomputers was debated, partly because the mechanics of getting the processors to work together had not been fully worked out[1].

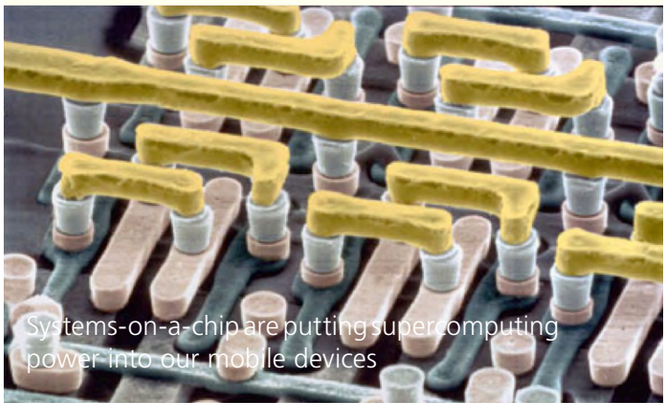
The message passing interface (MPI) standard, introduced in the early 1990s, allows processors to effectively 'talk to each other' and distribute computational load. The first implementation was worked on, among others, by Tony Hey, later head of the UK e-science programme. It has allowed software engineers to develop the concept by introducing bridging tools to make use of multiple processors, including commercial solutions such as Microsoft Compute cluster and e-infrastructure middleware such as gLite.

Chip manufacturers such as Intel, IBM and ARM have been switching to 'multicore' for several years, as heat restrictions have meant that cramming more transistors onto a die, or increasing the number of cycles per second, simply wouldn't work. More recently, system-on-chip (SoC) has meant that mobile technologies such as smart phones and tablets are virtual 'supercomputers in the hand', allowing such devices to run high-powered

The Web was celebrated in the 2012 Olympic opening ceremony

1/ Today, there would be no question – not only supercomputers, but even home computers, games consoles and mobile devices can contain more than one CPU core.

software. VisIVO, a project featured in the e-science Briefing on visualisation, has ported its astrophysics simulation system to Apple's iOS, for instance, and tablets are widely used in the medical sphere.


Systems-on-a-chip are putting supercomputing power into our mobile devices


Tony King-Smith, Imagination Technologies – *"Advancements in multicore and parallel processing technologies will continue to be the key driver in the mobile space as heterogeneous computing keeps blurring the lines between the PC and embedded market in terms of overall system performance. The graphics processor - the quintessence of what parallel computing is all about - has now become the most important part of a standard System-on-Chip. This has enabled smartphones and tablets to display resolutions higher than your regular TV or bring the quality of console gaming into the handheld realm while allowing consumer electronic products to get smaller yet faster with each generation. With new APIs, like OpenCL, that parallel processing capability will also enable more image recognition, photo enhancement and augmented reality applications to become mobile"*

Peopleware: Transferring Knowledge

Silvia Olabarriaga worked in industry before moving into academia, but is insistent that her peers, including those like her who work in e-science, have also learned to think like entrepreneurs: "To survive in academia, just doing good science isn't enough. You have to be able to lead people, to write grant proposals, and to manage the overall operations of a research group. All of those skills are useful to industry."

Olabarriaga leads the e-Bioscience group at the Academic Medical Center in Amsterdam, having come from the software and computer graphics industries. She holds a doctorate in medical imaging. She is also passionate about peopleware – her idea that people are absolutely critical to the functioning of e-science; as important as hardware, middleware and software. "Communication is extremely important," she says, "some people are more adept at solving complex issues, but you also need good communicators, just like in any organisation."

'Capacity-building' is an often bandied-about term, but the skills gained by working in large, multidisciplinary and geographically dispersed teams are manifold, she says: "the business of the future will also be very distributed. It's going to be about teams of people with complementary skills and ideas working together. That's where the nice 'apps' are coming from."

Olabarriaga thinks cloud computing is being adopted much more quickly by industry than grid (a model successful in academia where research institutes share computing power over a network) because it equates to a service model that business leaders understand. "In commercial settings you need a very clear business model. With cloud, the business model is clear."


Silvia D Olabarriaga, Academic Medical Center, Amsterdam, The Netherlands – *"People who are exposed to big e-science projects undergo a change in the way they communicate. When recruiting for my research group, I look for people who have worked in these projects – they give you the mindset to see that, 'OK – it's not not just my part of the programme, it's bigger than that…I have to understand the whole, and know how to communicate my part of the problem with others.' To be able to go beyond the desktop is important because this is also what the industries of the future will need."*

Cloud: Product to Service

Box, Dropbox, Google Docs and Apple iCloud: consumer platforms for cloud services. The key word here is 'service'. In many cases, that service is storage, whether that's from online backup utilities such as Crashplan, or simply document sharing. But it doesn't stop at storage or even document and calendar synchronisation, although these are the services that, as consumers, many of us use day-to-day and see as being synonymous with cloud. Indeed, the Centre for Economics and Business Research in the UK estimated in 2011 that cloud computing could add €763bn to the European economy before 2015.

Simon Wardley, geneticist, former CEO for a division of Canon, and now researcher at the Leading Edge Forum, sees cloud as simply representing "the evolution of a bunch of activities from across the computing stack from products to services"[2]. While businesses might be told by that replacing their IT departments with services cloud might save them money, Wardley warns against this. Referencing William Stanley Jevons' consideration of more efficient steam engines in 1865, Wardley explains that businesses adapt to changing economics of operation. Cloud is more efficient, because it allows businesses to focus on their defining activity – "[but] does this reduce IT spend? No, because consumption of services goes up to compensate." What cloud does offer, he suggests, is "greater agility. It's faster…and you can do more stuff".

Yelp, Foursquare, and Hipstamatic all make use of Amazon's elastic cloud, otherwise known as EC2. Compared to the cost of setting up their own data centres, this has undoubtedly reduced the time needed to bring their services to market. There are also long-term benefits to using standard systems in a business model, akin to the adoption of standard screw fittings developed in the industrial revolution.

2/ OSCON 2010: Simon Wardley, «Situation Normal, Everything Must Change»


Cloud fits business well because of its service model

If competitors are using standard services, "ubiquity means that such activities will have diminishing strategic value, and become just a cost of doing business," explains Wardley. The competitive gap that opens up by not using standard services means that you lose focus on your defining activity.

e-science is feeding directly into this cloud-based computing ecosystem. HPC-in-the-cloud allows scientists to access virtual supercomputers running on Amazon's EC2. Just as for businesses, using standard services gives researchers a competitive advantage This is something that has particular resonance in the world of corporate R&D, but whether privately or publicly-funded, scientists and those working in science policy need to understand the entrepreneurial requirement for competing on a world stage. At a recent conference on e-infrastructure standards in Denmark, EGI Director Steven Newhouse framed the issue succinctly: "How many people here are flying home after this conference… [OK] – and how many people are building their own airport and aeroplanes to do so?" Concentrating on core activities doesn't only give private and public-sector researchers an economic competitive advantage, it also gives them more opportunities to innovate: "The main benefit of cloud is accelerated innovation," explains Wardley.

Tony Hey, Corporate Vice President for Technical Computing, Microsoft – *"Just as the scientific community sees an emerging data deluge that will fundamentally alter disciplines in areas throughout academic research, a wide range of industrial and commercial businesses are now beginning to recognize that data-driven business decision-making and analysis will be necessary to stay competitive. The magnitude of the data and the complexity of the analytical computations will drive both scientists and business analysts to explore the use of cloud computing, high performance computing systems, multicore processors, and parallel programming to manage the computing workload for these efforts."*

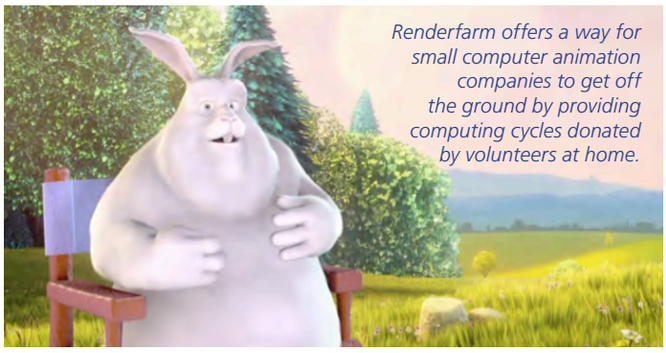
Computer Graphics: Innovation driving business

Computer graphics were once seen as a frippary by many computer scientists, but they have become increasingly important tools to communicate ideas and findings across disciplines (for example, see Briefing on Visualisations). e-science makes use of developments in computer graphics chips originally designed to be better for more immersive games. It's an example of commodity hardware, with prices driven down by economies of scale, feeding back into the e-Science ecosystem.

