Edinburgh Research Explorer

VERCE delivers a productive e-Science environment for seismology research

VERCE delivers a productive e-Science environment for seismology research

Malcolm Atkinson*, Michele Carpené†, Emanuele Casarotti‡, Steffen Claus§, Rosa Filgueira*, Anton Frank¶,
Michelle Galea*, Tom Garth‖, André Gemünd§, Heiner Igel**, Iraklis Klampanos*, Amrey Krause*,
Lion Krischer**, Siew Hoon Leong¶, Federica Magnoni‡, Jonas Matser††, Alberto Michelini‡,
Andreas Rietbrock‖, Horst Schwichtenberg§, Alessandro Spinuso†† and Jean-Pierre Vilotte‡‡

*University of Edinburgh, UK. Email: Malcolm.Atkinson@ed.ac.uk
†Supercomputing Applications and Innovation Department (CINECA), Italy
‡Istituto Nazionale di Geofisica e Vulcanologia (INGV), Italy
§Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Germany
¶Leibniz Supercomputing Centre, Germany
‖University of Liverpool, UK
**Ludwig-Maximilians-University, Germany
††Royal Netherlands Meteorological Institute (KNMI), The Netherlands
‡‡Institut de Physique du Globe de Paris (IPGP), France

Abstract—The VERCE project has pioneered an e-Infrastructure to support researchers using established simulation codes on high-performance computers in conjunction with multiple sources of observational data. This is accessed and organised via the VERCE science gateway that makes it convenient for seismologists to use these resources from any location via the Internet. Their data handling is made flexible and scalable by two Python libraries, ObsPy and dispel4py and by data services delivered by ORFEUS and EUDAT. Provenance driven tools enable rapid exploration of results and of the relationships between data, which accelerates understanding and method improvement. These powerful facilities are integrated and draw on many other e-Infrastructures. This paper presents the motivation for building such systems, it reviews how solid-Earth scientists can make significant research progress using them and explains the architecture and mechanisms that make their construction and operation achievable. We conclude with a summary of the achievements to date and identify the crucial steps needed to extend the capabilities for seismologists, for solid-Earth scientists and for similar disciplines.

Keywords-Science Gateway, HPC, Data-Intensive, Data Science, Metadata and Storage, solid-Earth Sciences, Virtual Research Environment, e-Infrastructure.

I. INTRODUCTION

The EU VERCE project[1], "*Virtual Earthquake and seismology Research Community e-science environment in Europe*", has developed a comprehensive and integrated virtual research environment (VRE) for computational and data-intensive seismology. This has been pioneered with particular simulation models and data-driven seismology examples. We report it here because the challenges it addresses are widespread and of growing prevalence and the solution strategy, which covers organisational and presentational as well as technical issues, is of wide relevance and applicability. VERCE typifies current challenges in combining simulation results with observational data as a research community grasps the opportunities presented by increased computational power and the growing wealth of data while using existing resources and practices. Climate modelling has already addressed such issues under the aegis of the IPCC. Sect. II presents comparisons and Sect. VI presents key lessons learnt and outstanding issues.

Seismology is one of the very few means of studying the sub-surface structure and phenomena of the Earth. It provides an opportunity to model physical processes and compare the simulation results with surface observations from digital seismometers. The field is made more complex because it also addresses a sustained societal challenge of natural hazard mitigation [1]. We hypothesise that there are many cognate disciplines that will need to develop integrated e-Infrastructures supporting their communities in order to bring simulations representing their models and diverse observations into a conveniently used framework to enhance the productivity and capabilities of their research teams. We anticipate that many of these will have other drivers, such as applications to benefit society or business.

In such contexts it is essential to empower researchers with capabilities to explore options and to make rapid innovations. Ideally they should be able to transfer the results from such creative contexts into large-scale production runs easily but that is constrained by two necessary compromises: *a*) the production systems and major resources need to be well protected (against accidents and malevolence), and *b*) some aspects of the work, such as preparing simulation codes to exploit the highest performance platforms well, require intensive team effort involving different experts. We report our path developing an effective VRE while handling these constraints in today's rapidly evolving digital environment.

The present VERCE e-Infrastructure involves the following major elements moving from the researchers' point of contact to the contextual digital resources:

- The *scientific gateway* that is Web accessible from anywhere, provides an integrated view of all available resources, handles continuity between sessions and

---

supports collaboration with shared data and methods, and with pervasive data access controls.

- The *mapping to multiple Distributed Computing Infrastructures (DCIs)* handling identity management, authority controls, transformations between representations and protocols, and access to resources.
- The *framework for computational science* that accommodates simulation codes, arranges for their use including the supply of input data and the acquisition of results with associated metadata.
- The *framework for data-intensive science* that delivers fine-grained composition of algorithms and provides a smooth transition between development and production.
- An *extensible coupling* with legacy and external services, and ingest of non-standard or bulk data.
- A *set of tools* to support validation, quality control and impact exploiting pervasive provenance records.

