Fostering Europe's eScience environment

Driven by the needs of earthquake and seismology data-intensive applications, a pan-European effort is building a platform for data mining and modelling. **Professor Jean-Pierre Vilotte** discusses their work

Why is the Virtual Earthquake and seismology Research Community in Europe eScience environment (VERCE) so important for improving understanding of earthquakes and seismology?

The nature of earthquake and seismology science is changing – new discoveries and new monitoring methods will emerge from statistical analysis of large amounts of complex data continuously transmitted by increasingly dense global and regional seismology and high-rate GPS networks. Today, our ability to acquire datasets outpaces our ability to manage, explore and analyse them. Data-intensive research challenges conventional methodology and requires a new and holistic approach. VERCE will develop a user-centric, data-intensive eScience environment specifically designed to enable data analysis and modelling tasks that are simply not possible today.

VERCE's objective is to provide a Datascope instrument of unprecedented capabilities, built upon a service-orientated architecture, and providing a set of work and data flow tools and services that allow efficient use of the European computer and data infrastructures together with the distributed seismology archives infrastructure.

Could you elaborate on some of the tools that VERCE has established to aid new emerging data-intensive applications of this community?

Driven by the needs of data-intensive applications in data analysis and modelling, VERCE's strategy aims to provide a comprehensive service-orientated architecture and framework adapted to the scale and diversity of those innovative applications and integrating the community data infrastructure with Grid, High-Performance Computing (HPC) and Cloud infrastructures.

Currently, VERCE is enabling a set of pilot data analysis and modelling applications within the Dispel data-intensive workflow environment, relying in particular on OGSA-DAI and other Grid services. The Dispel language is chosen for three main reasons: firstly, it is dataflow based for multi-site enactment; secondly, it has functions to describe work patterns; and finally, it is designed for human communication and to avoid details that inhibit automated mapping and optimisation.

Can you offer a brief summary of the concept of grid integration?

Grid integration, especially for data movement, is very important for VERCE. Greater support for GridFTP for reliable file movement has been integrated into OGSA-DAI in accordance with the specifications of the Globus project. Collaboration with the EUDAT project investigates the use of data management systems such as iRODS.

In order to promote uptake of any platform, ObsPy integration allows researchers to easily continue using the languages, tools and libraries with which they already work. Finally, a close collaboration with the ERC project WHISPER (see p17) focuses on enabling its data-intensive analysis library within the Dispel environment allowing enactment on Grid and HPC infrastructures.

How are you testing early prototypes?

A data-intensive seismic cross-correlation and an HPC waveform modelling use cases were put together in order to assess the prototype platform and identify early requirements in terms of data handling, data movement, and distributed process execution. A test case was trialled on the Edinburgh Data Intensive Machine (EDIM1) a data-brick computer cluster operated by the University of Edinburgh. The Seismic Data eXplorer (SDX) application – developed at Liverpool University, initially as part of the RapidSeis project – is a tool for seismic waveform analysis and has been further refined within VERCE.

Are you collaborating with any other countries or partners to meet your project requirements?

The VERCE project was launched in early October 2011 in Paris. It is composed of 10 European partners, all of whom have a specific added value, ranging from seismology to informatic research, to help achieve our objectives. Those partners will be working in close collaboration to complete this ambitious project during the next four years. Each of the partners brings important expertise to VERCE, and a number of them are leading contributors to the European Grid (EGI), HPC (PRACE) infrastructures, and international seismology data.