GPUs really began to be adopted by e-science with the advent of separate graphics 'cards', which were designed to meet demand for ever more realistic 3D worlds. The graphics co-processors in these cards were optimised to draw lots of polygons needed for 3D games very fast, making them highly suited to the parallel processing of mathematical algorithms. This makes them very useful for performing simulations of natural processes for e-science applications, such as molecular simulations for drug discovery in medicine and astrophysics simulations ever more realistic 3D worlds for computer games. This is called GPGPU (General Purpose computing on Graphics Processing Units).

Many of the techniques used for 3D visualisations in fields such as medicine were first developed for the entertainments industry. Realistic shading and depth of field can be crucial for clinicians deciding on a course of action when examining 3D recreations of internal organs from MRI scans, but the tools used to do this were perfected bringing characters like Woody from Disney's Toy Story to life.

Renderfarm.fi takes some of the ideas of desktop grids and volunteer computing and moves them into an entertainment sphere. The philosophy is that people want to contribute and like to help, whether that's towards solving scientific problems in citizen cyberscience projects such as SETI@Home or volunteering free computing cycles to creative works for entertainment purposes .


Renderfarm offers a way for small computer animation companies to get off the ground by providing computing cycles donated by volunteers at home.

GPU History

The first wave was discrete graphics processing units (GPUs) in some 8-bit and 16-bit 'home computers' in the early 1980s were produced to offload some of the burden of drawing objects on the screen from the CPU to a separate custom graphics chip (including the Atari 400/800 and Amiga[3]). These chips were very good for moving around 2D objects on-screen at a time when processor power was expensive, and 16-bit iterations were capable of broadcast video production and photorealistic static graphics. The second wave came with IBM PCs running DOS and Windows. Economies of scale associated with ubiquity meant their CPUs became much faster and cheaper, negating the advantage given by including graphics co-processors. Games such as Wolfenstein 3D and Doom, which ran on 386 and 486 processors,

3/ Maher, J. 2012, The Future Was Here

opened the floodgates in terms of what people expected from games, eventually leading towards high-powered dedicated graphics hardware, including graphics cards by ATI and NVIDIA. Today these are being used to build GPU supercomputers, such as EMERALD at the Rutherford Appleton Laboratory in the UK, which is used to model everything from pandemics to galactic simulations.

Ian Osborne, Knowledge Transfer Network – *"There are clearly big opportunities for those working in e-science to teach us in industry how to cope with large data – which has suddenly become a hot topic and which will only become more important with the 'internet of things'. There is suddenly a shared working space that industry and the e-science sphere can work together effectively in. Small companies can also use technologies developed in academia – properly licensed – to help fulfil their product pipeline without so much investment in R&D "*

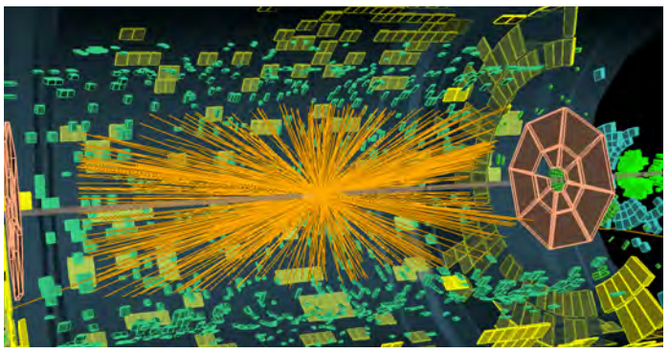
Intellectual Property in Tech Transfer

Intellectual property (IP) can be an issue for technologies coming out of e-science because it is largely embedded in academia. It is not always obvious who owns the intellectual property developed by an academic working in a university, although some institutions are now addressing this with agreements written into job contracts. Exactly who should own IP is still being debated, although a growing number of those working in science policy believe that it should be owned by the university rather than an individual professor. As engines of knowledge generation, its is argued, universities are well placed to properly licence technologies to industry and use the profits for future societal benefit. Indeed, one of the outcomes of individuals owning IP is that academic publishing could capitalise on it by asking them to sign away copyright on published research, which has hindered open access publishing. (This couldn't happen if the institute owned the IP).

The European Association of National Research Facilities (ERF) is a collection of European research facilities that have adopted an ERA open access policy, publishing their research for free for the benefit of those working outside of their walls. It includes members such as CERN, DESY and research centres of the UK Science and Technology Facilities Council

ATLAS: Not just measuring the massive

The technologies being developed for big experiments at CERN are, of course, widely disseminated thanks to an innovation in computer technology that went unpatented: the Web. But just as many of the advances for medicine came from the physical sciences (x-ray crystallography elucidating protein structures, radioisotopes from cyclotrons), experiments like ATLAS at CERN, which explores superheavy particles heavier than the Higgs boson and top quark, are themselves contributing to wider areas of human endeavour.



- Superconducting coils and cables are being explored in higher-density magnetic storage media, and could lead to even better power lines.
- Diamond-based sensors used to detect the fragments of particles yielded by high energy collisions could be used to monitor doses in hadron therapy, used to treat childhood cancers.
- 3D silicon detectors might be used for advanced medical imaging
- Pattern recognition technologies could be used for augmented reality, to allow automation of complex procedures by machines.
- Drift tubes can detect scattering of cosmic rays by dangerous materials, including those that could be used to make a bomb, hidden in containers.
- Medipix, a project based on ATLAS and other experiments at CERN, has been operating since the 1990s. The ability of detectors at CERN to count single photons with no noise is highly applicable to fields such as medical imaging, hence medipix.

Summary

Transfer or people, of ideas and of technologies continues to feed into and out of the e-science ecosystem. There are sometimes challenges in commercialising ideas coming out of academia, but scientists are becoming more adept at doing so as larger cultural changes take hold. Commercial models like cloud equally finding a place in public research settings. There are certainly exciting times ahead as academia and industry learn to work more closely to achieve their goals.

For more information:

- ***ktn.innovateuk.org***
- ***www.astp.net***
- ***research.microsoft.com***
- ***setiathome.berkeley.edu***
- ***www.renderfarm.fi***
- ***www.atlas.ch***
- ***www.cloudbroker.com/***
- ***www.stfc.ac.uk/About+STFC/51.aspx***

