Workshop on

Big Data and Open Data

7th – 8th May, 2014

Royal Museums of Art and History, Brussels

BOOK OF ABSTRACTS

Association of European-level Research Infrastructure Facilities

Elettra Sincrotrone Trieste

WORKSHOP ON BIG DATA AND OPEN DATA

Scientific Committee | Organising Committee | Scope

Scientific Committee

Anke Kaysser-Pyzalla

Donatella Castelli

Fabio Pasian

Giorgio Rossi (Chair)

Jesus Marco de Lucas

John V. Wood

Juan Bicarregui

Kimmo Koski

Nicolas Menard

Norbert Meyer

Sanzio Bassini

Volker Guelzow

Organising Committee

Fabio Mazzolini

Giorgio Rossi

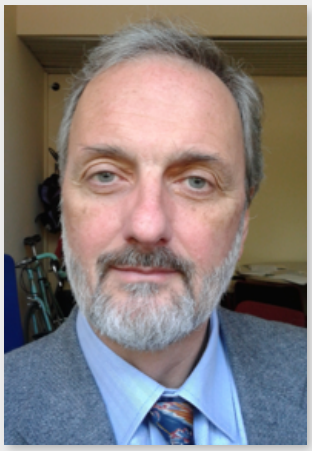
Ornela De Giacomo

Roberto Pugliese (Chair)

ERF-AISBL in collaboration with **Elettra - Sincrotrone Trieste**, is planning to address the ***big data*** and ***open data*** issues by organising a first workshop focused on the common problems that all data producing large scale research facilities are facing and will face in the years to come, and the ways to elaborate solutions. Purpose of the workshop is also to offer opportunities for discussions on specific proposals to be submitted to the Horizon 2020 program. The speakers are high profile experts in this field. We expect to attract over 100 highly concerned scientists and facility managers. All the sessions will be plenary.

PREFACE

Giorgio Rossi | Chair of the Scientific Committee



Giorgio Rossi is Professor of Physics at the Università degli Studi di Milano; he leads the APE group at IOM and Elettra performing research in surface and interface science and operating advanced beamlines and instrumentation open to users. He coordinates the Nano Foundries and Fine Analysis European infrastructure project since 2008. He is currently vice-president of ESFRI and Chair of the Physical Science and Engineering Strategy Work Group.

I welcome you warmly to the ERF Workshop "Big Data and Open Data".

The workshop focuses on a crucial aspect of all operational Research Infrastructures as well as of the new Projects: the data management and the open access to the data.

Data require specific infrastructures and specific treatment to become knowledge and value. The knowledge and value must be accessible in the most open and sustainable way. The data producing infrastructures and their users are engaged to extract knowledge and value from the data, and to manage their availability, beyond the traditional publication media, through interoperable repositories.

Storing knowledge and preserving value requires completing the data with the relevant metadata. This is certainly a challenging task. Metadata can be standard description of the detector sensitivity or of a trigger for serial observations under the same conditions, but can be extremely complex information if sample preparation steps, sample environment, simultaneous conditions of exposure, vacuum, probe beam energy and energy resolution, polarisation, instantaneous intensity and coherence must be recorded in order to make the data useful for extracting knowledge, in order to have value worth conserving and opening access to.

Beyond formats and transmission or storage protocols the critical step is in the metadata input. This requires definitions of the metadata set needed to give value to a data set: a variable with a large range. This imposes to design and implement instruments that will take care of collecting automatically the largest possible set of metadata while running an experiment. In many cases it is nevertheless needed to have an input by the researcher, and by the user, an efficient and adapted "logbook" vehicle to the stored information.

This last step is also part of the debate we must develop: the incentive to the researcher for doing this extra work that will add and maintain value in the stored data. A possible way worth discussing is the publication of data in "data journals" or "data repositories" that will be peer-evaluated as of validity and completeness of relevant metadata, with a corresponding "publication value" for the researcher who, perhaps only temporarily, renounces to analyse and publish a scientific paper, and rather publishes a well described data set, or both.

This ERF workshop addresses the multifaceted aspects of data management, and will certainly contribute to the maturing of a no-nonsense data policy that will increase effectively the value and the knowledge produced at the research infrastructures and its open availability.

Thanks to the organisers, chairs and speakers and to all participants of this potentially very productive workshop.

Giorgio Rossi

ERF-AISBL

Association of European-level Research Infrastructure Facilities

Members

ALBA
BNC BUDAPEST NEUTRON CENTRE
DESY
Elettra Sincrotrone Trieste
eli extreme light infrastructure
GSI
GANIL laboratoire commun CEA/DSM spiral2 CNRS/IN2P3
HZB Helmholtz Zentrum Berlin
Laserlab Europe
MAX IV LABORATORY
PRACE
diamond
PAUL SCHERRER INSTITUT PSI
SOLEIL SYNCHROTRON
FRM II Forschungs-Neutronenquelle Heinz Maier-Leibnitz

The **ERF-AISBL** (Association of European level Research Infrastructures Facilities) has the not-for-profit purpose to promote cooperation and joint project formulation between european-level research infrastructures (RIs) that are open, at international level, to external researchers. These Infrastructures include national infrastructures as well as european networks and consortia of research infrastructures.
To support and extend EU capabilities, ERF-AISBL wishes to engage not only with the communities represented in ERF-AISBL, but also more widely. This wider community includes governments, research funders, other RIs, higher education, industry and the citizen. To this purpose, ERF has organised a series of topical seminars and workshops on:

- Open Access (2009);
- Human Capital (2010);
- Energy (2011);
- Socio-economic impact (2012);
- Technology Transfer and Industrial Relations (2013)
- Energy -2nd workshop- (2013).

Visit our website: https://www.erf-aisbl.eu

Elettra Sincrotrone Trieste

ELETTRA SINCROTRONE TRIESTE

Elettra-sincrotrone trieste S.C.p.A

Mission

Its mission is to promote cultural, social and economic growth through:

- Basic and applied research;
- Technical and scientific training;
- Transfer of technology and know-how.

About

Elettra Sincrotrone Trieste S.C.p.A. is non profit shared company recognised of national interest; the partners are Area Science Park, the Autonomous Italian Region of Friuli Venezia Giulia, the National Research Council of Italy and Invitalia Partecipazioni S.p.A..

Elettra-Sincrotrone Trieste S.C.p.A. di interesse nazionale

Strada Statale 14 - km 163,5 in AREA Science Park

34149 Basovizza, Trieste ITALY

Tel. +39 040 37581

Fax. +39 040 9380902

www.elettra.eu

Elettra - Sincrotrone Trieste is a multidisciplinary international center of excellence, specialised in generating high quality synchrotron and free-electron laser light and applying it in materials science. The main assets of the research centre are two advanced light sources, the electron storage ring Elettra and the free-electron laser (FEL) FERMI, continuously (H24) operated supplying light of the selected "colour" and quality to more than 30 experimental stations on 32 beamlines. These facilities enable the international community of researchers from academy and industry to characterise material properties and functions with sensitivity down to molecular and atomic levels, to pattern and nanofabricate new structures and devices, and to develop new processes. Every year scientists and engineers from more than 50 different countries compete by submitting proposals to use the experimental stations at Elettra and FERMI.