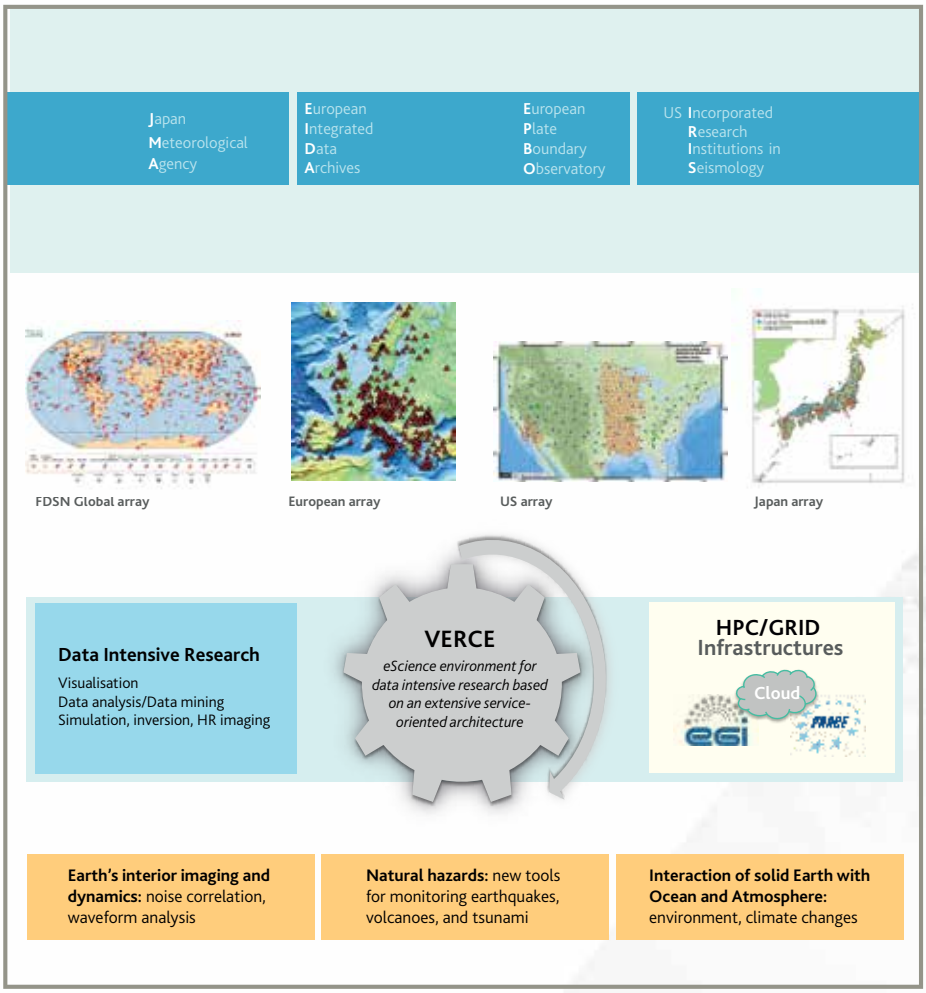
VERCE takes advantage of their visibility and established network to disseminate major news, events or promotional contents. The project has strategic synergies and collaborations with a number of related European projects in seismology: the ERC projects WHISPER and WAVETOMO; the ITN project QUEST; the research infrastructures projects NERA, SHARE and REAKT.

Through KNMI-ORFEUS and EMSC – the two seismological European NPOs of the consortium – VERCE has collaborations with other international NPOs in seismology like IRIS and Earthscope in the US; JAMSTEC and NIED in Japan. VERCE is also collaborating with large-scale European data infrastructure projects such as EUDAT, and with the European data and computing infrastructures: EGI and PRACE. Through EPOS, VERCE is involved in bilateral collaborations with the US National Science Foundation by way of the new COPEUS and i-CORDI initiatives.

Improving environmental **eInfrastructures**

By boosting earthquake and seismology research competitiveness, a collaborative initiative called **VERCE** is focused on improving Europe's eInfrastructures and information sharing in a highly data-intensive field

EARTHQUAKE AND SEISMOLOGY research addresses both fundamental problems in understanding the Earth's internal wave sources and properties, and augment applications to societal concerns about natural hazards, energy resources, environmental changes and national security. The earthquake and seismology community has for decades pioneered the prevailing philosophies of global, open data access and sharing. The nature of science is changing – new discoveries will emerge from the analysis and modelling of the large amounts of complex data generated by the increasingly dense global and regional seismology monitoring systems in operation all over the world. Today our ability to acquire observational data outpaces our ability to manage, explore, analyse and model them. To exploit the full potential of this wealth of data, and to guarantee optimal operation and design of the high-cost observation and monitoring systems, a new and holistic approach is required.



Funded by the EU Seventh Framework Programme (FP7), an initiative kicked off in late 2011 to provide a data-intensive eScience environment specifically designed to enable data-intensive analysis and modelling tasks that are simply not possible today. The project, known as Virtual Earthquake and seismology Research Community in Europe eScience environment (VERCE), aims to develop a comprehensive user-centric service-orientated architecture, and a platform of work and data flow tools and services that meet the differing scales and the diversities of innovative data-intensive analysis and modelling applications in seismology. VERCE also aims to provide 'intellectual ramps' for earthquake and seismology researchers to engage incrementally with the tools, techniques and methods of data-intensive research.

The work is a central contribution to the European Plate Observing System (EPOS) and the European Strategy Forum on Research Infrastructures (ESFRI) initiative in Solid Earth Sciences. In the words of Project Coordinator Professor Jean-Pierre Vilotte, the programme will also be delivering a Datascope, a new scientific instrument with unprecedented capabilities having a service-orientated architecture and built upon the European computing and data infrastructures – Grid, Cloud, High-Performance Computing (HPC) – and the international distributed archives infrastructure of the seismology community.

One of the first tasks for the VERCE team is to integrate the service-orientated architecture with an efficient communication layer between the distributed European public data and computing infrastructures, and private seismological resources. Another key task is linking high-throughput data analysis and HPC data modelling applications which is being

VERCE

VIRTUAL EARTHQUAKE AND SEISMOLOGY RESEARCH COMMUNITY IN EUROPE ESCIENCE ENVIRONMENT

OBJECTIVES

To develop a data-intensive eScience environment to enable innovative data analysis and data modelling methods that fully exploit the increasing wealth of open data generated by the observational and monitoring systems of the global seismology community.