The paper has the following structure. Sect. II examines the requirements from researchers and their modes of using the VRE. It examines how this may evolve and extend into other disciplines. Sect. III outlines the architecture and key technologies used for each element of the VERCE VRE. Sect. IV reports on the difficulties encountered setting up this coherent, European-wide accessible research platform—those overcome and those persisting. The benefits of building the VERCE e-Infrastructure and delivering its capabilities as a VRE are presented in Sect. V. Sect. VI summarises the progress made, evaluates it in a contemporary context and identifies priority research directions.

II. USE OF THE VERCE VRE

There are today large networks of digital seismometers, many of them permanently deployed, and some being moved periodically or deployed temporarily for a specific purpose. All of the seismic traces that they record are collected, standardised and made accessible under the aegis of the FDSN[2], provided in Europe by ORFEUS and EIDA[3]. This growing wealth of data is used to characterise earthquakes (*e.g.* energy, location, depth and source mechanism). This serves two main purposes: *a)* alerting relevant responder authorities if the magnitude and proximity to populations warrants triggering further analyses immediately, *e.g.* calculating tsunami scenarios for earthquakes in sea areas; and *b)* building catalogues of event properties, *e.g.* GCMT[4] [2], [3] that are used by computational seismologists studying the Earth's structure and earthquakes' sources, and for estimating hazard and risk. As in many sciences, seismologists observe data, infer possible physical models at the origin of the data and then compare the results of the modelling with the observations. They proceed by formulating a "*mesh*" as a finite element model (FEM) of a volume to be studied—see Fig. 1. This involves estimates of density and wave speeds

for each element. Building such meshes takes considerable effort and skill [4]. For each event in the region, the propagation of seismic waves is then simulated generating *synthetic* seismograms for each seismometer. The observed and synthetic traces are then compared in a process called *misfit analysis*. The detected differences can then be back-propagated to refine the model in a process called *inversion*. This can be done for all events in the region, and as the model of the Earth's structure improves it can be extended to finer resolution. This is computationally intensive; current inversions require up to 10^7 CPU hours *e.g.* [5]–[8].

Figure 1. Forward wave propagation, misfit analysis and inversion.

The forward wave propagation is well understood, but it requires considerable skill, effort and engineering to develop codes that simulate this efficiently by exploiting the power of modern HPC systems. A prime example, scalable to more than 10^5 cores, is SPECFEM3D_Cartesian[5] [9], which has been made much easier to use by the VRE. Prior to VERCE, each seismologist undertaking such a study needed to find their own resources, understand the documentation of a simulation package, install and configure it, organise all of their runs, marshalling inputs and collecting results. The VRE automates all of this and presents it as a simple operation in the science gateway, where a seismologist can select a region, choose an existing mesh or supply their own, to run the forward model of a chosen event for a specified time. The science gateway enables seismologists to specify misfit analyses, selecting signal sources and comparison methods, as the next step in tomographic inversion. The VERCE platform keeps provenance records to help scientists examine or repeat what they have done (Fig. 1).

Also, the collection and archiving of the continuous traces from seismometers has shown another route to seismic tomography. The previously disregarded ambient noise, which comes predominantly from ocean generated energy, can be correlated for long time intervals to yield the Green's function[6] between each pair of seismometers, *e.g.* by stacking a year's worth of day-long trace correlations. This can reveal variation of the Earth's material properties with time. It is data-intensive, requiring $O(n^2)$ correlations of n trace channels. VERCE provides a toolkit that automates and packages the required variations of this process and provides provenance tracking, saving seismologists many detailed data-management steps.

However, in both cases, scientists need to be able to select the region and frequencies of interest, and choose the processes to be run. This particularly applies in both cases to the quality controls and data preparation on the recorded seismic waveforms [10] and to the post-processing, identification and preservation of result data together with any additional annotations that the seismologists require. An extensive scientific Python library, ObsPy [11], provides a repertoire of data-processing steps for this purpose and for ambient noise correlation, improving consistency and saving much re-implementation, thereby allowing seismologists to focus on their research rather than handle many details. The dispel4py Python library, described below, provides a data-intensive system to deliver scalability and flexibility for data-intensive processing [12] with a simple abstract model. These empower researchers to formulate the data handing and transformations they require. Currently seismologists use these toolkits directly to explore their potential prior to packaged options being provided in the science gateway.

Seismic traces have further uses, *e.g.* to detect landslides and glaciers calving, to monitor and analyse induced seismicity and other anthropogenic activity [13], and to measure the effects of global warming on ocean wave energy. The set of solid-Earth sciences needs interworking of models and data, for example geodesic measurements, such as GPS data, LIDAR surveys, clinometers and photogrammetry, can be combined with seismic traces at fault-zone and volcanic observatories. They can then be compared with computational models of the geophysical processes. Similarly, models of mineral nanostructures, geochemistry and brittle-rock fracture can be compared with laboratory induced fractures and geology in former fracture zones now exposed.

An integrated VRE that incorporated such simulation systems and observational data will result in significant advances in geosciences. The integration of geophysical resources planned by EPOS (the ESFRI infrastructure for solid-Earth sciences[7]) will provide the critical organisational context and many of the key resources. The geo-specific examples portend a much wider application of the VERCE

approach that will require access to observational and experimental data. It will depend on the simulation codes and data analyses being made efficiently scalable, and on provision of resources for simulations and for data.