WORKSHOP PROGRAMME

Global initiatives in this domain | Big Scientific Data in related domains

Wednesday, 07 May, 2014

Elettra Sincrotrone Trieste

Opening Session	***Chair*: Carlo Rizzuto**
13:00 - 14:00	Registration
14:00 - 14:30	Welcome **Giorgio Rossi, Carlo Rizzuto, Wolfgang Sandner**
14:30 - 15:00	Carlos Morais Pires **Data e-Infrastructures Horizon2020**
Global initiatives in this domain	***Chair*: Kimmo Koski**
15:00 - 15:20	Leif Laaksonen **Research Data Alliance**
15:20- 15:40	Peter Wittenburg, Damien Lecarpentier, Kimmo Koski **EUDAT: Shaping the Future of Europe's Collaborative Data Infrastructure**
15:40 - 16:00	Nobert Meyer, Maciej Brzezniak **e-IRG data challenges**
16:00 - 16:15	Coffee break
Big Scientific Data in related domains	***Chair*: Juan Bicarregui**
16:15 - 16:35	Francoise Genova **Big data in the astronomical community**
16:35 - 16:55	Alf Game **ELIXIR - Big data in Bioinformatics**
16:55 - 17:15	Jamie Shiers **Long Term data preservation for HEP (CERN)**
17:15 - 17:35	Massimo Cocco **Big Data in solid Earth sciences Observatories**
17:35 - 17:55	Brian Wee **Big, Complex Environmental and Biodiversity Data**
17:55 - 18:15	Stefano Cozzini **Data and Metadata for Nanoscience**
16:00 - 16:15	**Summary and discussion**
20:00 - 22:30	**Social Dinner**

WORKSHOP PROGRAMME

National and International initiatives in big data domain | Large Scale Facilities and ERF members initiatives in this domain

Thursday, 08 May, 2014

Elettra Sincrotrone Trieste

National and International initiatives in big data domain	***Chair*: Jesus Marco de Lucas**
08:30 - 08:50	Niinimäki Sami **Big and Open Scientific Data in Finland**
08:50 - 09:10	Sanzio Bassini **Italian Scientific Big Data Initiative**
09:10 - 09:30	Norbert Meyer, Maciej Brzeźniak, Maciej Stroiński **Polish National Data Storage**
09:30 - 09:50	Sergi Girona **PRACE Big Data challenge**
09:50 - 10:10	Tiziana Ferrari **EGI services for high throughput big data processing in a secure federated environment**
10:10 - 10:30	Juan Bicarregui **PaNdataODI: an Open Data Infrastructure for Photon and Neutron based research**
10:30 - 10:45	Coffee break
Large Scale Facilities and ERF members initiatives in this domain	***Chair*: Giorgio Rossi**
10:45 - 11:05	Roberto Pugliese **Elettra Big Data and Open Data Challenges**
11:05 - 11:25	Jean-Francois Perrin **ILL Open and Big Data challenges**
11:25 - 11:45	Thomas Stibor **Storing and Analyzing Efficiently Big Data at GSI/FAIR**
11:45 - 12:05	Volker Guelzow **Big Data management for large experiments at DESY**
12:05 - 12:25	Rudolf Dimper **BIG Data at ESRF, big problems, big opportunities**
12:25 - 12:45	Giovanni Lamanna **EU-T0, Data Research and Innovation Hub**
12:45 - 13:00	Bill Pulford **Diamond Big Data and Open Data Challenges**
13:00 - 13:15	**Summary and discussion**
14:00 - 16:00	**Facilities available for specific project proposal meetings**

ABSTRACTS

Global initiatives in this domain | Data e-Infrastructures Horizon2020

Carlos Morais Pires joined the European Commission in 1998 and is the Head of Sector in DG CONNECT for "Scientific Data e-Infrastructures" activities, as part of the European R&I Programme Horizon 2020. Before, he was lecturing computer networks and signal processing at the technical university in Lisbon. Carlos holds a PhD in Electrical Engineering from Instituto Superior Tecnico (1996) after his work on video coding in Telecom Italia Labs.

In 1949 Bertrand Russell wrote:

"Men of science […] could formerly work in isolation as writers still can; Cavendish and Faraday and Mendel depend hardly at all upon institutions and Darwin only in so far as the government enabled him to share the voyage of the Beagle. But this isolation is a thing of the past. Most research requires expensive apparatus […]. Without facilities provided by a government or a university, few men can achieve much in modern science."

This is still true today. At the dawn of the 21st Century the frontiers of knowledge are getting remarkably closer to understand the physics of the infinitesimally small and extremely large scales of the universe. Scientists are addressing extremely complex systems and interactions in macro and micro biological ecosystems. Funders have put considerable resources in very ambitious scientific projects for modelling the Human Brain which requires large amounts of data and processing with high performance computation. Large scale societal phenomena are being modelled to support better quality of life in urban areas and to transform our lives with the support of sophisticated technology.
The need for "expensive apparatus" is something that modern science intensified (need for more powerful telescopes, light sources, research boats, geological probes etc), and information is still the key element of scientific activity.
Horizon 2020 workprogramme contains a number of topics addressing data and computing e-infrastructures that if taken with appropriate resources and critical mass, can project Europe into the new world of data driven science exploiting to its full the potential of information and communication technologies.
The approach is to combine the expertise of scientific communities that know best their needs and the meaning of the data produced in their fields with the expertise of ICT communities capable of exploring the limits of high bandwidth communication, high-performance computing, open scientific software and virtual research environments.
The available instruments in Horizon 2020 will bring together different expertise ensuring the best balance of competences, critical mass and European dimension.

ABSTRACTS

Global initiatives in this domain | Research Data Alliance

Dr. Leif Laaksonen is a Director at CSC - IT Center for Science Ltd. in Espoo, Finland. He joined CSC in 1985 and is currently the Project Director for the RDA Europe project. Laaksonen contributed to the startup of FP7 projects PRACE, EGI_DS (European Grid Initiative Design Study) and was the Project Director for the e-IRGSP2 (e-Infrastructure Reflection Group Support Programme 2) project. He was the Chair of the e-Infrastructure Reflection Group (e-IRG) 2007 – 2010 and has supported the Finnish Ministry of Education and Culture in e-Infrastructure related matters. His training background is in computation chemistry and molecular graphics.

The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data. The Research Data Alliance aims to accelerate and facilitate research data sharing and exchange. The work of the Research Data Alliance is primarily supported and undertaken through its short term (up to 18 months) working and interest groups activities. Participation in the working groups and interest groups, starting new working groups, and attendance at the twice-yearly plenary meetings is open to all subscribing to the RDA principles. Coupled with this RDA boasts a broad, committed membership organisations dedicated to improving data exchange.

ABSTRACTS

Global initiatives in this domain | EUDAT: Shaping the Future of Europes Collaborative Data Infrastructure

Dr. Peter Wittenburg got his Diplom-Ingenieur Degree in Electrical Engineering in 1974 at the Technical University Berlin with computer science and digital signal processing as main topics.

In 2011 he became the head of the new unit called The Language Archive that was built as a collaboration between Max-Planck-Society, Berlin-Brandenburg-Academy of Sciences and the Royal Dutch Academy of Sciences and is Senior Advisor of TLA since 2012.From March 1st 2014 he joined the Max Planck Data and Compute Center in Garching/ Munich as Senior Adviser.