EGI : www.egi.eu

Real Time Monitor: rtm.hep.ph.ic.ac.uk

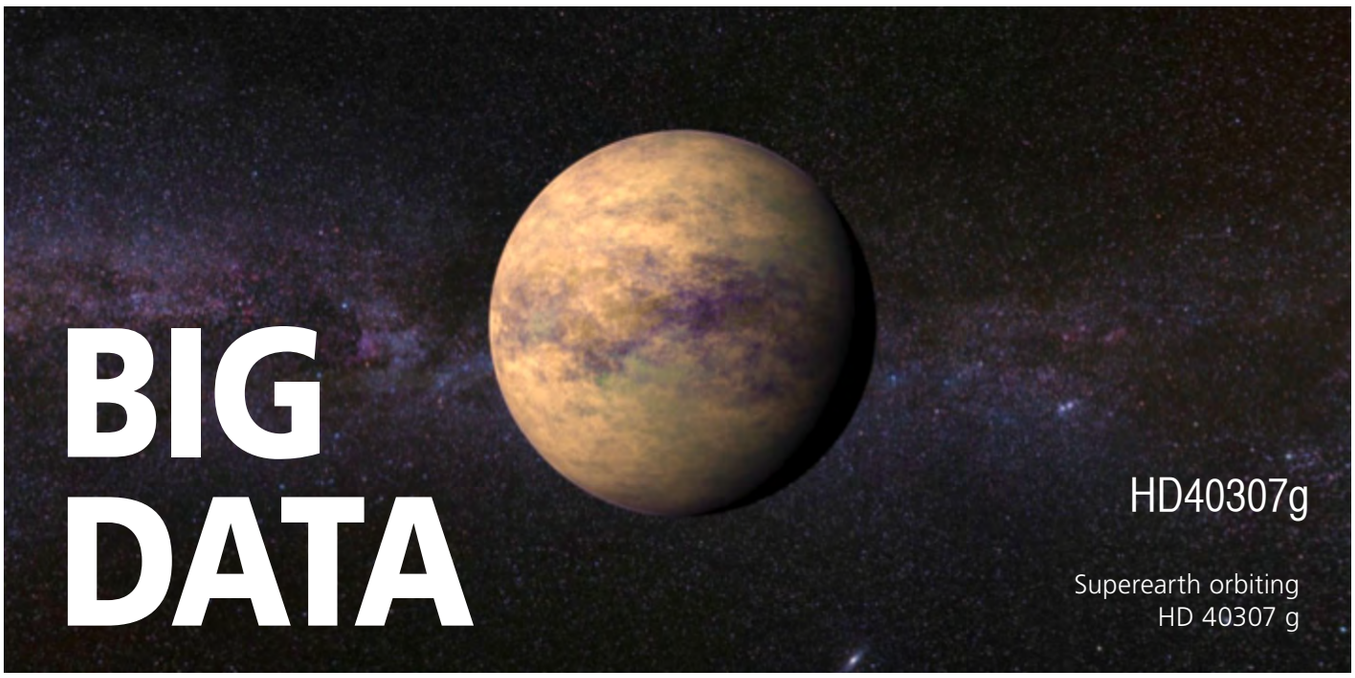
iSGTW: www.isgtw.org

e-ScienceTalk: www.e-sciencetalk.org

email: info@e-sciencetalk.org

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733



In November 2012, Mikko Tuomi of the University of Hertfordshire and Guillem Anglada-Escude of the University of Göttingen announced the discovery of a new 'Super-Earth' – a rocky planet five times larger than ours – orbiting around the habitable zone of its parent star, where surface water would be liquid. They did this by analysing old data sets using new methods. This discovery demonstrates the importance of keeping and curating data so it can be reused later. But as science continues to produce a deluge of data, is keeping it all even viable – and will a future researcher from a different or even completely new field be able to understand it? This challenge has led to the concept of 'Big Data'.

Big Data is about the petabytes of results from particle physics, systems biology and Earth simulation science – how we deal with that volume of data and how we use it. But it's also about the variety of data being produced. Life sciences, social sciences and cognitive sciences produce data of many different types, including images, for example, as well as text-based data, so categorising and storing it all becomes a challenge. And in medicine, as data becomes obtainable at an ever-faster rate, there is an opportunity to mesh data from different source – from physiological feedback and genetic screening – to determine the course of intervention particular to individual patients.

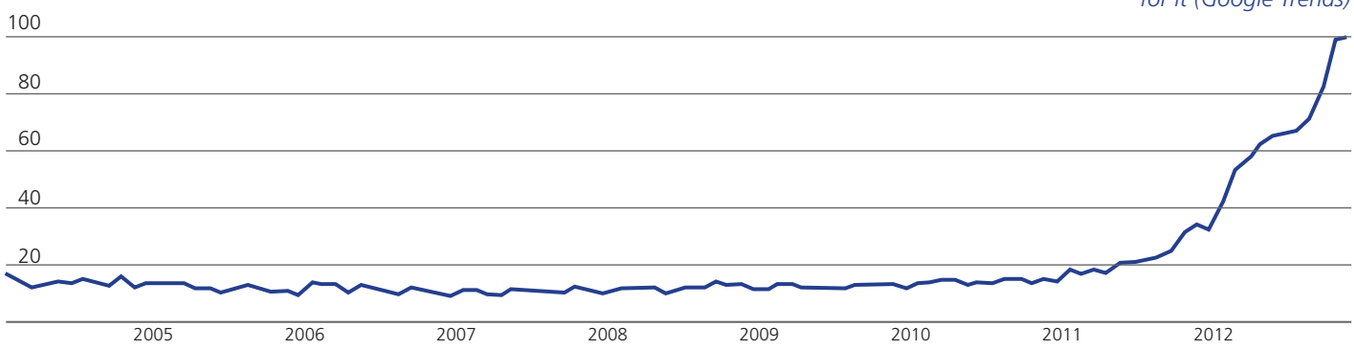
Big Data is not just confined to science: it pervades other areas, including commerce and government. Many online retailers and search engines know so much about our interests and buying habits that they feel confident enough to tailor advertising and suggest products that we might like to buy. Some of the time, at least, they get it right. Science works differently to commerce, however. Science still needs theories to operate and to make sense of that data. One area where that distinction may be blurred is smart cities, where science and technology are employed to regulate our living processes. Already, masses of data once kept under lock-and-key is being shared by governments openly, allowing app developers, for instance, to tap into data on public transport, or refuse collection, and then present it in a useful way to the consumer.

Interest over time

The number 100 represents the peak search volume

Forecast

Big data is a big deal, and more people are searching for it (Google Trends)



Data as Infrastructure

The term 'data infrastructure' was used by US President Bill Clinton in 1994, but it has actually taken some time for the idea that data is infrastructure to catch on. In the global knowledge economy, data is a basic component that leads to commerce and economic growth. Big Data owes its existence to e-infrastructures that governments have invested substantially in, but it was to facilitate the handling of Big Data (principally from particle physics at first) that e-infrastructures were built in the first place. Since that initial conception, Big Data has widened its scope to include diverse data including all manner of text- and number-based data, graphics and audio files, and increasingly metadata.

Data is infrastructure: astronomical data are encapsulated in the gears of an orrery, while data is the basis of the stock market, and of whole economies

Laurence Field, CRISP Data Management – *"Data management is a key aspect of what we're trying to do and covers many topics: data archiving, data preservation, persistent identifiers of data, data access and identity management. Physics research infrastructures have their own bespoke solutions, so we're trying to provide common solutions where possible. There are also new research infrastructures coming online that have additional requirements and we are trying to include these in those common solutions. Even though the requirements of research infrastructures may be different, we're able to identify common challenges, for example identity management, so there are always areas where we can work together."*

Metadata: Making sense of data

Data provides information and information leads to knowledge. To make sense of the data in the first place, you need to describe and manage the data. Metadata is data about data. 'Tagging' photographs, weblinks, even music in the apps we use is something that we're all used to; it helps us organise our lives. For the website delicious.com, tagging Web pages with useful category-words has helped users build their own signposted guide to the web and share those guides with others – here, metadata is a way to signpost points of interest[1]. This concept was later adopted by other social media platforms such as photosharing and blogging sites, and a particular implementation of metadata – the hashtag, used by Twitter and other microblogging sites – has entered the popular consciousness[2]. Even the file-and-folder analogy that was useful for personal computers and small networks has given way to the new paradigm of search, whether on our personal machines or 'in the cloud', helping users find files even if they can't remember what they called them.

In science, metadata is an important way of organising data to deal with it now, but it also makes it more searchable, and therefore useful (and, crucially, reusable) in the future. This is especially important when researchers, funders and governments are under pressure to demonstrate 'added value' to publicly funded research. Without metadata, the rush to make large datasets open – which is one way they are being asked to add value – would leave future researchers and those from other disciplines with a meaningless glut of information. Furthermore, solutions to the Grand Challenges of climate change, pandemic, biodiversity and energy security are expected to require innovation at the interface between traditional scientific and engineering disciplines, necessitating cross-disciplinary collaboration.

It is extremely important, therefore, that metadata is universally understood by scientists, no matter what field they work in. This means that standards are important and should be flexible enough to make sense to researchers from fields not originally anticipated to have an interest in the data, and even from fields that didn't exist when the data was collected. Collecting data costs money curated and preserved so that it can be reused to extract the maximum possible benefit from it.

"Without context, data rots," said Ross Wilkinson of the Australian National Data Service. "It needs to be integrated into other datasets and publicised." Creating a scientific 'data commons' not only aids discoverability, connectivity and value, explained Wilkinson, but is also incumbent on scientists funded by the public: "Data coming off an instrument that is publicly funded is intended for the community…it's not for the sole use of the grant holder, but is intended to be shared."

Metadata is data about data, helping scientists to find it in the future

1 /Scrupulous employment of metadata in what is termed 'white hat search engine optimisation' helps Web users find the content they're looking for, and is enshrined in principles of good Website design

2/ having initially been widely adopted by IRC earlier in internet history

EUDAT

Enshrined within European Union Law is the free movement of goods, capital, people and services between its member states. To take on the grand challenges of climate, energy, food and global health requires a fifth freedom: data. The European data infrastructure project EUDAT, which has been running since 2011, was set up in response to the recommendations of the High Level Expert Group on Scientific Data. It aims to meet the challenges of the rising tide of scientific data in Europe. The project has benefitted from being able to take lessons learned from other initiatives around the world, such as dataone in the US (since 2009) and the Australian National Data Service (ANDS, since 2007).

Peter Wittenburg, Max Planck Institute for Psycolinguistics and EUDAT Scientific Coordinator – "During the early internet it became obvious that there are cross-disciplinary common services that could be used by all research disciplines. The same email system could be used by physics and humanities researchers, for example. In the same way, we're realising now that there are common components – building blocks – for global data infrastructures. One crucial building block we're working on is a world-wide registration and resolution system for persistent identifiers. Every single piece of data would have a PID, just as computers connected to the internet have a unique IP address."