This will depend on a well-founded incorporation of diversity in software, data, computational systems and institutional goals—an organisational as well as technical challenge to deliver a coherent research environment in the context of loosely coupled systems and independent institutions that serve many other goals. Research changes understanding and therefore goals. Commercial pressures and innovations drive changes in technologies. Diversity will therefore continue. It will not be subdued by emergent business models or standards; these become extra members of the digital ecosystem in which research must thrive. Without an integrating framework that embraces inherent and evolving diversities researchers will continue long and tedious struggles facing diversity alone, never reaching the full power available. VERCE pioneers a shared investment delivering a coherent research environment in such an evolving diverse ecosystem and makes the advantages of the coherence and the technical strategy available to others. For example, to assess and respond to natural hazards or to respond to the societal challenges for renewable energy and resources.

The SCEC project[8] strives for better prediction of strong ground motion, which requires better 3D models. It includes work on simulation codes, whereas, in principle, VERCE hosts independently developed code. It differs from VERCE in delivering the services from one HPC platform and site, thereby simplifying the e-Infrastructure required, but missing the potential combination of multiply owned resources. Its users may have less opportunity to provide their own finite-element models and to shape their own collection, preparation and post-processing of seismic observed and simulated data—a tradeoff still being explored between flexibility, complexity and community ownership in VERCE.

The Shake Movie service run by Princeton University[9] uses SPECFEM_GLOBE to generate simulations corresponding to recent earthquakes. Here users can access the simulation results but cannot set up and steer their own simulations.

III. THE VERCE E-INFRASTRUCTURE

Fig. 2 shows the major components of the e-Infrastructure needed to deliver the VERCE VRE. Simply integrating, presenting and tailoring technologies does not complete a VRE. That requires communities driving the research, collections of relevant data, application-software tuned and maintained to meet the latest research requirements and to exploit hardware advances, and teams of ICT experts maintaining its leading capabilities. However that framework and its usability and sustainability is key to attracting

---

[6]The response at one seismometer to an impulse at the other.
[7]www.epos-eu.org

[8]scec.usc.edu/scecpedia/Geoinformatics_Project
[9]global.shakemovie.princeton.edu

these researchers, enabling their collaboration, creating effective interplay between ICT experts and domain-focused researchers, gaining access to the resources required and amortising costs over sufficiently broad communities. As already noted this has to embrace challenging diversity in virtually every element.
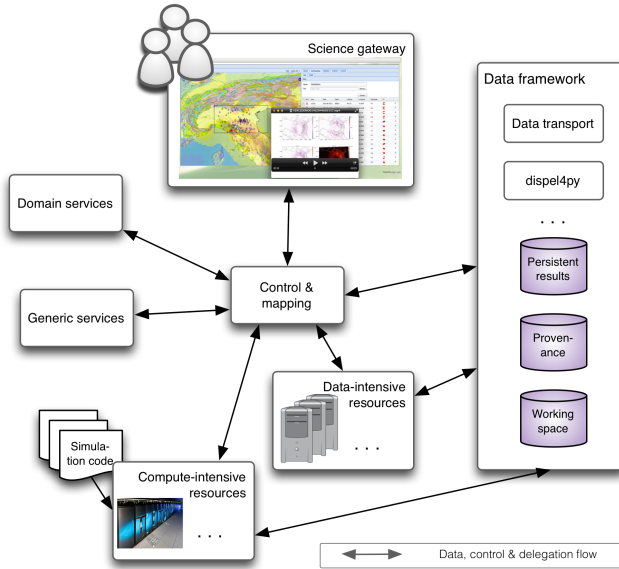

Figure 2. Overview of the VERCE e-Infrastructure architecture.

A. The VERCE science gateway

A science gateway has to support all the required modes of interaction with all aspects of the VRE needed by the geographically dispersed and often mobile research community. It has to support training and induction into this community, organisation of sharing, distribution of responsibilities and attribution of credit. It needs to provide convenient and understandable access to tools and functions that enable all stages of the research methods, such as those illustrated above, from their creation to their application to build evidence for conclusions, publications and actions. Whilst enabling the domain scientists to be as productive and agile as possible is the highest priority, the work of the technical teams supporting them should also be facilitated, with many of the technical chores automated to improve efficiency, accelerate delivery of new facilities and reduce error rates.

The science gateway is built on SCI-BUS [14] which itself builds on Liferay[10] and gUSE/WS-PGRADE [15], which provide an Internet portal framework and a task-oriented workflow system respectively. Their wide use promises shared future maintenance and their maturity and integration implied reliability. The gateway organises and structures the set of available options as portlets with drop-down menus of the required functions, different sets of functions being

visible depending on the authority and mode of the logged-in user. A typical function will be illustrated using the modelling of seismic wave propagation introduced above—see Fig. 1. Via standard geospatial interfaces seismologists choose their region of interest, which then exposes the existing models and available seismic stations, from which they choose a mesh and velocity model. They then select a set of seismic stations to be used for simulations and misfit analysis, and select a relevant event, *i.e.* an earthquake record from GCMT, or from other seismic source solution catalogues, e.g. TDMT[11] from INGV [16]. Users can supply their own mesh, velocity model, seismic observations and pre- or post-processing steps at each stage. They then submit a simulation code run (currently using SPECFEM3D). The gateway *a)* supports region selection by geographic name, by map zoom and by return to previous session contexts, *b)* it provides structural-model and mesh upload or selection from an existing library of meshes and models, *c)* seismic stations are selected by clicking their icon on the region map, *d)* the gateway queries the GCMT or TDMT Web services and displays a list of seismic events.