Among others, his later relevant activities include leading the technical infrastructure work in the CLARIN research infrastructure, from 2009 to 2012, being member of EC's High Level Expert Group on Scientific Data (Riding the Wave report) in 2010/11 and since 2011, acting as scientific coordinator of the EUDAT data infrastructure and the DASISH SSH cluster project as well as being member of the Steering Board of Research Data Alliance.

In 2011 he received the Heinz Billing Award of the Max-Planck-Society for the advancement of scientific computation and in 2013 was awarded Dr. H.C. from University Tübingen.

In recent years significant investment has been made by the European Commission and European member states to create a pan-European e-Infrastructure supporting multiple research communities. In the data area, efforts are being driven by the EUDAT project, a pan-European data initiative that started in October 2011. The project brings together a unique consortium of 26 partners – including research communities, national data and high performance computing (HPC) centers, technology providers, and funding agencies – from 13 countries. EUDAT is laying out the foundations of a European Collaborative Data Infrastructure (CDI) comprising a network of collaborating, cooperating centres across Europe, and combining the richness of community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centres. Cross-disciplinary and cross-national data infrastructure services for accessing and preserving research data are being designed within the project, addressing the needs of a broad range of users, from small and medium size scientific communities, to larger scientific organisations. Internal users of the CDI are those concerned with the management of community-specific data repositories containing large data collections. They can join their repositories formally with the CDI network, instantly benefitting from the persistence and resilience offered by the EUDAT partner network. Internal users are interested in archiving, replicating and cataloguing data on behalf of the research community they support. External users are those wishing to share data with colleagues or collaborators, or those wishing to discover and re-use data as part of their ongoing research. External users are anybody – researchers, citizen scientists, policy makers, members of the public – anyone wanting to share or re-use European research data, typically derived orphan data, in simple, powerful ways.

ABSTRACTS

Global initiatives in this domain | e-IRG data challenges

Norbert Meyer is the head of the Computing Data Services and Technologies at Poznań Supercomputing and Networking Center (Poland). His research interests concern resource management in distributed environments, accounting, data management, technology of development graphical user interfaces and network security, mainly in the aspects of connecting independent, geographically distant domains. He is the author and co-author of 60+ conference papers and articles in international journals, member of programme committees of international IT conferences. Norbert Meyer is the member of the e-IRG (e-Infrastructure Reflection Group), chair of the data management task force, co-author of several white papers, member of STRATOS group. He coordinated EU projects: DORII and RINGRID and several national projects.

The e-Infrastructure Reflection Group is formed by official delegations of ministries of science from various European countries. The e-IRG mission is to pave the way towards a general-purpose European e-Infrastructure. The vision for the future is an open e-Infrastructure enabling flexible cooperation and optimal use of all electronically available resources. The mission is to The e-IRG also coordinates activities with international initiatives outside of Europe.

Data services and its infrastructures are key parts of the European infrastructure. An analysis made by the group of experts showed the importance the context of data management, e.g. access and management of data infrastructures, reliability of services, metadata structures, unified access and interoperability of data structures, security.

The requirements collected from end users show that important is data curation - the access to infrastructure, which guarantees its stability over the next 20 - 30 years of archiving it in a predictable long-term business model. This applies to the scientific communities' experimental data, raw data and final results.

The data infrastructure is critical for applications, and its durability and reliability are vital for the quality of services provided on it.

Authors: Norbert Meyer, Maciej Brzeźniak

ABSTRACTS

Big Scientific Data in related domains | Big data in the astronomical community

Françoise Genova is the director of the Strasbourg astronomical data centre CDS, and one of the founding parents of the astronomical Virtual Observatory project. She has been the coordinator of several European projects dealing with the European Virtual Observatory, the current one being Collaborative and Sustainable Astronomical Data Infrastructure for Europe (CoSADIE, DG-CONNECT project 312559). She was a member of the High Level Expert Group on Scientific Data set up by the European Commission in 2010, and one author of the 'Riding the wave' report published in October 2010. She is a member of the Technical Advisory Board of the Research Data Alliance (RDA) and of the RDA/Europe project.

Astronomy relies on Big Data produced by space-borne and ground-based telescopes, and disciplinary data centres produce value-added services. Care is also taken of long tail data, in particular by Strasbourg astronomical data centre CDS in collaboration with the academic journals (data validated by a refereed publication).

The change of paradigm permitted by the on-line availability of data already took effect in this discipline, and scientists routinely use remote data from observatory archives and services in their daily research work. Key parameters have been open data and early international collaboration to define exchange standards, in particular a common data format, FITS, in the 70s.

On-line resources begun to be networked soon after the advent of the world wide web, and since 2000 the astronomical Virtual Observatory (VO) project develops a framework to allow seamless access to data and services. The European incarnation of the VO, Euro-VO, has been funded by the European Commission in a series of projects, the current one being Collaborative and Sustainable Astronomical Data Infrastructure for Europe (CoSADIE, DG-CONNECT project 312559). VO Standards are defined by an international body, the International Virtual Observatory Alliance (IVOA, http://ivoa.net), which gathers national VO projects from the five continents, Euro-VO and ESA-VO. Generic standards are used when possible, in particular for the Registry of Resources (OAI-PMH) and for Semantics (SKOS/RDF). The framework is inclusive, and anyone can register a resource or provide an access tool.

The open, widely used, global data infrastructure of astronomy can thus be considered as one of the research infrastructure of the discipline. The Virtual Observatory is the “glue” between the astronomical physical infrastructures. It is also a powerful mean to integrate the scientific community at large, by giving all researchers, wherever they are, the capacity to use the best data and tools, and to bring the research infrastructure in schools and universities.

ABSTRACTS

Big Scientific Data in related domains | ELIXIR - Big data in Bioinformatics

Dr. Alf Game

At present I have overall responsibility for BBSRC's strategy for development and support of research infrastructures. I am UK government representative on ESFRI (the European Strategy Forum for Research Infrastructures), and vice-chair of the board of the ELIXIR pan-European life sciences data infrastructure project.

In recent years I have led for BBSRC the conception, funding and delivery of programmes to establish national research capacity and capability in bioinformatics, genomics, stem cells, systems biology and bio-energy. I am also involved in various programmes to develop European- and international-scale activity in these fields, and to address the socio-economic, ethical and public perception issues surrounding them. My other recent scientific interests include plant and crop science, biodiversity and bioscience of ageing.

Much of this work involves assisting the transition of biology to "big science" and equipping the UK bioscience community with the facilities, skills and working practices to tackle problem-oriented large-scale, integrated and interdisciplinary science.

A substantial aspect of my role is the understanding and encouragement of change in the UK science base and I am interested in the interaction between funders, government, research institutions and individual scientists in this dynamic.

I retain a strong interest in European research funding and chair the UKRO Board.

European countries, companies and funding bodies invest heavily in biological research, seeking solutions to the many serious challenges facing society today.

Modern high-throughput biological research technologies such as DNA sequencing and various types of imaging have given individual investigators access to experimental technology that generates vast amounts of data. This needs to be analysed and stored – sometimes permanently. Moreover, there is need to ensure wide access to these data and to provide the means to use data from many different sources and of many different kinds together, for comparative study, modelling and simulation.