EUDAT offers common data services relevant to a wide spectrum of communities, which it has achieved in part by having worked from inception with five exemplar projects from different research areas: LifeWatch, covering biodiversity; ENES, covering climate modelling; EPOS, seismology and vulcanology; CLARIN, for linguistics, and Virtual Physiological Human (VPH) for medicine. Working with a wide range of disciplines has helped EUDAT to specify the requirements those initial projects have from a pan-European data infrastructure. This in turn has helped codesign a range of services useful to researchers across all major fields. EUDAT have also extended their reach to communities from across the biomedical, environmental, physical and social sciences and humanities.

Matthew Dovey, Programme Director, JISC – *"In some branches of applied science, being able to make accurate predictions can be of more practical importance than understanding the underlying models. For example: determining future weather patterns, or choosing between different but established medical treatments based on a patient's lifestyle. Here, Big Data can be used to identify trends and patterns with improved reliability. Ever increasing sophistication of analytical tools may even one day replace the role of the theoretical scientist in hypothesising new models. Scientists then have the task of devising experiments to challenge and test these computer-generated models."*

Coordinating effort

What specific skills are required to make use of these emerging data infrastructures? The social infrastructure is something the UK body JISC[3], which looks after digital strategies and standards for post-secondary education in the UK, is looking into. The SIM4DRM project is gathering evidence at an international scale in order to develop a model of best data management practices. The project will provide a 'cookbook for all stakeholders, organisations, policy makers and funders'. Another initiative, the project Knowledge Exchange, has a vision of 'making scholarly and scientific content openly available on the internet." They are particularly interested in the areas of Open Access, Licensing, Repositories, Research Data and Virtual Research Environments. Many changes are needed to establish the open-access publishing environment including quality training/incentives (e.g. more journals for data publications) and understanding of the benefits and costs of re-using publishing and archiving data sets.

Life Sciences

One of the biggest challenges for Big Data in life sciences is the variation in terminology - even groups working on the same disease but in different model organisms will often use different terms for the same thing. BioMedBridges aims to link together datasets so that researchers can find the information they're looking for even if they don't know the terminology specific to the model organism it has come from.

Stephanie Suhr, BioMedBridges – *"BioMedBridges is about making the most efficient use of life sciences data, linking up existing resources and creating bridges between different research communities to get them to agree on common data standards and formats. This involves cultural as well as technological challenges – different communities can use different terminologies. In one project use-case we are linking diabetes and obesity-related data from human patients and from mouse models, which requires translation between the terms used by both research communities. Systematic use of extensive mouse data resources by clinical researchers will be an extremely powerful tool for new scientific discoveries and therapies."*

Standards

Kimmo Koski, Director of CSC in Finland, joked at EUDAT's first conference in Barcelona that metadata in Europe would be in Finnish, but actually imposing such a top down standard for metadata is problematic. Recognised by EUDAT and others is that the scientists themselves must be involved in the development of standards, otherwise they could run the risk of being irrelevant to researchers. The eventual goal is to have metadata auto-generated and data tagged with standard, searchable terms as it is collected. A basic example already in place is the way some search engines allow searching for images with a particular colour, which is achieved by simple image analysis as the Web is crawled.

3/ Joint Infrastructure Services Committee
4/ Support Infrastructure Models for Research Data Management

CLARIN – Speaking the right language?

A big step towards the 'semantic web', where information is linked together in a smart or intelligent way, is turning natural language queries, such as "list all the instances of e-science in European policy reports from 1998", into filtered searches. Commodity services that aim to provide natural language search have improved vastly over the last few years. CLARIN's aim is leverage the nuanced precision of natural language queries into accurate searches of data in the social sciences, a field dubbed eHumanities. In addition to being one of the exemplar projects for EUDAT, CLARIN has linked up with other projects working in linguistics to form the DASISH consortium, which takes its name from the projects it comprises: DARIAH, CESSDA, ESS and SHARE, as well as CLARIN. All work in complementary areas of linguistics.

One of CLARIN's tools that is already in place is the Virtual Language Observatory – a search tool linking to a vast Europe-wide corpus of linguistic datasets covering everything from psycholinguistics – how the brain learns and interprets language – to endangered languages, especially from rapidly changing regions such as the Amazon basin, including those the indigenous Trumai, Aweti, and Kamayurá peoples.

Tape library, CERN, Geneva 2, holding masses of data. By Cory Doctorow / CC BY-SA

Virtual Physiological Human

Peter Coveney, Director, UCL Centre for Computational Science – *"Big data is precisely 'where its at' for the medical domain. It's feasible to generate vast quantities of data, especially from gene sequencing – which can now be done very quickly – potentially a few minutes for an entire human genome. We're faced with the challenge and opportunity of marshalling it. Virtual Physiological Human is concerned, like many projects, with accessing patient data and using data mining and analytics techniques, but also in the business of modelling and simulation, which is used quite extensively in the physical sciences and engineering, but far less common in medicine and biology. Merging these together – e.g. a CT scan and simulation – will allow truly personalised medicine where a surgeon can be armed with the best information to make the right decision."*

How Big Data is changing science itself

To Galileo, *esperienza* – what his senses could tell him about the Universe – was key to unlocking its secrets. Trusting his own experience (a word that shares the same root as experiment) over the doctrines of classical scholars, whose ideas were 'common-sense' but often scientifically inaccurate, caused a knowledge revolution and lay the foundations of the modern scientific method. Experience, refined by controlled experiment, has allowed scientists to determine the nature of reality and describe it with ever-greater clarity. Data drives decisions in science and other areas of human endeavour, challenging scientists and policymakers to refine their understanding of How Things Really Work.

Big Data in science is a challenge requiring input across and between disciplines, and even outside the realms of academic science towards the citizen scientist. But there are tremendous benefits to having so much data available to science: for one, it allows us to test and modify theories as never before, with greater accuracy and agility. Big Data, like Galileo's *esperienza*, could be more revolutionary than evolutionary, because the availability of large data sets could spark off new areas of enquiry. It is already doing so in the field of systems biology, which couldn't exist without Big Data[5].

The solutions that e-science comes up with, in terms of e-infrastructures, will help lay the foundations for environmentally responsive smart cities that rely on the Internet of Things – where every electronic device is networked to make our lives easier and the impact we have on the environment smaller. Two years ago scientists, perhaps worried about the data tsunami, were invited to ride the wave[6]. Big Data is about meeting the challenge of riding that wave and sharing how to do so with others.

For more information:

- ***www.sim4rdm.eu***
- ***www.icordi.eu***
- ***rd-alliance.org***
- ***www.clarin.eu***
- ***verc.enes.org***
- ***www.epos-eu.org***
- ***www.eudat.eu***
- ***www.dataone.org***
- ***www.ands.org.au***

EGI : www.egi.eu

Real Time Monitor: rtm.hep.ph.ic.ac.uk

iSGTW: www.isgtw.org

e-ScienceTalk: www.e-sciencetalk.org

email: info@e-sciencetalk.org

Scan this QR code into your smart phone for more on this e-ScienceBriefing

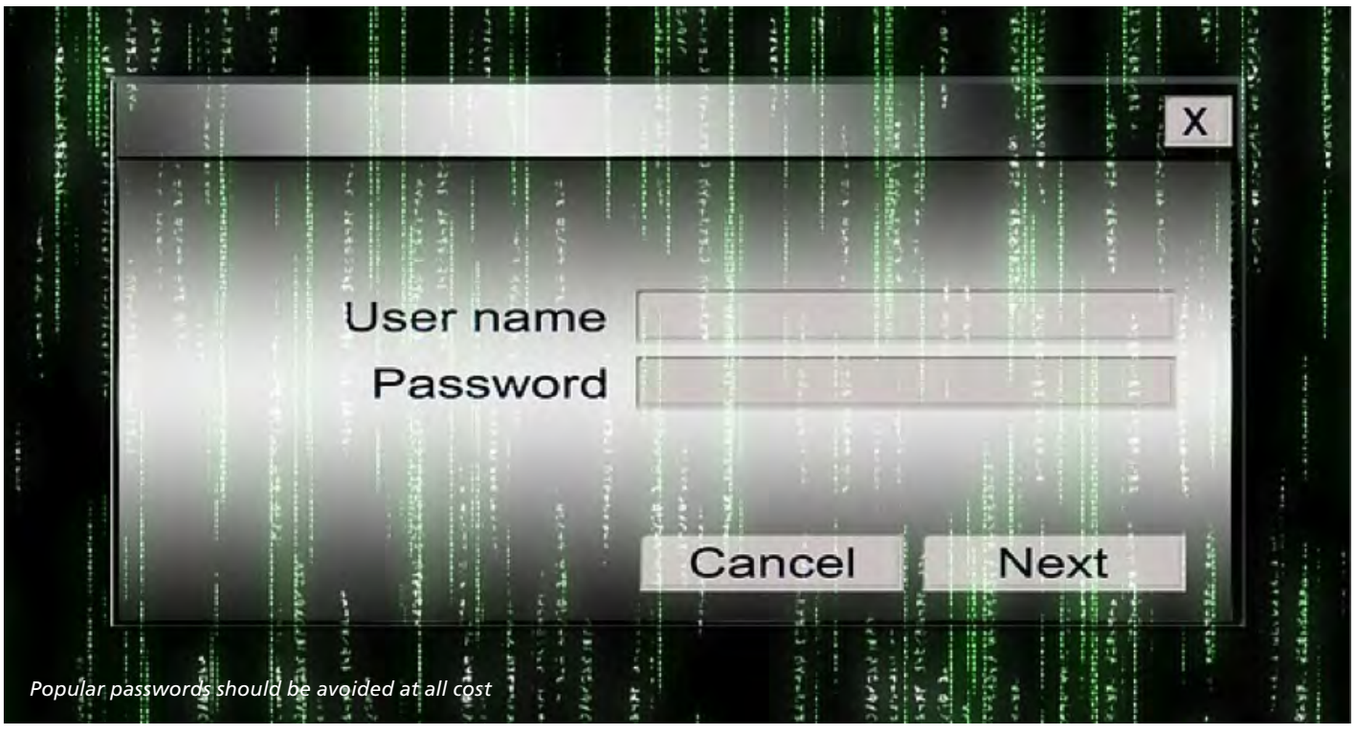
5/ "Systems Biology wasn't possible until we had the data" – Ross Wilkinson, Australian National Data Service