To perform the submission, it uses the SCI-BUS DCI-Bridge [17], which handles many aspects of submitting a workflow to different targets, transporting data and security credentials, manages job submissions and collects status updates—an example of reducing the ICT chores. Arrangements to match the simulation code, the workflow, the credentials and the data handling, still require substantial effort from ICT experts in current target systems—see below. Obtaining relevant traces and station metadata for the identified stations for the propagation time of the waves for each event is implemented as queries to seismic archives. The preparation (*pre-processing*) of these selected traces and the synthetic traces then uses a common dispel4py pipeline that seismologists have previously designed and installed. This can also be fed to a misfit analysis, also using dispel4py, that yields visualised results. A seismologist makes a decision as to whether these are worth preserving and whether to use the computed misfit measurements to validate the geological model or as a source of back-propagated wavefield that interacting with the forward wavefield, highlights the required corrections to improve the structural model. This back-propogation step is not currently set up in the VERCE gateway.

To accomplish the above functionality the gateway has to support all of the user interactions by generating graphical presentations and by interpreting users' actions. The interpretation involves the generation of workflows shaped by those actions and by the session history. It maps from workflow templates (developed by experts) to target-specific workflows and job descriptions in submission format.

The sites all require that identity, security and authorisation issues are handled and their diversity must be addressed. The gateway attempts to provide uniform access to a federation

of sites. Two strategies are available: integration of user identity and account data into a centralised database or use of a decentralised schema. For example, PRACE[12] distributes user-account information across LDAP servers replicating users' information on each site, which can then have local policies. This forces administrators to maintain their own authentication service. Consequently, to use resources users have to register with each site via its system administrators. General software products support the centralised approach; *e.g.* Shibboleth[13] seeks a complete solution to authentication using SAML [18] to allow people to sign in with just one identity to independently run systems on which they have an account. An identity provider supplies information as SAML assertions, while resource providers use this information to adjudicate access. This offers flexibility but sufficient trust and wide adoption is needed. For VERCE such consistency is not yet available, *e.g.* GridFTP (data transport) uses proxy certificates that do not support SAML. UNICORE[14] (middleware used by some HPC providers) supports Explicit Trusted Delegation (ETD) [19] while iRODS[15] (used for federated data and metadata storage) provides authentication via Pluggable Authentication Modules (PAM) [20].

Addressing such diversity is essential. The computations require substantial resources and different researchers have different sites at which they have allocations. These change as a result of competitions, policy changes and transfer between projects. Careful tuning of each simulation code to exploit the latest advances in parallel architectures is needed; consequently for each run different codes have different targets where they perform best. Economic and technical motivations warrant service amortisation over multiple communities, *e.g.* data transport and preservation delivered by EUDAT[16]. We revisit such issues in Sect. IV and Sect. VI.

B. Composing DCIs

We use diverse HPC e-Infrastructures and access remote storage systems such as B2SAFE[17] from EUDAT and FDSN services from EIDA; these use different DCIs. Further diversity is required *a)* to transition between phases of research from exploration and development, on small convenient facilities, to production runs, and *b)* to efficiently handle loads across the compute-intensive to data-intensive range. Changes in the user expectations appear as a consequence, *e.g.* they expect a simple and familiar log-in to their own machine or institution's DCIs. During field work and when collecting observations from remote temporary deployments they may work disconnected from the Internet and independently from reference services. Normally they expect a closely coupled system with automated access to many services. Transitions between such contexts require bulk ingest

of new observations into the framework. Simulation code optimised to fully exploit advanced parallel architectures needs to be located on the architecture for which it was optimised for production but can be hosted on lower performance clusters and elastic clouds for smaller-scale development runs. Data-intensive workloads (see Sect. III-D) will under-utilise CPU resources on HPC platforms, so data is moved to appropriate clusters for processing. Specialised data-intensive clusters, such as the Terracorrelator[18], outperform standard clusters and clouds [21]. Parallel I/O and dynamic data compression is needed for production scale runs [22] but standard data operations found locally are adequate for test runs. Commercial clouds are used for initial deployments, *e.g.* OPENSHIFT for the GCMT FDSN service development. VERCE has to integrate diverse DCIs—DCI-Bridge [17] substantially reduces the effort needed.