It is estimated that the annual rate of bioscience data generation will increase by 10^6 by 2020. The numbers of individual sources and users of these data is also potentially vast: the total number of hits to the European Bioinformatics Institute (EBI) website from commercial users alone in 2013 was 110 million. EBI represents only a small fraction of the bioscience databases within Europe.

ELIXIR will link these databases into a larger bioinformatics infrastructure, connecting them with one another and with tools that enable researchers to interpret the data they contain. ELIXIR is defining universal standards and best practices in order to present users with a single, transparent interface to a world of resources that are in fact widely distributed. For users, this represents a major improvement in the bioinformatics landscape.

The realisation of the vision for big biology and its economic and societal applications is dependent on e-science – storage and distribution, cloud computing and HPC.

ABSTRACTS

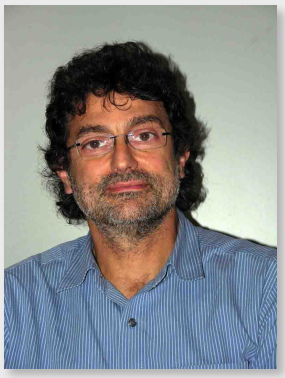
Big Scientific Data in related domains | Long-Term Data Preservation for HEP (CERN)

Dr. Jamie Shiers has worked at CERN since the time of the construction of the Large Electron-Positron (LEP) collider – the machine previously housed in the 27km tunnel currently used by the Large Hadron Collider (LHC). He is currently actively involved in the Research Data Alliance, where he serves as a member of the Technical Advisory Board, in the setting up and running of a service-provider independent e-Infrastructures User Forum, and as Project Manager for the Data Preservation in High Energy Physics collaboration (DPHEP). He has previously worked in many different areas in CERN's IT department, from operations, through software development and support, databases, data management, grid services and deployment, as well as collaboration with other disciplines (notably Astronomy and Astrophysics, Life Sciences and Earth Sciences, through EGI-InSPIRE and other EU-funded projects).

This talk will describe how the world's High Energy Physics laboratories, institutes and experiments are collaborating together to ensure the preservation of their data, as well as the knowledge required to re-use it. This is done based on standards and best practices that have been developed together with – or by – other disciplines and projects. This multi-disciplinary collaboration has advanced our implementation schedule by between 3-5 years, based on a "2020 vision", prepared just one year go. The talk will further explain the Business Model that has been developed to motivate long-term sustainability of the solutions proposed and how these solutions may benefit other communities. The targeted data volumes range from a few tens of TB to a few tens of PB for previous experiments and from hundreds of PB to several EB for on-going ones, i.e. those at CERN's Large Hadron Collider. Some detailed cost figures will be given, and areas were future projects are needed will be high-lighted. These will include the development and maintenance of complex software across disruptive changes – such as those that we are all facing today – as well as the all-important topics of Open Data and Big Data, the themes of this workshop.

ABSTRACTS

Big Scientific Data in related domains | Big Data in solid Earth sciences Observatories



Massimo Cocco is a Director of Research at the Istituto Nazionale di Geofisica e Vulcanologia, sezione Seismology and Tectonophysics, Rome. His research interests are focused on the physics of earthquakes and faults. More specifically, his work deals with earthquake dynamics and fault interaction, seismicity patterns and fault frictional properties. He is interested in both theoretical studies and observational researches. He has interests in all aspects of the mechanics of earthquake and faulting from observations of natural faults through geophysical and geological measurements to experimental faults at the laboratory scale. His expertise also covers the management of seismic networks and monitoring research infrastructures. He is presently coordinating the Preparatory Phase of a European Project named EPOS: European Plate Observing System. Its mission is the long-term integration of research infrastructures for solid Earth Science (www.epos-eu.org).

Progress in the understanding the physical processes controlling earthquakes, volcanic eruptions, unrest episodes and tsunamis as well as those driving tectonics and Earth surface dynamics requires a long-term plan to facilitate integrated use of data, models and facilities from existing, but also from new, distributed research infrastructures for solid Earth science. The European Plate Observing System (EPOS) represents such a plan. EPOS expresses a scientific vision and an IT approach in which innovative multidisciplinary research is made possible for tackling this challenge. One primary purpose of EPOS is to take full advantage of the new e-science opportunities coming available. The aim is to obtain an efficient and comprehensive multidisciplinary research platform for the Earth sciences in Europe.

The EPOS mission is to integrate the existing continental, national and local research infrastructures (RIs) in solid Earth science warranting increased accessibility and usability of multidisciplinary data from monitoring networks, laboratory experiments and computational simulations. This is expected to enhance worldwide interoperability in the Earth Sciences and establish a leading, integrated European infrastructure offering services to researchers and other stakeholders. EPOS is promoting open access to geophysical and geological data as well as modelling/processing tools, enabling a step change in multidisciplinary scientific research for Earth Sciences. The EPOS Preparatory Phase (funded by the European Commission within the Capacities program) is succeeding in leveraging the project to the level of maturity required to implement the EPOS construction phase, with a defined legal structure, detailed technical and financial plans.

In this presentation, we will describe the RIs to be integrated in EPOS and present the EPOS IT architecture in order to illustrate the integrated and thematic core services to be offered to the users. Some of the thematic services at community level already exist and are operational. The core services include not only user access to data, software, services, equipment and associated processing but also facilities for interaction and cooperative working between users, and storage of history and experience. The EPOS e-infrastructure is going to operate a full e-Science environment including metadata and persistent identifiers.

The presentation will also deal with the implications for the user community and funding agencies associated with the adoption of open data policies and access rules to facilities as well as the implications for the proper assessment of socio-economic impact of distributed, multidisciplinary RIs. We will also discuss the resources needed to tackle the challenge of fostering data driven research and big data applications. For Earth scientists, the prevalent problem is represented by the need of data, which must be promptly discovered, made accessible and downloadable, curated, minable and transferrable together with appropriate processing software and e-infrastructure resources. In general, there are a number of overlapping issues that regard data organisation and their access, data transfer from (and to) super computing centres (HPC) and among the platforms of the federated communities. We will also present examples of combined data analyses and we will address the importance of opening our research infrastructures to users from different communities and stakeholders - not limited to solid Earth science - using open data techniques.

Finally, the presentation will also discuss the international cooperation initiatives and the global perspectives for solid Earth data infrastructures.

ABSTRACTS

Big Scientific Data in related domains | Big, Complex Environmental and Biodiversity Data

Dr. Brian Wee
As NEON, Inc.'s Chief of External Affairs, Brian is the organization's liaison to Congress, US Federal agencies, and other scientific organizations. He also represents the informatics needs of the large-scale environmental sciences before the computer science and Federal data community. Brian joined the NEON Project Office at the American Institute of Biological Sciences in 2004 as a post-doctoral associate, then became a staff scientist before transitioning to the role of Administrative Director. Previously he worked for Andersen Consulting (now Accenture) designing and implementing IT solutions and then served as Senior Instructional Designer leading instructional design, knowledge management, business-process redesign, and web development projects.
Brian holds a Ph.D. in Ecology, Evolution, and Behavior from the University of Texas at Austin, a M.Sc. degree in Computer Science – Artificial Intelligence at Northwestern University, Evanston, IL and a B.Sc. in Information Systems and Computer Science from the National University of Singapore. His M.Sc. studies focused on designing and implementing computer augmented learning solutions for high-school classrooms and corporate training at the Institute for the Learning Sciences. His Ph.D. focused on investigating the relative effects of behavioral, physiological and landscape barriers on the genetic structure of insect populations by integrating genetic, behavioral, and GIS analyses.