6/ "Riding the wave: How Europe can gain from the rising tide of scientific data" EC, 2010

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733

Security and e-Science

'Password'; '123456'; '12345678'. The top three most popular passwords of 2012, as published in lists by hackers, were identical to the top three of 2011. When it comes to security, popular passwords aren't to be celebrated – not only are these passwords easy to guess, but it's a safe bet large a majority of easily guessed passwords in 2012 were 'protecting' the very same files they did in 2011. And, more often than not, those same passwords are also duplicated across a range of online services. This creates an easy target for identity thieves, whose intent is much worse than those who publish passwords online. What the fact that such lists can be published highlights is that total security is elusive: whether passwords are easy to guess or not, they are sometimes liberated, even from the biggest sites in social media and online gaming. Is the password reaching a crisis point? And what could replace them?



Popular passwords should be avoided at all cost

e-Science faces the same challenges as the rest of the online world – not least because many researchers are online outside of work, just like everybody else. But there are specific concerns: e-Health will herald a new era of personalised medicine, but having your personal file compromised could reveal more about you than you would like. Any large corpus of knowledge could be vulnerable to attack by cyberterrorists, and the hyperconnected 'smart cities' of the future might be an attractive target for acts of cyberwarfare. That is, if they're not quite smart enough to outsmart the bad guys.

e-infrastructures such as the grid also have a long history of managing security, access to services, and controlling privileges. These concepts are becoming more and more important to the rest of the online world, as the idea of universal 'web identities' takes hold. Indeed, multifactor authentication solutions ('two stage sign-ins') employed by sites such as Facebook and Google generate one time passwords to be provided in addition to your normal password.

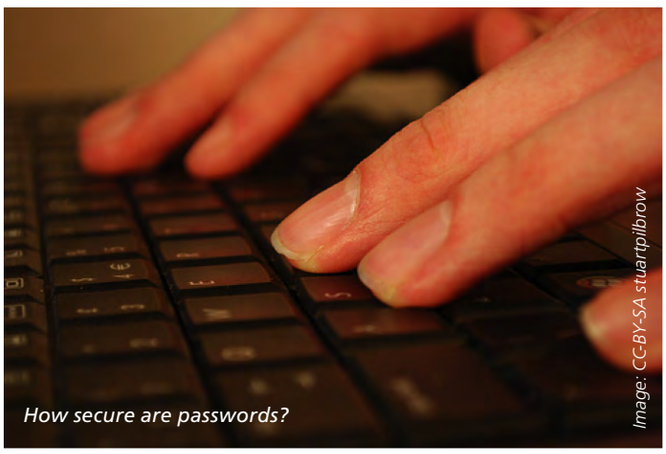
At the same time, e-science services are beginning to adopt social media credentials to allow users access. While this may simplify access from a new user's perspective, opening up grids and academic clouds to more researchers in the life sciences and e-Humanities, it can present new security challenges.

Stephan Lüders, CERN Security – *"Computer security is a sociological problem. It is time to teach our users and colleagues to stop-think-click when browsing the Internet as they've been taught to look both ways when crossing a road."*

Grid Security: Certificates

With many people becoming overwhelmed by the growing number of web-based services they use daily, the concept of a universal web identity seems like a sensible solution. For researchers using the grid, this idea is familiar in the form of digital certificates. Certificates are files that reside on a user's personal device that contain, alongside information such as date and duration of validity, a special key that is unique to the user, generated by a certificate authority. When a user accesses the grid, their credentials are checked to see if they are authorized to do so by the certification authority.

Certificates may be familiar to people outside of the grid community using services such as OpenVPN to securely access their employer's network. This typically requires a certificate issued by your employer to be installed; some proprietary Virtual Private Networks, such as that provided by Cisco, are configured to use your work login and password for authentication. Websites also use security certificates, but this may only become apparent to the average user when the certificate expires.



How secure are passwords?

Image: CC-BY-SA stuartpilbrow

Certificates have some advantages. Because they identify an individual, losing or having a laptop stolen that is validated by having a certificate installed only requires a single certificate cancellation request be made. This stands in contrast to the many different logins and passwords the typical user has for web services, which would all have to be changed individually if a laptop was stolen – some browsers contain easily accessible lists of logins and passwords used, for example.

The idea of a single web identity, therefore, with a single sign-on seems to have some value in the rest of the online world. Several protocols including OpenID, OAuth Connect and Facebook Connect (a proprietary protocol) have arisen. The latter two, respectively, allow Twitter and Facebook credentials to be used to sign in to all kinds of web-based services. Due to the prevalence of social media these are becoming de facto standards, even for some grid services. There are concerns about the security risks of using social media as a universal sign on for e-science services (which

Roberto Barbera, Chain project – *"With the same simple sign on, a user could access everything from their campus network via Eduroam to the entire global grid. This is tremendously powerful"*

is plausible, because social media is such a huge target for online fraud). However, these "are removed by retaining a distinction between authentication and authorization," says Roberto Barbera, Technical Coordinator of the CHAIN project, who are providing access to grid services using social media authentication.

That authorization could come in the form of certificates or portable IDs such as the Shibboleth system from the UK organization, JISC, which is what CHAIN is using. The user can then access grid infrastructures around the world, including EGI, Open Science Grid, Teragrid, GISELA, SAGrid and Garuda. "You have to remember that authentication is completely separate from authorization," says Barbera, "identity federations allow us to control access, but we can control the privileges a user has separately." The report 'Advancing Technologies and Federating Communities' produced by Terena suggests that more researchers are using the social web to collaborate, and that e-science services should provide access via social credentials. However, a research institute needs to be sure that the individual presenting social credentials is the same as the individual they have authorized to use the service.

OpenID Connect

At the time of writing, OpenID Connect, a suite of lightweight universal ID standards, is in the implementer's draft phase. It aims to offer an alternative to Facebook Connect and OAuth and is said to be an improvement of the previous version of OpenID.

Clouds: The Safest Place?

It's a recurring story: an individual working for an organisation providing some sort of public service loses a USB key, or has their laptop stolen. The files contained on the stolen item contain the personal details of thousands, tens of thousands – or more – individuals. Worse, the data was unencrypted, meaning anyone could access it easily. Amidst public anger that their personal information could be so easily accessed, new security measures are put in place. Usually the solution chosen is encryption, so all data, whether sensitive or not, must be encrypted, whether on a laptop's hard drive or on a USB stick.

But does that solve the problem? It's a technological solution that's fairly easy to implement, but it ignores the fact that, in many cases, poor practices undermine its usefulness. Password-based encryption is only as effective as the strength of the password. And if the

device is left logged in when not in use, it may as well not have been encrypted.

Clouds have the potential to offer much better security. Access can be controlled to files, to greater or lesser degrees for the individual collaborators working on those files, as required. Thanks to desktop synchronisation and version control (which saves incremental changes to files at many points in time over the lifetime of the file) storing everything in the cloud seems like the perfect solution. So much so that some netbooks only allow saving to the cloud. But which cloud?

Keeping data on USB keys is a risky business. CC-BY-SA Scienceatlife (Flickr

Popular cloud services such as Dropbox, Box.com and Google Drive have been widely adopted by researchers (and many others) because they offer easy ways to collaborate on or share files and data. The reliability and of these services is very high, but as with any online service there always remain vulnerabilities and potential for attack at every level (from a user's personal machine and network, to outsourced online helpdesks with lax security)[1].