C. The computational science framework

Researchers need to use institutional and national facilities where they have routine access, and transfer to the highest performance facilities, such as the PRACE sites, when they are granted time to tackle demanding simulations. They require help from the VERCE framework moving their work, with its associated code sets, workflows and data, between computational facilities, and they need to retrieve results and provenance records into their sustained working context. Beyond facilitating the transfer of work and data, this requires careful tuning of code and workflows for the highest end context. Once systems and software engineers with specialist knowledge have done such preparations the VERCE framework supports the transitions and presents the computational services uniformly to researchers entitled to use them. This entitlement is facilitated for training and induction, but it remains challenging for significant use because: *a)* the cost, particularly in energy, of such use is very substantial, *b)* the risk of malevolent misuse of such powerful systems has to be minimised, and *c)* constraints are imposed on owners of such resources.

D. The data-intensive framework

The data-intensive facilities build on the dispel4py framework and on a set of common data services—see Fig. 2. dispel4py enables scientists to develop their own data-intensive applications using Python, familiar tools and their normal work environment, and to map automatically to a range of target systems adapting dynamically to the scale of data [12]. There is a core library to support initial development and mappings to four scalable target enactment models: MPI [23], Apache Storm clusters[19] and shared-memory multi-core machines using Multiprocessing (a Python package for concurrent interpretation) and Apache Spark[20]. These mappings employ established middleware

---

[12] www.praceri.eu
[13] shibboleth.net
[14] www.unicore.eu
[15] wiki.irods.org
[16] www.eudat.eu
[17] www.eudat.eu/services/b2safe
[18] www.wiki.ed.ac.uk/display/Terra/Terra-correlator+wiki
[19] storm.incubator.apache.org
[20] spark.apache.org

to benefit from its substantial development effort and implementations on multiple DCIs. dispel4py integrates with provenance mechanisms—see Sect. III-F and has a registry to promote method and component description and sharing.

dispel4py encodes data-intensive methods as a graph of *operations* coupled together by *data streams*. The operations are instances of Processing Elements (PEs) that consume units of data from their inputs and emit units of data on their outputs. They may have zero or more inputs and zero or more outputs. The units of data are application dependent, and may be of any size, from a small scalar to a complex object, such as a multi-dimensional array with its metadata. The data streams may use any communication mechanism. They may use compression and provide sufficient buffering. They deliver, preserving order, the units of data supplied by an output or external source to all of the inputs to which they are connected. The operations are instances of Python classes and the stream connections are made by Python operations on these classes. Initial sets of these PE classes can be downloaded as a Python library or obtained from the registry. Both scientists and experts extend the set of PEs by writing their own algorithms or wrapping legacy algorithms using Python's extensive scientific libraries and call outs to other languages. A PE may also wrap a subgraph so that it may be used as an operation. The connections use local in-memory communication if this is available, and direct communication over local or wide-area protocols; they do not go via disk I/O unless buffering overflows. Explicit writes to disk for final results or diagnostics require parallel I/O in production contexts.

When the script is run, the graph is built as Python objects. In a development context this is interpreted locally. When larger scale runs are required, the graph is optimised, *e.g.* PEs are clustered to minimise data transport within loading constraints, the target DCI and intermediate technology is chosen, and the graph is mapped onto these targets with appropriate pipelining and parallelisation. The use of data streaming has three benefits: *a*) the low overheads, particularly for local coupling, means that it is efficient for scientists to compose low-cost steps, called "*fine-grained workflows*", *b*) the direct handling of streams means that scientists can develop methods for live continuously flowing observations, and *c*) the load placed on the increasingly limiting bottleneck of disk I/O, Kryder's law [24], is minimised. The abstract form of dispel4py together with automated mapping means that scientists are not distracted by technical details and their encoded methods are not locked into a particular target. The automated mapping also relieves experts of the tasks of optimising and mapping for particular DCI architectures. The choice of DCI target and mapping is critical for production performance, *e.g.* Filguera *et al*. report for a 1000 station cross correlation, the Terracorrelator running slightly faster than SuperMUC[21] using MPI and 5.5 times faster than Amazon EC2 [21]. This becomes more significant

when we note the low cost and energy consumption of the Terracorrelator—many organisations are now acquiring such machines designed for data-intensive loads [25]. The widespread availability of Python, including on supercomputers, increases the number of potential targets and reduces requirements for ICT experts to perform installation.

VERCE also needs to integrate diverse data-management systems, including file systems on the many computers in use, DBMS and shared data services. These include multiple data storage systems and services for finding, referencing and moving data between them, as shown in Fig. 2. Data needs to be in the right place at the right time, *e.g.* on a cluster about to run a simulation if it is input and removed from that cluster immediately the job completes if it is a result set. This requires careful scheduling of many data movement operations; these are automatically incorporated into workflows, saving scientists from such details. The underlying mapping system needs to choose the right mechanism for the job, *e.g.* GridFTP for large volumes as it handles recovery and retry, but ftp and http for small transfers. In the longer term, shared services provided by EUDAT and generic technology such as Data Avenue [26] should reduce the ICT chores involved, but understanding and specifying the required movements, resource management, authorisation, preservation and persistent identification will still be needed. The distributed system for results and their metadata is built on iRODS [27], and this links closely with provenance, described below, which uses MongoDB[22] as its distributed NoSQL DBMS. The broad user bases of iRODS and MongoDB motivated their selection, support for iRODS from EUDAT helped sustainability, and MongoDB combined flexible data strutures with scalability.