The National Ecological Observatory Network (NEON) is a US investment designed to listen to the pulse of the continental ecosystem and how it is responding to large-scale environmental changes. Once fully commissioned in 2017, NEON (the Observatory) will collect data from 106 sites across the United States (including Alaska, Hawaii and Puerto Rico). These sites were strategically selected to represent 20 eco-climatic domains, which include distinct landforms, vegetation, climate and ecosystem processes. Each site includes a variety of sensors placed in the soil, water, and on a tower. Information is also collected on plants, animals, invertebrates, and microorganisms around the site. An airborne remote sensing platform flies over sites annually collecting aerial data. Over 700 Observatory data products will be freely web accessible to enable regional to continental-scale ecological forecasts, an essential capability if we are to better manage critical ecosystem services that humans rely on. NEON's open access approach to its data and information products will enable scientists, educators, planners, and decision makers to map, understand, and predict primary effects of humans on the natural world and effectively address critical ecological questions and issues.
NEON is also developing formal partnerships with environmental observatories outside the United States to foster greater interoperability between observatories. To date, the Observatory has executed agreements with European Union and Australian partners that focus on collaborations to harmonize cyberinfrastructure approaches, environmental measurement techniques, education best practices, and other areas. NEON is funded solely by the US National Science Foundation. The Observatory is currently under construction, and slated for completion in 2017.

ABSTRACTS

Big Scientific Data in related domains | Data and Metadata for Nanoscience

Dr Stefano Cozzini has over 15 years' experience in Information Technology, mainly in the area of scientific computing and HPC computational e-infrastructures.
He has been involved in many scientific e-infrastructure projects and is currently now coordinating a team which is setting up a data repository for nanoscience within the NFFA project.
As from this year, he is Coordinator of the Master in High Performance Computing (www.mhpc.it) promoted by Sissa and ICTP.
He served as Scientific Consultant for the International Organisation Unesco–ICTP, from 2003 to 2012, and UNDP/UNOPS during 2011 and 2012.
At the end of 2011, he cofounded a spin-off company of CNR/IOM institute eXact lab srl. The company provides advanced computation services by means of HPC and CLOUD infrastructure.

In this short talk I will describe the current status of the Data Repository for nanoscience.
Such a repository system aims to be able to manage scientific data produced by experimental large scale facilities which are part of the Nanoscience Foundries and Fine Analysis (NFFA) initiative.
The goals of NFFA Data Repository are to provide to its users not only high performance data access,
but also advanced search capabilities between data from different sources with an extended metadata structure and data sharing to promote collaboration and interoperability.
The core of the DR architecture will based oon the KIT Data Manager of Karlsruhe Institute of Technology, a multi-layered service architecture integrating software and technologies to build up repository systems able to manage Big Data.
In this initial stage of the Data Repository scientific data will mainly acquired from the following NFFA facilities: a scanning electron microscope (SEM) equipped with a detector for Energy Dispersive Spectroscopy (EDS), the APE beamline
instrumentations and finally from the NFFA Theoretical facility mainly based on the open source Quantum-Espresso Package.
For all this sources, an injection plug-in to the repository is under implementation allowing an automatic harvest of metadata of the data produced by NFFA users.
Moreover, one of the most challenging purpose of our project is to realize a semantic search engine that will allow users to search scientific related data among different resources. In this respect it is of great importante a clear identification and organisation of metadata associated with scientific data coming from different experimental and theoretical sources.

ABSTRACTS

National and International initiatives in big data domain | Big and Open Scientific Data in Finland

Sami Niinimäki works as a senior advisor at the Finnish Ministry of Education and Culture and more specifically the Department for Higher Education and Science Policy. His duties revolve around open science and e-infrastructure of higher education institutes. Niinimäki has a background in environmental and forest economics. Prior to joining the ministry he worked at the Finnish Forest Research Institute and the University of Helsinki. His research combined a highly complex ecological tree growth model with economics and optimisation.

The Finnish Government Programme from 2011 promises that information resources produced using public funding will be opened up for public and corporate access. The goal is to make digital materials, big and small, produced by the public sector available for everyone in a digitally reusable format via information networks. Within the Finnish Government, the Ministry of Finance is in charge of the Open Knowledge Program 2013-2015. The most important goals of this program are to develop best practices and information architecture and to create a common metadata model and a public data portal. The Ministry of Education and Culture (MinEdu) is among other things responsible for developing science policy and for advancing interoperable scientific information infrastructure. In a recent initiative MinEdu set out to incorporate open science and research to the whole research process by setting guidelines for opening not only publications and research data but also methods and tools. Also important are increasing skills and knowledge and support services in open science domain. The benefits of openness include better and faster access, cost-efficiency, equality, and quality improvements. Goodwill also requires supporting infrastructure services. MinEdu offers services to Finnish universities and polytechnics mostly through a contract with CSC – IT Center for Science Ltd. which is a non-profit company owned by the Finnish state. The hottest topic at the moment is the data-intensive scientific discovery, or the so called 4th paradigm of science. To address this MinEdu has invested through CSC in state-of-the-art solutions like fast networks, tiered computing, scientific applications, information management, data services as well as expert support. These services benefit not only national science and research but, through various collaborations, the whole international community at large.

Author: Sami Niinimäki[(1)] and Pirjo-Leena Forsström[(2)]

[(1)] Ministry of Education and Culture, P.O. Box 29, FI - 00023 GOVERNMENT.
[(2)] CSC – IT Center for Science Ltd., P.O. Box 405, FI-02101 Espoo, Finland.

ABSTRACTS

National and International initiatives in big data domain | Italian Scientific Big Data Initiative

Sanzio Bassini joined CINECA in 1979; in 1981 he was nominated responsible for the scientific computing systems installed at CINECA, and in 1984 he joined the Italian Supercomputer Project that introduced the first supercomputer of this class in Italy. In 1996 he was appointed High Performance Systems Division Manager. In 2006 he was appointed Director of the newly formed System & Technology Department that is in charge of the whole CINECA data-centre infrastructure and services management and starting from 2009 he is Director of the SuperComputing Applications and Innovation Department.

We will give an overview of the initiatives at National and International level about the data management and life cycle management of the big data for the research.The presentation will address curation, replica, long term archiving and data processing as well of enabling service for the scientific research. Some attention will be posed also for technology transfer action towards industries and added value service to promote the innovation process in the private sectors. Some considerations will be presented for some vertical domain including environment, NGS and bioinformatics, and neuro informatics in the context of human brain project.

ABSTRACTS

National and International initiatives in big data domain | Polish National Data Storage

Norbert Meyer is the head of the Computing Data Services and Technologies at Poznań Supercomputing and Networking Center (Poland). His research interests concern resource management in distributed environments, accounting, data management, technology of development graphical user interfaces and network security, mainly in the aspects of connecting independent, geographically distant domains. He is the author and co-author of 60+ conference papers and articles in international journals, member of programme committees of international IT conferences. Norbert Meyer is the member of the e-IRG (e-Infrastructure Reflection Group), chair of the data management task force, co-author of several white papers, member of STRATOS group. He coordinated EU projects: DORII and RINGRID and several national projects.