PARTNERS

Centre National de la Recherche Scientifique, France • **Institut de Physique du Globe de Paris**, France • **Université Joseph Fourier de Grenoble**, France • **The University of Edinburgh**, UK • **Koninklijk Nederlands Meteorologisch Instituut**, The Netherlands • **Euro-Mediterranean Seismological Centre**, France • **Istituto Nazionale di Geofisica e Vulcanologia**, Italy • **Ludwig-Maximilians-Universitaet Muenchen**, Germany • **The University of Liverpool**, UK • **Bayerische Akademie der Wissenschaften**, Germany • **Fraunhofer-Gesellschaft zur Foerderung der Angewandten Forschung e.V.**, Germany • **Consorzio Interuniversitario Cineca**, Italy

FUNDING

EU Seventh Framework Programme (FP7) – contract no. 283543

CONTACT

Professor Jean-Pierre Vilotte
Project Coordinator

Institut de Physique du Globe de Paris
1 Rue Jussieu
75238 Paris Cedex 05
France

T +33 18 3957583
E vilotte@ipgp.fr

www.verce.eu

JEAN-PIERRE VILOTTE is Professor of Seismology at the Institut de Physique du Globe de Paris and head of the Parallel Computing and Data-intensive Analysis Centre of IPGP. He is Scientific Deputy for High Performance Computing and Information Technology at the CNRS-INSU. His research interests include computational seismology, mechanics and dynamics of earthquake sources.

achieved through a number of workflow and data sharing tools.

TESTING THE PLATFORM

The VERCE research platform is intended to operate on top of a set of distributed public and private data and computing resources provided by European and national grids, HPC eInfrastructures as well as a number of VERCE partners.

The team has completed analysis of nine pilot applications and use cases. Two have been chosen to be enabled by the VERCE platform because of their scientific importance and significance to the seismology community. The first is a data-intensive statistical analysis application that uses innovative noise cross-correlation methods to detect seismic sources, image the Earth's structure, and monitor transient changes in its properties. The VERCE use case was organised using re-usable macro-modules and processing libraries with parallel data management and staging, integrated into a distributed data and work flow environment allowing multi-site enactment and the use of Grid and Cloud infrastructures. The second is a data-intensive modelling application that enables high-resolution 3D seismic imaging. This use case includes data and workflows across community data archives, Grid and HPC eInfrastructures with different identification, authentication and data management policies. Such pilot applications are critical for testing the effectiveness of the programme and ensuring its robustness in data platform that can cut across scientific methodologies and applications.

SHARING THE KNOWLEDGE

One of the core work packages is the training and user documentation package which includes a number of plans focused on both internal and external training and education. The first of these has already been designed and planned and includes a shared community of practice across the consortium as well as targeting the wider seismology community. This is built in synergy with the ITN Quest and the ESFRI-EPOS projects and is coordinated with other EU eInfrastructure projects. Training workshops, documents, videos and webinars, from both VERCE and other relevant EU projects are collated and made available on the website. In addition, a helpdesk has been set up and a first internal workshop held late last year focused on the pilot applications and the work and data flow engines. All of this is intended to support the wider integration of data and information across the discipline and beyond. "This programme will provide a collaborative environment between the earthquake and seismology research community and computer scientists, system architects and data-aware engineers, fostering the emergence of research technologists with sustained mastery for data-handling methods and a thorough understanding of the research goals," expounds Vilotte.

In order to create this environment, a strong focus on dissemination and public outreach is required. Target audiences have been identified, including research organisations, industry representatives and national agencies. The dissemination plan included a review of communication tools that have already been set up by VERCE partners and it is proposed that two well-established communication channels, ORFEUS and EMSC, are to be utilised. Links to other relevant international projects, such as JAMSTEC and NIED in Japan, will help to broaden VERCE's reach. The partners are involved in presenting the project at a number of international conferences and European meetings.

The work package dedicated to knowledge and method sharing is planning to develop a community-specific web portal. This will enable access to the underlying set of HPC, Grid and data-intensive resources available within the VERCE consortium. Vilotte observes that the development of the gateway will require regular communication and information sharing with the seismology stakeholders so they can ensure the components are as relevant as possible for each use case: "Based upon the experience gained, the right balance between those looking for services to connect to and those who need graphical tools to interact with will be driven by the analysis of the scientific use cases through an iterative process involving several work packages".

With the programme due for completion in late 2015 there is still much work to be done. However, the team has progressed well and are on track to release the platform on time. Development of the VERCE architecture and tools for data analysis and data modelling applications are in progress. Several issues with regards to the orchestration of the data workflow across different eInfrastructures and private resources with heterogeneous access and data management policies have been addressed. A collection of resources and infrastructures to be integrated into the initial test bed has been described and a 'Plan-Do-Check-Act' cycle has been prepared to help manage the final platform release process.

Ultimately, it is hoped that the key outcomes and impact of VERCE will build Europe's scientific standing in earthquake and seismology research as well as improve and enhance the data exploitation and the modelling capabilities of this community. In turn, the research should make an important contribution to the European and national eInfrastructures.