E. Connecting external facilities

These fall into two groups: *a*) domain-specific services, in this case seismology, geoscience and geographic services, and *b*) generic services that support these and other communities, and facilitate interworking between software systems. These categories form a continuum as some services, such as those providing map and political data are widely used. The challenge is to achieve a balance between making it easy for specialists to use their particular sources of data and facilitating interdisciplinary boundary crossing—both access to other disciplines' data from VERCE and access to VERCE's data from other disciplines.

The immediate needs of the research activities described above are mainly concerned with getting access to seismic observations, geological models, the meshes and earthquake event parameters. The seismic traces and the metadata associated with the seismic stations, often come from FDSN compliant services, however they can come from *ad hoc* deployments organised for specific purposes, *e.g.* to study structure exploiting aftershocks, or as international

---

exchanges of large collections of data, *e.g.* the Japanese data relating to the 2011 Tohoku earthquake. The FDSN sources can be interrogated by Web services, but the other data needs careful ingest into the system, to adopt relevant standards, to generate necessary metadata and to build indexes that allow their use via the same framework and underlying workflows. This is an open-ended requirement that has to involve experts who understand the form and origin of the data. However, they are provided with a set of processing elements that they can select, combine and parameterise and if necessary augment using dispel4py to establish such *data ingest* methods, which can then be reused and refined. These tools are used directly at present, while understanding about how best to package them in the science gateway develops.

The geological models and meshes are less standardised and strongly depend on the simulation code adopted. The mesh generation by a computational seismologist requires topography and geological data, which is often obtained from external services, FEM building tools, *e.g.* CUBIT[23], and knowledge about how mesh properties affect running times and model resolution. It is difficult to integrate a meshing generation toolkit into the VRE due to licensing costs and limitations of Web interaction. On the other hand, we are exploring possibilities to host validation workflows that would allow at least a sanity check of the user-supplied mesh, before its adoption for simulation. VERCE has initiated a useful catalogue of meshes and corresponding tomographical models, which others may use and to which they may contribute. Such Web accessible resources will be key elements in motivating collaboration, while a mesh generating workflow is being investigated.

The event parameters are obtained from sources such as the GCMT or the INGV catalogues, which use different estimation methods, or are the result of local analyses. These will increase in complexity as the models of the physics of the slip processes are developed. Geodesy, interferometry (InSAR—Interferometric synthetic aperture radar) to measure ground motion and space-image data may be used to support further analyses in those cases. To orient the simulation setup and results, *e.g.* as visualisations or videos, we make use of more geospatial and geopolitical context-setting data and appropriate visualisation services, typically those supporting OGC[24] standards as mandated by the INSPIRE directive [29]. These services are provided for public use from a number of organisations and we currently integrate those offered by European initiatives including OneGeology[25] and SHARE[26] and governmental institutions such as the KNMI, providing respectively geological information, regional hazard maps and other environmental visualisation overlays. Other disciplines, *e.g.* civil engineering, mineral extraction and urban planning, may access the resulting

data, for instance to investigate anthropogenic seismicity. They may even use the platform for their regional studies, adopting private models with restricted access policies. This last case is supported by the authorisation-based access control to the data, which is in place throughout the storage elements of the platform.

F. Provenance-driven tools

To achieve quality, science must be replicable, *i.e.* the same or a different scientist should be able to repeat a digital experiment. This is enabled in VERCE by consistently handling provenance records across the diversity of subsystems. This has a further payoff; it results in a consistent model of the relationship between methods and data that is presented and exploited by tools that help scientists explore and validate their data and processes.

The VERCE science gateway combines workflows interactively behind its tailored user interfaces. This integration of specific and generic tools led to a monitoring and validation system that is used to evaluate *at run-time* the progress of the computations for all of the interactive services accessed via the gateway. This needs to present information to users specific to their application domain that enables them to make decisions that have immediate effects. For instance, if we consider long-running cross-correlation analyses or high-resolution simulations, detecting errors in results or metadata during the run, can trigger reactions that can save users from unfruitful waits and from wasting energy or expensive computing cycles. Such direct control makes their research more efficient and progress faster.

Pauw *et al*. [30] argue that the generation of huge quantities of provenance traces is excessively demanding for an effective visualisation tool, and it overloads users. To reduce this cognitive overload and to provide orientation when investigating properties of the processes and data, we specify and implement a Web API, targeting common interrogation patterns, which are combined and orchestrated via a user interface. This provides unified access to fine-grained provenance, as well as summary information and both user and domain-specific metadata, as recommended in PROV-DM [31]. Its implementation takes advantage of the flexibility of NoSQL technologies, such as MongoDB.