The National Data Storage (NDS) is a distributed national data infrastructure delivered via services in the Polish academic fibre optic network PIONIER to the Polish scientific community. The basic service is remote backup and archivisation of data with on-line access at anytime. Additional applications which make the functionality more attractive are as follow: appliance, CryptoBox, CryptoDroid and CryptoFS. Security, Privacy and sustainability are the main features which characterise the NDS. The availability allows to send and download files from your computer, tablet or smartphone wherever you are, at home, at work, at school, on vacation. The services allows to use the infrastructure by institutional organisations but also by users which may want to share the data within a group or simply make a backup of data.

Authors: Norbert Meyer, Maciej Brzeźniak, Maciej Stroiński

ABSTRACTS

National and International initiatives in big data domain | PRACE Big Data challenge

Sergi Girona is Chair of the Board of Directors of PRACE, as well as Director of the Operations Department of the Barcelona Supercomputing Center (BSC). He belongs to the BoD of PRACE since its creation in 2010, and currently is both its Chair and Managing Director.

He holds a PhD in Computer Science from the Technical University of Catalunya. In 2001, EASi Engineering was founded and Sergi became the Director of the company for Spain, and the R&D Director for the German headquarters.

In 2004, he joined BSC for the installation of MareNostrum in Barcelona. MareNostrum was the largest supercomputer in Europe at that time, and it maintained this position for 3 years. Sergi was responsible for the site preparation and the coordination with IBM for the system installation. Currently, he is managing the Operations group with the responsibilities for User Support and System Administration of the different HPC systems at BSC.

The success of PRACE in providing unprecedented allocations of resources for computational science projects of the highest scientific quality has identified challenges in the storage, analysis, visualisation and transfer of data connected to large-scale simulations.

In order to provide the best possible services to the scientific community and ensure the largest possible impact of these large-scale simulations, the Scientific Steering Committee (SSC) and the Board of Directors (BoD) of PRACE decided to ask the users in November 2013 for Expressions of Interest for data services connected to large-scale computational projects.

The EoI should include the answers on following questions: What is the scientific justification and why must the handling of data be closely integrated with the PRACE computing services? What is the expected amount of data and the expected computing resources needed for the analysis? For how long must the data be accessible? Should the data be open or not to the wider scientific community and how should the scientific access be organised?

The presentation at the Big Data and Open Data Workshop will give an overview on the PRACE's strategy on Big Data, in order to better understand the level of demand, the scientific justification for such data services and the best way to structure these data services on Tier-0 machines.

ABSTRACTS

National and International initiatives in big data domain | EGI services for high throughput big data processing in a secure federated environment

Dr. Tiziana Ferrari is Technical Director at EGI.eu and she is responsible for the direction of the EC project EGI-InSPIRE, which is supporting the operations and evolution of the European Grid Initiative. She was formerly Chief Operations Officer of EGI, and she participated in the definition of the European Grid Initiative governance and services. She has been active member of the Open Grid Forum and Internet Engineering Task Force, contributing to the testing and design of various generations of the pan-European research and education network that interconnects Europe's National Research and Education Networks. Tiziana holds a PhD in Electronics and Data Communications Engineering from the Universita' degli Studi in Bologna.

The European Grid Infrastructure with 365 data centres offering more than 400,000 cpu cores and 190 PB of disk space, is one of the largest distributed e-Infrastructure facilities in the world providing services for data management and high throughput computing. In order to better address the emerging new challenges of the European research data factories, this infrastructure is evolving to include new types of platforms that will provide researchers the needed flexibility to run customised and tailored services for their discipline.
The talk will describe how the federated, standards-based IaaS and PaaS Cloud platform will deliver reliable and advanced services for scientific computing and e-Research across Europe and worldwide. Use cases and the needed capabilities for big data analytics in a distributed environment will be described.

ABSTRACTS

National and International initiatives in big data domain | PaNdataODI: an Open Data Infrastructure for Photon and Neutron based research

Juan Bicarregui is Head of the Data Division in the Scientific Computing Department at STFC. The division has responsibility for research and development of the data systems which handle much of the huge volume of scientific data which is produced by the STFC research facilities. Juan leads the PaNdata collaboration which is developing a shared computing infrastructure across all 13 major European Photon and Neutron laboratories. He plays a key role in formulating STFC and RCUK policy on open data and is the UK representative in a G8+O6 working group on data. He is chair of the RDA Organisational Advisory Board and also chair of the Alliance Permanent Access and a Director of the Digital Preservation Coalition. Juan has over 100 publications which can be found through Google Scholar or through the STFC Institutional Repository.

The PaNdata collaboration brings together thirteen large multidisciplinary Research Infrastructures which operate hundreds of instruments used by over 30,000 researchers each year. Scientifically, neutron and photon laboratories are complementary facilities which are applied in a wide spectrum of research disciplines. They support fields as varied as physics, chemistry, biology, material sciences, energy technology, environmental science, medical technology and cultural heritage. Applications are numerous, for example, crystallography can reveal the structures of viruses and proteins important for the development of new drugs; neutron scattering can identify stresses within engineering components such as turbine blades, and tomography can image microscopic details of the 3D-structure of the brain.

PaNdata-ODI is developing an Open Data Infrastructure across the participating facilities with user and data services which support the tracing of provenance of data, preservation, and scalability through parallel access. Historically, the situation at many of the facilities, and in particular at the photon sources, has left data management largely up to the individual users who often literally carried data away on portable media. Not only is this becoming unfeasible due the dramatic increase in size of some of the data sets, it is also counterproductive for the scientific workflow, verifiability of the data analysis and ultimately constitutes a dramatic loss for the scientific community. Our vision is to standardise and integrate our research infrastructures in order to establish a common and traceable pipeline for the scientific process from scientists, through facilities to publications. At the heart of the vision is a series of federated catalogues which allow scientists to perform cross-facility, cross-discipline interaction with experimental and derived data, with near real-time access to the data. This will also deliver a common data management experience for scientists using the participating infrastructures particularly fostering the multi-disciplinary exploitation of the complementary experiments provided by neutron and photon sources.

The two main services being deployed in PaNdataODI are catalogues of users and data. For users, a major component of the project is an authentication system that is normalised to include scientific users across the collaborating facilities and able to interoperate with similar systems across the ERA. For data, the deployed data catalogues will capture details of data files generated by facility instruments during experiments, during analysis, and through to publications.

Research is being undertaken into provenance and preservation which will enable the tracking of data along the complex path "from proposal to publication", and into scalability as volume of data is becoming increasingly challenging with frame rates in the kHz to MHz range, resulting in data rates of Terabytes per day.