As e-Health, which will offer unparalleled diagnosis capabilities and personalised therapies, becomes a reality, there is a need for clouds that operate independently of the privately run clouds. Stratuslab in Europe produces software that allows researchers to build academic clouds on their own hardware. These are not subject to the same terms (or potential for change of terms) as commercially available services. For researchers in many fields, a long-term goal is the establishing of repositories for 'Big Data' coming out of computationally intensive science that stand apart from the 'free' commercially run cloud services, with the different terms of service they entail.

Just as for jobs on the grid, controlling access is a key issue. Strong passwords, certificates, or logins tied to a machine's hardware all have their place.

1- Dropbox, for instance, had a major security flaw in 2011, and in February 2013, Zendesk, an online help system for Twitter, Tumblr and Pinterest, which also looks after Box.com support, experienced a security breach: http://www.wired.com/threatlevel/2013/02/twitter-tumblr-pinterest/

The Password Problem

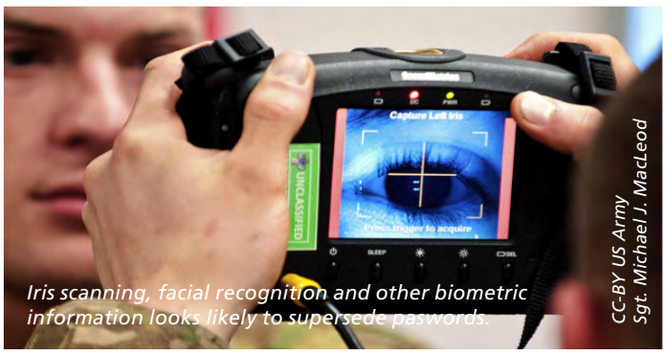
In spite of major efforts to educate users around the importance of using unique passwords, it seems that many users disregard warnings that their data is at risk of being compromised. According to reports given in several high-profile hacking cases involving attacks on state-level systems there are some serious short falls in security. Passwords are routinely distributed indiscriminately, rarely or never updated, even displayed on post-it notes in areas that can be physically accessed by individuals who would not otherwise have been given the password.
When it comes to judging the security of a password-protected system, a concept called Kerckhoff's principle is often invoked: a system should be secure even if everything about it, apart from the password, is public knowledge. For open standards advocates, this is a cornerstone of good encryption practice – the 'security by obscurity' used by proprietary software engineers avoids Kerchoff's principle precisely by not making the workings of the system public knowledge. However, if there are loopholes (and there often are), hackers can find them and circumvent any security measures in place. In open source software, the 'many eyes' working on the same code are more likely to spot loopholes, and many minds working on fixing the loopholes will do so more effectively.

But in practice, passwords in software, whether proprietary or open source, suffer the problem that people forget them. Many online services offer a 'I forgot my password' option at login, which will send a password reset code to your email address. While this might seem sensible it could be the first step to having your entire online life hijacked. Even services that offer extra layers of protection, such as requiring you to provide personal details, can be easily duped into sending a password reset to an unauthorised third party. Service providers have to manage high levels of password reset requests; people just have bad memories, so its unsurprising that they are so ready to be helpful in helping us get to our accounts.

Passwords are either so difficult to remember, therefore, that they have to be reset, or so easy to guess at that they provide no protection whatsoever. Perhaps passwords alone are no longer the answer. Two-step verification, now being adopted by services such as Google and Facebook, combines passwords with device-dependent ID keys that have to be set up for first time use. This is only secure if users set their devices to lock out during periods of non-use.

Sven Gabriel, NIKHEF – *"In a distributed environment like the European Grid Infrastructure, operational security has an additional dimension, since here we have to coordinate the activities of many different security teams involved in a multisite incident" See how distributed teams combatted a simulated virus on the grid at: http://v.gd/gridsecurity*

In 2012 researchers used the computing prowess of supercomputers to show that even the strongest password keys could be broken by brute force. The falling cost of computing power means that, in the long term, the age of the password is drawing to a close.
Biometric ID authentication based on facial or iris recognition looks likely to play some role in how we use devices in the future. Such technologies have been in place for a number of years at national border controls, and are becoming more commonplace in mobile devices. How web freedoms can be maintained in a future where our bodies become our logins and passwords is likely to be an area of intense discussion.



Iris scanning, facial recognition and other biometric information looks likely to supersede passwords.

How secure is your platform?

The choice of hardware and software platforms used in working environments greatly affects their susceptibility to infection by malware, the safeguards against data loss, and their overall security. UNIX-derived operating systems favoured by the e-science community have historically tended to be more secure than proprietary alternatives like Microsoft Windows for the simple reason that the former offer greater granular control of access and editing privileges to files. UNIX-like systems including the many derivations ('distributions') of GNU/Linux and also Mac OS X (from FreeBSD/Mach) observe a clear distinction between users and administrator. Many of the security loopholes in Windows were historically caused by systems being installed in administrator mode by default, which lets malicious code be surreptitiously written to locations in the system without the owner's knowledge; many of these instances of code being written would have alerted users on UNIX-like systems by demanding a password. However, Microsoft's increasingly swift update cycle has improved security greatly in recent years.

Indeed, Microsoft's Windows is conspicuously absent from a recent list of top ten vulnerabilities produced by security software firm Kapersky, yet is still the most targeted platform, mainly due to market share. Linux is the most popular platform in e-science for other reasons that also benefit its security: as it is an open platform, 'many eyes' are involved in checking the code for loopholes. And as this is performed openly, it is the embodiment of Kerckhoff's principle. Additionally, the many different distributions allow users to tailor the platform for their research. This creates a diversity that is as sound a defence against virulent malware in the online world as it is in agriculture. Monocultures in plants or computers means threats can propagate quickly with devastating results.
Any operating system that grants application plugins such as Java or Flash special privileges can, however, compromise security. They effectively create a virtual monoculture. Examples of malware on various systems has subsequently led to the latest versions of the Google's Chrome and Mozilla's Firefox browsers being released with Java turned off by default. HTML5, an open standard, is being promoted as an alternative for developers building web-based services.

A growing trend for employees to 'BYOD' – bring your own device – also presents challenges that differ from those experienced before. BYOD presents challenges to organisations aware of the productivity boosts adopting such a policy can make, because it can introduce security risks depending on the device. Apple's iOS ecosystem is generally more secure than Android due to the fact that each submission is checked for content and functionality, leading to a closed app ecosystem. Users wishing to 'jailbreak' out of the so-called walled garden who have not subsequently protected their devices from unauthorised access can present security risks accidentally, but the Android ecosystem, where applications can be freely distributed without undergoing checks, presents the bigger risk.

Summary
e-Science faces the same challenges of authentication, universal identity management, and authorisation (including privileges) as many other web services. But with the number of researchers using such services in light of the growing importance placed on Big Data for life sciences and e-Health, for example, it is important that access to them is properly and securely controlled. The changing nature of how people use the web for the rest of their online lives is also influencing how people access e-science services, and it looks likely that universal web identities might prevail over the anonymity of the early days of the web.

For more information:

Stefan Lüders' blog post on passwords: ***http://security-blog.switch.ch/ 2013/01/09/password-awareness/***

Advancing Technologies and Federating Communities : ***http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf***

CHAIN: ***https://www.chain-project.eu/***
OAuth: ***http://oauth.net/***

EGI : ***www.egi.eu***

Real Time Monitor: ***rtm.hep.ph.ic.ac.uk***

iSGTW: ***www.isgtw.org***

e-ScienceTalk: ***www.e-sciencetalk.org***

email: ***info@e-sciencetalk.org***

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733

e-Science in Horizon 2020

Horizon 2020, the European Commission's next funding cycle, is set to launch in January 2014. With less than a year to go, you may be wondering: what is Horizon 2020? What makes it different to the frameworks that preceeded it? Why the break from the simple numbered iterations, FP7, FP6…which began, as you might expect, with FP1…all the way back in 1984?

The commission is the executive arm of the European Union, a unique union of nations that has recently found itself the recipient of the Nobel Peace Prize – an acknowledgement of the successes of the European experiment – just as economic difficulties threaten to cause social and political unease across the union. It is safe to assume that politicians throughout Europe will be scrutinising the outcomes of EU-funded projects more closely than ever before. The overarching goal is to strengthen the economic and social ties within Europe, and to bolster the European economy in the longer term.