Users search for runs and for data products by submitting metadata queries as value ranges. They examine search results by browsing the metadata associated with the data stream, including the data relationships. This is illustrated in Fig. 3. The dynamics of this exploration develops very detailed insight into a workflow's logic. Generated files are linked by the provenance metadata and made available for visualisation or download, by accessing the underlying data management system. For rapid and large size download operations, a bulk download script matching the search results is automatically generated. This allows users to fetch their selected data in batch mode, into their local facilities

[23]cubit.sandia.gov
[24]www.opengeospatial.org
[25]www.onegeology.org
[26]www.share-eu.org

adopting appropriate data transfer protocols. Portability is supported by enabling users to export full traces of runs in W3C-PROV format. This will foster the hand over of a final research product to institutional and multidisciplinary archives, where provenance is included in their data-curation processes.

Provenance systems typically collect metadata about the data transformations [32] during workflow execution. We have introduced additional features to trigger actions using third-party components to react to provenance information. An operational example in the VERCE, is the runtime shipment of the intermediate data from a DCI to external systems, to facilitate rapid diagnostic evaluation or visualisation. In many cases this behaviour can't be instrumented within the workflow itself because of the lack of external connectivity from HPC clusters.

IV. BUILDING THE VERCE VRE

Science gateways are one of the most discussed, often criticised, but yet required components of a VRE. The most challenging task that we had to face was how can we make it actually useful and transparent for our community. Our solution was to engage people in *task forces* where committed users, developers and service providers use co-design and co-development to overcome limitations that seismologists would otherwise experience. This changed the focus from HPC production environments and community data services, where middleware standards, security procedures and connectivity dominated, to a user-focussed scenario. Without these task forces, users would have run away rather than engaged with the VERCE platform.

An example task force focused on misfit calculation. It involved two computational seismologists, an expert developer of the ObsPy Python library, two dispel4py experts, a data expert, an expert in provenance and the developer of interaction templates and portlets. After intensive preparation, it had a three-day kick-off meeting, at which current systems were tried and ideas for improvements were hatched and critical steps evaluated. Two months of intense development and collaboration followed, with daily communication between pairs attacking subgoals. A further face-to-face accelerated the integration of the components, towards exhaustive tests and refinement before production deployment. These task forces adopt agile development methods but have a greater focus on teasing out what new scientific methods are becoming feasible as the digital context improves.

We had to accommodate users with different levels of expertise. Some require an advanced programming environment. Others require results from simulations with a variety of different input configurations and parameterisations. This requires significant house keeping, detailed technical knowledge and awareness of the potential offered by the external domain-specific services and facilities which are out there to serve the community with data and tools. During our training sessions we discovered that, regardless of their level of

expertise, users are often unaware of or ignore such services. By automatically organising data movement, house keeping and resource use, VERCE not only accelerated and simplified the functional steps, it also migrated work to suitable targets. This was demonstrated from the trainees questions which, as a result, were focused on the parametrisation of the simulation and the modelling methodologies. Also the novel interactive paradigms for result validation via metadata and provenance exploration, were used enthusiastically.

Some parts of the VERCE e-Infrastructure implementation are shown in Table I[27] The first two columns use technology, such as simulation codes and FDSN Web services, that is specific to computational and data-intensive seismology respectively. We anticipate that placing equivalent data, technology and functions here for another discipline would now be straightforward as the relationships with the other two columns have been established. In many cases there are templates and workflows that could be suitably modified and reused. These last two columns provide an e-Infrastructure and many generic services that would be equally necessary and applicable in other contexts. As far as possible existing technologies, *e.g.* SPECFEM3D, Liferay, DCI-Bridge and iRODS, and services have been used to reduce delays and share maintenance costs. This should be taken further as systems such as the products of the EUDAT projects become available. In most cases these need to be deployed on hardware, systems software and networks that are shared or used for multiple purposes.

Resilience and sustainability depend on appropriate insulation from the changes in software services, data representations, network protocols and hardware architectures. We start with traditional insulation, such as: *a*) high-level languages, *e.g.* Python, and operating systems, *e.g.* Linux, to span diverse computers, *b*) computational frameworks, *e.g.* UNICORE and Globus, to span operating environments, *c*) data management frameworks, *e.g.* iRODS, to span storage systems, *d*) job submission and workflow enactment frameworks, *e.g.* DCI-Bridge, protecting against diversity and change in computing services, and *e*) Web and document management, *e.g.* Liferay, providing a consistent environment for interaction development. VERCE builds on this to protect investment and composes large algorithmic elements using WS-PGRADE/gUSE workflow enactment system, that works with all the DCIs that we encountered. It also handles many of the detailed requirements of credential, authority and identity management. VERCE composes smaller-scale tasks and fine-grained operations using dispel4py because of its low overheads and multiple mappings to multiple data-intensive middleware systems. These two composition methods offer two advantages: *a*) the development of templates