ABSTRACTS

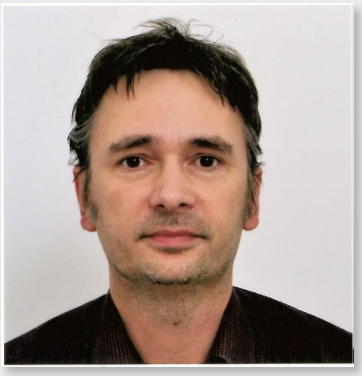
Large scale Facilities and ERF members initiatives in this domain | Elettra Big Data and Open Data Challenges

Prof. Roberto Pugliese is a computer scientist with an MBA. He is the coordinator of the Technology Platform of Elettra Sincrotrone Trieste.
He has been also Professor E-Commerce at the University of Udine. At the moment he is teaching project management a the SISSA - International School for Advanced Studies (www.sissa.it) and at the MIB School of Management (www.mib.edu) of Trieste.
He has published more than 20 articles in scientific journals and is co-author of 4 books about Instrumentation, Grids and Clouds and ICT research Infrastructures edited by Springer.
His research interests include e-Science, Human Computer Interaction, Mobile Robotics, Project Management, Management Control, Strategic Control, Personal Development, High Performance and High Throughput Computing and Big Data and Open Scientific Data.
He is member of the International Scientific Advisory Committee of International Conference NOBUGS since 1998 and since December 2003 and since 2006 Co-chair and member of the Program Committee of the INGRID International Conference. Since 2002 he has been involved with leading roles in regional, national and international ICT projects (BIOXHIT, IA-SFS, EUROTeV, GRIDCC, EGEE-II, DORII, MADBAG, I-SOI, PANDATA-EU, PanDataODI).
Since November 2011,he is member of the Project Management Institute, Northern Italian Chapter, (http://www.pmi.org) and since June 2013, certified Project Management Professional by PMI. Since May 2012 he is a MIUR Industrial Research Expert. He has recently co-funded two high-tech startups.

Elettra is notably the core partner of a distributed research facility (CERIC-ERIC). The members countries include Austria, Croatia, Czech Republic, Hungary, Italy, Poland, Romania, Serbia, and Slovenia. This initiative aims at providing researchers with access to synchrotron light and other microscopy probes. The access should promote analytical and modification techniques for fields such as materials preparation and characterisation, structural investigations and imaging in Life Sciences, Nanoscience and Nanotechnology, Cultural Heritage, and Environmental and Materials Sciences. This collaboration renders Elettra a research accelerator hub as it can provide a common platform and infrastructure for the all the other partners of the CERIC-ERIC. This paper describes how Elettra is preparing to for this task. The Elettra Scientific Data policy will be presented as well as a set of use cases that describe specific big data issues that modern beamlines have to face. In order to capitalise the national investment and maximise the utilisation of resources, the Elettra data storage infrastructure will be integrated with the off-site Italian supercomputing center CINECA. The architecture of this interconnection and integration is outlined in this paper. Additionally it is presented a review of important aspects of Big and Open scientific data such their financial costs. Finally we present the features that are offered today by the Elettra Virtual Laboratory and the future plans to move towards a sustainable solution for big and open data challenges in Elettra.
Cluster file systems are the heart and crucial component for storing and accessing efficiently massive amount of data.
Deploying cluster file systems on commodity hardware introduces challenges such as robustness and scalability. GSI/FAIR is running one of the largest Lustre file system deployment in Europe with high-speed connections to partnered research institutes and universities for analysing data from the accelerator for heavy ions.
Monitoring such a large cluster file system for finding IO bottlenecks and even making predictions on upcoming issues given the past history is a very challenging task which will be addressed in this talk. In addition, I will provide a brief overview of how massive computational power provided by GPU's in combination with large data sets allow us to tackle problems which seems to be infeasible for being tackled in the past.

ABSTRACTS

Large scale Facilities and ERF members initiatives in this domain | ILL Open and Big Data Challenges



Jean-François Perrin is the Head of the IT Service at the Institut Laue - Langevin, he is responsible for the maintenance and improvement of the general aspect of informatics and telecommunication at ILL, and has previously been involved in a number of different EU funded projects (PANData-Europe, ESRF-UP), he is also currently leading work packages on data management in the PANData-ODI and CRISP projects. Jean François holds a MSc. in Fundamental Physics.

The Institut Laue Langevin (ILL) is an international research centre at the leading edge of neutron science and technology. It operates the most intense neutron source in the world, providing intense beams of neutrons to a suite of 40 high-performance instruments.

Over 750 experiments are completed every year, in fields including magnetism, superconductivity, materials engineering, and the study of liquids, colloids and biological substances.

An ambitious modernisation programme (2001-2014) was launched in 2000, through the design of new neutron infrastructure and the introduction of new instruments and instrument upgrades. The first phase has already resulted in 17-fold gains in performance. The second phase has started in 2008, it comprises the building of 5 new instruments, the upgrade of 4 others, and the installation of 3 new neutron guides. New instrument (ThALES) and the major upgrade of another four instruments (SuperADAM, D16, D22, IN15) will be rolled out in June.

The IT department not only provides ICT support and solutions but also network, archival and curation for the dataset acquired from scientific experiments carried out since 1973, and also provides analysis infrastructure. Since the publication of the ILL data policy, in 2011, data is publically available to the scientific communities following a 3 year period where the dataset is exclusively available to the initial experimenters.

Until recently, visiting scientist were able to easily transfer their data to their home laboratory for further analysis (using hard drives or standard network transfer protocol). Other scientists used locally available infrastructure for processing their data.

Nowadays with the recent growth of the volume of experimental data generated at ILL, where some experiment are generating more than 35 TB of raw data, transporting data to most users home facility is no longer feasible. Providing a modern solution not only for the storage and archiving but also for the data analysis has become a paramount objective for the IT department.

We need to improve our analysis facility by providing more; capacity, flexibility, greater performance, and user friendliness. Hopefully cloud technologies are now mature enough to help us to achieve our goal.

ABSTRACTS

Large scale Facilities and ERF members initiatives in this domain | Storing and Analyzing Efficiently Big Data at GSI/FAIR

Thomas Stibor received his PhD degree in computer science from the Technical University Darmstadt, Germany.
He spent during his PhD time a half year as a visiting researcher at University of Kent, England, where he worked on artificial immune systems
and machine learning. After that, he was a post-doc at University of California at Davis, USA and later a lecturer at Technical University Munich, Germany where
he worked in the field of machine learning. Currently, he is with GSI Helmholtz Centre for Heavy Ion Research, Germany
as a researcher and software developer for HPC file systems.

Cluster file systems are the heart and crucial component for storing and accessing efficiently massive amount of data. Deploying cluster file systems on commodity hardware introduces challenges such as robustness and scalability.
GSI/FAIR is running one of the largest Lustre file system deployment in Europe with high-speed connections to partnered research institutes and universities for analysing data from the accelerator for heavy ions. Monitoring such a large cluster file system for finding IO bottlenecks and even making predictions on upcoming issues given the past history is a very challenging task which will be addressed in this talk. In addition, I will provide a brief overview of how massive computational power provided by GPU's in combination with large data sets allow us to tackle problems which seems to be infeasible for being tackled in the past.