In EU member states, the economic situation means there is a renewed focus on investments giving returns in the short term, and this includes research funding. Indeed, for the EC too the economic vision has always been more pragmatic than purely Keynsian: there is room for 'blue sky' research, but projects have long been required to aim for sustainability post-funding; now, there is likely to be a greater emphasis on public-private partnerships. For research e-infrastructures, new funding models are being tested. Variations of pay-for-use, a model familiar both to industry and increasingly to academic research when it comes to cloud services, are being tested for grid. The need to act synergistically; to coordinate at the level of national research centres; minimise overlap when it comes to large-scale funding; emphasise centres of expertise and optimise governance, are all now acknowledged.

Neelie Kroes, Vice President, European Commission – *"As the Commission President has stressed, this budget can still be a catalyst for growth and jobs, and a tool to boost our competitiveness. The significantly increased investment Horizon 2020 will be making in EU research and innovation, including in the field of ICT, is a very vivid illustration of that. This is investing in tomorrow's growth; and by acting at European scale we can ensure research and its benefits spread as widely as possible, including across borders."*

e-infrastructures in 2020

At its heart, the focus of Horizon 2020 rests on three pillars: excellent science, competitive industries, and better society. These are the broad objectives that are hoped to be achieved in Europe by Horizon 2020. At the 10th e-infrastructure concertation meeting in Brussels, Kostas Glinos, Head of GÉANT & Infrastructure within the EC's DG Connect information society and media provided an overview of Horizon 2020 and what it would mean for e-science. e-infrastructures being developed need to reflect the societal and policy needs of Europe; they must integrate into the planning phase the specific innovation activities to be supported – the scientific projects they will allow. They must also go beyond science, reach out to industry and work for the benefit of society.

An important part of the Horizon 2020 strategy is a review of how e-infrastructures used for e-science operate; to maximise coordination and identify and build upon synergies across member states, so that successful areas can be developed while minimising overlap of effort. De-siloing is important: infrastructures built by and for specific research communities are often of use to the wider community, and they should be identified and made universally available where appropriate. Similarly, governance should be optimised: expertise in managing

projects should be identified and nurtured, or streamlined where required. The Horizon2020 framework preparations have identified areas of e-infrastructure development that already have solid foundations, in addition to areas where work is still required (such as the full HPC strategy implementation, still in progress), with the ultimate goal to 'make every researcher digital' by 2020. A set of funding instruments to help achieve this has been outlined (see box). The requirement for Open Access publication of data from ERC-funded science has been mandated since 2002, and the successful development and deployment of the OpenAIRE framework allows researchers to deposit their publications online where they can be freely accessed by other researchers. Open Access is an important stepping-stone to Open Data, the open sharing of data that will help researchers to more efficiently focus efforts when working to tackle some of the biggest challenges of the 21st century (see e-Sciencebriefing 22: 'Open Data, Open Science' . In addition to this, the international Research Data Alliance has been set up to push for convergence of standards on how data is stored and categorised. This data about data is called metadata, and was covered in e-ScienceBriefing 24:'Big Data'.

Research institutes distributed across countries and continents can benefit from the connectivity e-infrastructures allow

In addition to the 'instruments' or tools available within the funding structure to help foster innovation in e-science, there is a HPC strategy, aiming at an integrated approach to HPC for industry and academia. HPC, or high performance computing, is often referred to as supercomputing. There are scientific and technological goals, notably the move into exascale computing, which in simple terms means a thousandfold increase in the number of calculations possible. This will lead to better and more accurate simulations across a range of scientific fields, from biomedicine to engineering and climate models. From the industrial side, public-private partnerships in HPC should bring economic benefits by tapping into supply and demand, for example by pushing forward innovation more rapidly and allowing for quicker adoption of disruptive technologies. This includes the use of GPGPU (general purpose computing on graphics processors), acknowledged to offer better performance when modelling complex phenomena such as climate, with the added benefit that graphics processors

Kostas Glinos, Head of e-infrastructures unit, EC *"The goal to make every researcher digital, through the development and deployment of e-infrastructures, and to achieve the digital ERA [European Research Area]. We need to look at not just how e-infrastructures can help with scientific research, but also how they can be used to address the needs of society overall."*

are becoming much cheaper much more quickly than CPUs (e-ScienceBriefing 23: Transferring Technology and Knowledge). Finally, there will be a focus on fostering centres of excellence for HPC applications or specialities in particular fields. These will form central hubs much like the established research institute (e.g. CERN, DESY) and help realise the ERA or European Research Area, allowing researchers from across Europe to tap into expertise in a very streamlined fashion.

The right tool for the job: Funding Instruments for Horizon 2020

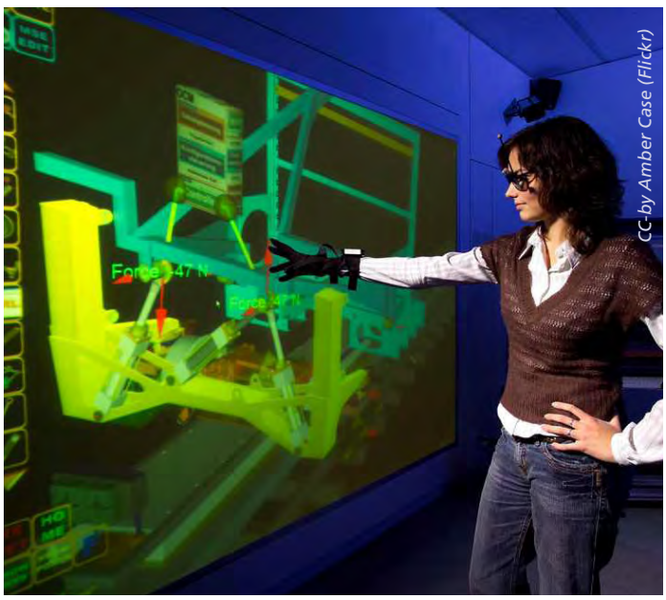
2020 Under Horizon 2020, there are a number of instruments available to the EC. For research and development, grants can be given up to 100% for each objective in a work package; for innovation, it's 70%. There remain places for coordination and support actions, examples in FP7 being the e-infrastructure reflection group (e-IRG) and e-ScienceTalk projects. Small-to-medium enterprises are allowed to apply to participate for grants at three stages: concept and feasibility assessment; R&D, demonstration and market replication (including prototyping, piloting); and commercialisation. To stimulate innovation in the private sector that can benefit the public sector, pre-commercial procurement (PCP) can steer development to fit public sector needs, while public procurement of innovation solutions (PPI) ensures that innovation in the market is rewarded by the public sector being the first buyer of these innovations. Prizes both recognizing innovations already taken to market, and also inducing companies to take those that are promising to market, also have their place.

Advancing Synergies with e-Science

Global connectivity between researchers themselves and superfast connections between existing centres of excellence have, arguably, a more important role to play than ever. Funding the building of new research institutes is less justifiable in tough economic times when centres of excellence already exist, but clearly superfast research networks are not simply a cheap substitute for a new CERN. The goal of making every European researcher digital and the changing model of how research is being done makes these networks, such as GÉANT, an integral part of the fabric of the European Research Area. 'New'

institutes are likely to be centres of excellence with particular synergistic specialisms, geographically spread out but linked together through networks and virtual research environments.

Identifying synergies in research is an important first step. The Cluster of Research Infrastructures with Synergies in Physics (CRISP), for example, identifies and builds on commonalities in four areas: accelerators, instruments and experiments, detectors and data acquisition, and IT and data management. By sharing expertise and finding common solutions, CRISP hope to contribute to improving the efficiency of spending in large physics-focused research institutes. In biomedicine, on the other hand, BioMedBridges are linking up various life sciences communities through innovative e-infrastructures that, for example, allow sharing of disease data between scientists studying models of diabetes in mice and known data from humans. This kind of comparison, which was previously very difficult, is now possible and through these kinds of innovations, synergies are also being developed and exploited in the life science communities.



Virtual research environments could take any shape: from an electronic lab book to a virtual reality physics lab.

Virtual Research Environments (VREs) are a diverse range of collaborative platforms for researchers from across the sciences and humanities that support users in a variety of ways. On one level, they can act as an enabler of open-notebook science – either within specific communities or completely open to the public – allowing data to be recorded and shared with others. VREs can automate aspects of data recording, including metadata – for example by using a smartphone's GPS ability to record geographic location alongside other data recorded by researchers, such as for biodiversity studies. Modular, customisable workflows are another aspect of some VREs, which can be used by researchers to speed up the process of getting to the experiment and data collection stage. This is being done both for individual communities, such as solar sciences (CASSIS, for heliophysics) and hydrology (DRIHMS), to more generic tools for building gateways for a range of research communities (e.g. SCI-BUS). The diversity of nature and scope of VREs means that they represent more of an evolving ecosystem than a specific type of platform. There is a need to continue this flexible approach to user-configurable e-science environments, to help build communities and define what is needed.