ABSTRACTS

Large scale Facilities and ERF members initiatives in this domain | Big Data management for large experiments at DESY

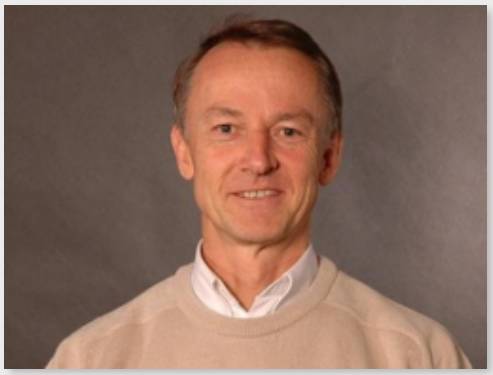
Volker Guelzow studied Mathematics and Physics at the university of Goettingen from where he received his PhD in 1987. At that time he was as scientist at the german aerospace establishment DLR. From 1988 until 2001 he became head of the software department at the Climate Computer Centre (DKRZ) in Hamburg. From there, he became director of the computer centre of the University of Kiel. Since more than 10 years he is the Head of Computing at DESY. His special interests are Big Data, Grid&Cloud and HPC computing and networking.

DESY, a large German national laboratory, is supporting a large variety of demanding scientific communities. The most outstanding ones are certainly particle physics, supported through their large T2 centres, and the different photon science experiments, e.g. at PETRA III & Flash. The enormous amount of data produced, requires an extremely reliable and high performance computing and data infrastructure. With the European XFEL becoming operational soon, even a new class of data management challenges need to be solved. Although we benefit from certain similarities between the communities, we observe severe and significant differences on which this presentation will be focused. Based upon the dCache solution, developed under the lead of DESY in an international cooperation, DESY's Big Data concepts for tomorrow will be described.

Authors: Patrick Fuhrmann, Volker Guelzow

ABSTRACTS

Large scale Facilities and ERF members initiatives in this domain | BIG Data at ESRF, big problems, big opportunities



Rudolf Dimper earned his degree in chemical engineering in 1981 in Hamburg. His diploma work was done in the European Molecular Biology Laboratory (EMBL) on the design of a real-time data acquisition system for X-ray muscle diffraction experiments. After his studies he held a position at the Institute Laue-Langevin (ILL), Grenoble, to design assembler programs for real-time data display of neutron detector data, followed by a position at the Institute for Millimetre Radio Astronomy (IRAM). Here his work focused on software design for a radio wave correlator and in accompanying the construction of the Plateau de Bure Radio Interferometer (2550m altitude) as the Station Manager. Since 1987 he works at the European Synchrotron Radiation Facility where he initially held various positions in the Computing Services Division. In 2004 he was appointed Head of this Division and member of the ESRF management team. In 2010, following an internal reorganisation of the laboratory, he was nominated Head of the Technical Infrastructure Division, bringing together services encompassing the computing infrastructure, management information systems, building construction and maintenance, electrotechnics, the vacuum systems as well as geodesy and alignment.

As a service institute, the ESRF invites more than 5000 scientists per year to carry out peer reviewed experiments in addition to an internal world-class scientific programme. The ESRF is a typical example in the scientific landscape where science and computing are intimately coupled and interdependent. New experiments are becoming possible because of an enabling high-performance computing environment; other experiments push the computing infrastructure to the limits and encourage us to explore new ways for staying abreast of an unprecedented data avalanche.
With the help of one or two scientific examples I will highlight the challenges of data intensive experiments and their underlying IT issues. IT is a key enabling technology allowing to control sophisticated instrumentation, read data from high resolution imaging detectors, and process data on CPU/GPU clusters. ESRF is currently building new experimental stations which will allow exploring samples in the nano-scale with extraordinary precision and detail and will require even more cutting edge IT services. It becomes obvious that being able to analyse very large data sets is increasingly difficult because scientists are not IT specialists and the IT environment is not always optimised or well adapted. Analytical facilities must provide additional assistance for data management and data analysis of very large data sets for increased scientific output. I will conclude my presentation by an overview of current efforts to implement a European wide initiative to pool resources in an effort to develop a data analysis eco-system.

ABSTRACTS

Large scale Facilities and ERF members initiatives in this domain | EU-T0, Data Research and Innovation Hub

Giovanni Lamanna is a physicist, Directeur de Recherche at CNRS (Centre national de la recherche scientifique), leading the high energy Astroparticle Physics research team at the LAPP (Particle Physics Laboratory in Annecy-le-Vieux) CNRS laboratory in France.
Some of his areas of research are fundamental and applied physics, experimental research instrumentation and data processing. He has published over 200 research papers.
He coordinates the Data Management project of the Cherenkov Telescope Array (CTA) international consortium. He has been cooperating during the last years with APPEC (the Astroparticle Physics European Consortium) on issues concerning optimal and efficient large-scale data analysis and management in Astroparticle. In 2013 Giovanni Lamanna became computing, e-science and e-infrastructure policy manager at IN2P3 (national institute of nuclear and particle physics). IN2P3 is part of CNRS and through its reference large computing centre (CC-IN2P3) and some other national distributed centres, provides e-infrastructures and know-how in data archive and processing to international research collaboration and scientists worldwide.

Some research institutes and funding agencies engage in the EU-T0 initiative with the purpose of making the most efficient use of their computing infrastructures and all the accumulated experience, by encouraging and extending cooperation in support of research in multiple fields. The EU-T0 is a Data Research and Innovation Hub: a European Tier 0 data-management and computing center, implementing a point of connection and coordination among major national e-infrastructures. The EU-T0 partners aims for: development of modern data management services and solutions, deployment and operation of the federated computing infrastructure and interoperable services to support research workflows. This is considered of particular interest in the fields of Particle, Nuclear, Astro-Particle Physics, Cosmology and Astrophysics, and then extended to other disciplines, the research projects of which require the handling of very large volumes of very complex data, present critical-challenges in computing and data processing models/services and in the long-term data preservation.
The communication will focus on the big data and open access issues concerning in particular current and future Astroparticle projects and the implications for the EU-T0 collaboration.

ABSTRACTS

Large scale Facilities and ERF members initiatives in this domain | Diamond Big Data and Open Data Challenges

Dr. **Bill Pulford** was head of the Data Acquisition and Scientific Computing group and is now Science I.T. coordinator at the Diamond Light Source.

Bill has performed similar roles first at the ISIS neutron facility and later at the European Synchrotron Radiation Facility. He has very extensive experience at most aspects of data acquisition with both X-Rays and Neutrons and was one of the earliest instigators of data management at ISIS. In addition he is a prime mover for a common authentication system for scientific users across Europe.

Diamond Light Source is the UK's national x-ray synchrotron science facility, located at the Harwell Science and Innovation Campus in Oxfordshire. The facility has currently 24 scheduled beamlines rising to 32 in 2017; these are used by over 3,500 academic and industrial researchers to study a wide range of disciplines including structural biology, energy, engineering, nanoscience and environmental sciences. Since the light source was opened to users in 2007, since then there have been over 10,000 experimental visits leading to the acquisition of more than 300,000,000 raw data files with a volume in excess of 1.3 Petabytes; with very few exceptions all of these files are catalogued and are available to our users for continuing analysis. We now face the ever increasing challenge to provide this service where possible for data volumes rising to the 10 Petabytes projected in 2017 and then beyond.

This talk will provide an overview with focus on some potentially interesting details of the experiment flow at Diamond from experimental proposal submission, review, scheduling, data acquisition, evaluation and analysis through to data storage. Discussions will include metadata and file format and storage issues together with authentication and authorization of the individual experimental visits.