Current EU initiatives have been been successful in the fields of cross-border collaboration, particularly in respect of mobility and harmonised access. TERENA's (Trans-European Research and Education Networking Association) eduroam® allows researchers to access wi-fi facilities in institutions across Europe and the rest of the world. GÉANT's web Single-Sign-On (SSO) service, eduGAIN, provides secure access to remote collaborators whilst maintaining data and access security. The ELCIRA initiative, meanwhile, will provide Web Conferencing, Wiki collaboration and data sharing secured by eduGAIN between Europe and Latin America.

Michael Krisch, CRISP coordinator, ESRF, *"Competition in science has been something that has been very beneficial in the past because it stimulates progress. But as research infrastructures are financed by taxpayers' money, we have to be increasingly vigilant to spend funds in a very efficient way. One way of doing that is to identify commonalities across the european research infrastructures and find common solutions to problems, which we can then implement."*

Pay-for-use

Pay-for-use is a model familiar to many in the world of cloud computing; one of the reasons that cloud has taken off in industry while grid has struggled is that businesses understand the idea of paying for a service – paying for what they use. The marketplace in cloud is subsequently enormously diverse, a benefit due in part to the free marketplace that can quickly adapt to changing customer requirements. Competition means that commercial cloud providers now sometimes provide free cloud services to individuals with the offer of upgrading to a 'pro' package – more suited to a heavy user, or a business – for a recurring fee. The Helix Nebula project is piloting the use of commercial cloud services for eScience with a pay-per-use model in a public-private partnership that brings together major public research organisations, such as CERN, EMBL and ESA, with a range of European cloud service providers. So could the pay-for-use model work to bring similar benefits to grid computing?

Scientific research is not a business: there is the need to be cautious when considering pay-for-use, as it is a big change to scientists who have become used to central funding such services. On the other hand, some parts

of the community are already using commercial cloud providers to do everything from supercomputing, to storing large datasets. As they are already exploring the supposed greater agility, dynamism and adaptability of these services, they could be ready for a pay-for-use grid.

"Until e-infrastructure providers try these models, how would we know what works?" asked Sy Holsinger, Strategy and Policy Officer at the European Grid Infrastructure (EGI), during a workshop at the EGI Community Forum in April 2013. After a paper from 2012 was endorsed by the EGI Council, the time was right to turn the 'thought experiment' into a real-world test. Exploring potential brokering models, EGI have tested 'matchmaker' and 'one stop shop' models, in addition to the 'independent advisor' model currently used. Matchmaker sees the federator (EGI) allocate resources from resource providers that match customer requirements, with the resource provider paying the federator for establishing the contractual agreement and the customer paying the resource provider. One stop shop sees the federator doing everything including collecting payment from the customer, then paying the resource provider. For pay for use, both of these models work better, with lower overheads, than the independent advisor model currently used by EGI, where interactions are more decentralised (incurring more overheads); the federator listing services offered by resource providers and being funded by a membership. So pay for use may mean a shift in the role that EGI – and other infrastructure providers – adopt.

Problems and challenges ahead? There are some. Some academic research centres cannot, by law or contract, resell services to third parties. There is the question of taxation. Many member countries have VAT, which could have implications for the purpose of invoicing.

Identifying Success: Going Beyond Science

Horizon 2020 comes at a very interesting point in the history of Europe. Despite the current financial situation across Europe, public perception of science and wider academic research is very positive, and support for funding science remains high. There has never been a better time to engage the public in scientific discourse, and e-science projects are learning how to better interact with the public. The project iMarine, for example, has released an app called Applifish that can be downloaded onto a mobile device, which contains species and location data for many types of fish through an attractive and easy-to-use interface. This can help the average consumer make decisions about the sustainability of the fish they put on their plate. Global Excursion, the extended curriculum for science infrastructures, is aiming to introduce e-infrastructures – what they are and what they have achieved – to school students aged 14-18. Their virtual science hub has also spawned a successful spin-off project, Mash-Me.TV, which has now been commercialised

At the 10th e-Infrastructure Concertation meeting, Sonia Spasova, Communication Officer for DG Connect at the EC, highlighted the success of the e-ScienceTalk project in bringing the world of e-Science to a broader audience. Through the diverse communications channels that e-ScienceTalk employs including iSGTW.org, Gridcast.org, e-ScienceCity.org and GridGuide.org, the success stories of e-infrastructures in Europe and beyond are reaching a wider audience and hopefully inspiring young people to consider e-science as a career.

Going beyond science isn't just about communicating success stories. There is a need to formulate a translation pipeline to encourage instances of adoption of innovations and to recognise contributions outside of academia. Researchers may still require up-skilling and some hand-holding to help bring products to market, something that has been integrated into the 'rich toolbox' of instruments available within the EC funding framework.

Catherine Gater, e-ScienceTalk, *"e-ScienceTalk have pushed the message of e-science to a broader community. By delivering training programmes on communication, they've also been able to pass on what they've learned to other projects so that this incredible science can be made more public."*

Summary

It is clear now that 2020 is a goal-line. By shifting the emphasis to a date we can imagine, the EC is reframing the tasks ahead: to take on the Grand Challenges of the 21st Century with renewed focus. There is much change to come, but also an array of reassuring success stories that demonstrate the dynamism of e-infrastructures in the past. Taking on the Grand Challenges will require more coordination and strategic developments, but as synergies are identified within and between research communities, it is clear that research will continue to be more digital, more interlinked and more accessible than ever before.

For more information:

'Digital Science in Horizon 2020' (DG Connect 2013) http://ec.europa.eu/digital-agenda/en/news/digital-science-horizon-2020

Neelie Kroes blog - A budget for European growth: http://blogs.ec.europa.eu/neelie-kroes/eu-budget-innovation-cef/

Adding Pay-for-Use Models within EGI: https://www.egi.eu/blog/2012/12/03/adding_pay_for_use_models_within_egi.html

CRISP: www.crisp-fp7.eu

GÉANT: www.geant.net

eduroam: www.eduroam.org

EGI: www.egi.eu

Real Time Monitor: rtm.hep.ph.ic.ac.uk

iSGTW: www.isgtw.org

e-ScienceTalk: www.e-sciencetalk.org

email: info@e-sciencetalk.org

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733

e-ScienceTalk projects

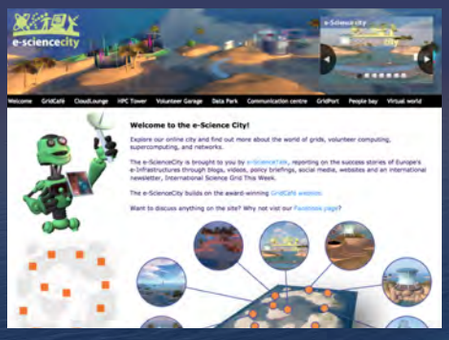
Along with the e-ScienceBriefings contained in this report, e-ScienceTalk runs a number of communications initiatives.

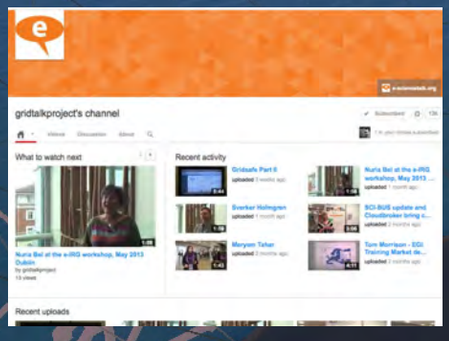
International Science Grid This Week (ISGTW) – our online weekly magazine – has expanded its scope from scientific research enabled by grids to include coverage of other forms of distributed computing, including supercomputing and cloud-based computing.

Blogging live from distributed computing and e-research events around the world, GridCast invites its audience 'behind the scenes' to share in the discussion and excitement of developing international e-science.

e-ScienceCity expands on the award-winning GridCafé – the place for everyone to learn about about power and potential grid computing – with some exciting new areas that take this learning tool into a whole new dimension. The 3D virtual world, based on OpenSIm technology, allows the user to explore e-ScienceCity with open source software.

There is a wealth of new content: Cloud Lounge explores the benefits and challenges of remote data storage for research; Volunteer Garage explains how you can donate your computer's spare CPU cycles to benefit researchers; HPC Tower focuses on supercomputers – what they are, and why (with grids and clouds) we still need them; Data Park takes the idea that scientific 'big, open' data is infrastructure – the foundation for better science. We've also integrated GridGuide – the sights and sites of distributed computing – into e-ScienceCity with the brand new addition of GridPort.





www. e-sciencetalk.org
www. e-sciencecity.org
www. isgtw.org
www. youtube.com/gridtalkproject
http://rtm.hep.ph.ic.ac.uk/
http://gridcast.web.cern.ch/Gridcast

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733