---

[27]All subsystems in that table are operational in the actively used system. Three are still under development: (i) misfit analysis still requires some user interfaces in the science gateway and is used directly at present, (ii) the pre-configured library of pipelines continues to have additions, and (iii) the bulk data ingest kit is used directly and not yet packaged in the science gateway. Security requirements remain an issue as described below.
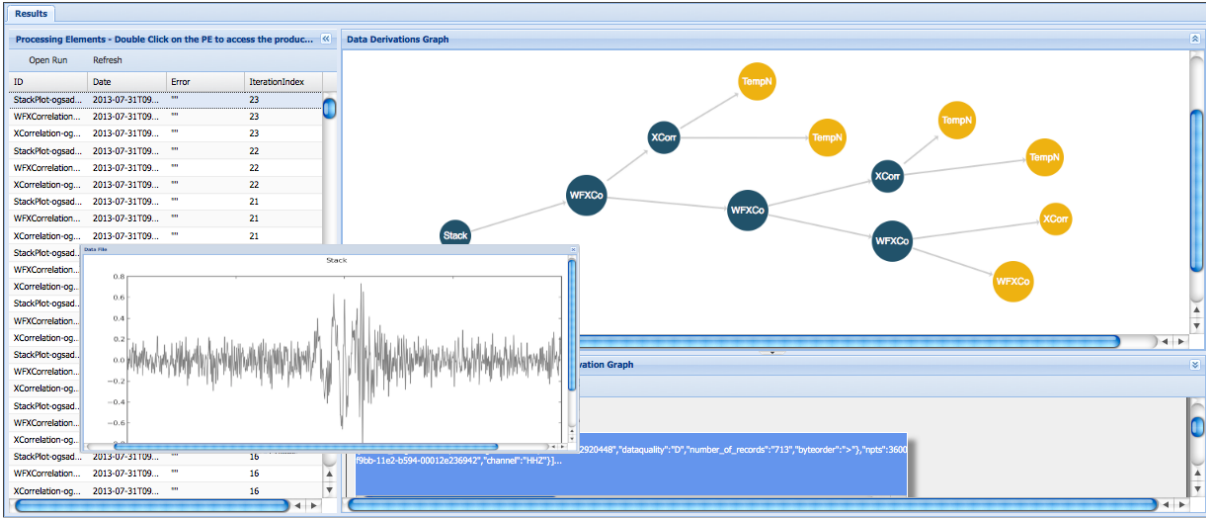
Figure 3. Provenance Explorer GUI: Visualisation of the interactive derivation graph and data preview for an intermediate trace of a cross-correlation workflow. The dots represent data at a transformation stage. Yellow dots are expandable to allow the further navigation through the dependency graph.

that can be reused for other domains and adapted at a high level for new target environments, and *b)* connection with larger user communities that will invest in mappings to new technologies. Inevitably, using a series of insulating layers also reduces the scope for innovation and new approaches as identified by Gesing *et al*. [33].

Management of security, identity, authorisation and accounting is still problematic. We attempt to provide a consistent continuum of facilities and resources so that researchers can experiment and innovate freely in a low-cost, low risk environment and move their work fluently into powerful production contexts. The former context imposes very few security hoops for researchers to jump through. The latter require careful protection because of their expense and potential for misuse. In VERCE we try to combine open-data and free to use services, with security mechanisms adopting the VOMS[28] model, which is an advanced and effective system, technically and conceptually, based on the grid X.509 certificates. It presents some weakness, *e.g.* in the definition of roles so that they are semantically consistent across contexts. It is adopted by service providers involved in the EGI initiative, but not yet fully recognised by others. Consequently, today there are different security hoops to leap through in almost every context, and many of them require lengthy set up. A wide variety of computational and data-intensive researchers and innovators face this problem. It is surprising that there is not a European-wide or internationally accepted, sufficiently trusted, sufficiently multi-scale robust solution to achieve easy migration between research contexts. The impediments are largely organisational and political. Lack of it continues to inhibit uptake; thus depriving researchers of significant potential. Steps to resolve this disharmony need to involve many communities, many categories of provider and all e-Infrastructure campaigns;

VERCE exposes the issue, others must solve it.

V. BENEFITS FROM THE VERCE VRE

Researchers report completing computational seismology tasks in a few hours that would have previously taken days or weeks. They use a wide array of computational facilities to satisfy their simulation requirements. Daily research activities are satisfied on the institutional resources where they have routine access while ground breaking research activities that require higher performance capabilities are achieved on national and European facilities, such as PRACE*. The VERCE framework enables the researchers to work seamlessly on such facilities without having to learn the heterogeneity of the underlying environment. With VERCE's support, potentially any associated code sets, workflows and data can be administered.

Indeed, one leading seismologist described the progress in this way, "*We have experienced three phases of computational seismology. 1) Write your own code, parallelise it and do science organising everything yourself. It doesn't work anymore—it takes too long and the hardware is too complicated. 2) Download community code developed by a team. Install it and run your jobs. It still requires organising everything yourself and much technical knowledge. 3) Use a platform like VERCE. The big benefits from the platform are you can focus on your science and productivity improves by one or two orders of magnitude.*" [34]

The VERCE platform tackles commonly experienced cross-disciplinary issues encountered by users as they develop, test and run *in silico* experiments and simulations. It provides immediate access to tools to explore the metadata, the dependencies and the diagnostic logs related to any application activated via the science gateway. We move from a very practical and science-driven way of setting up an *in*

Computational seismology	Data-intensive seismology	Data management	Framework
SPECFEM3D wave simulation	ObsPy Python library	iRODS federated store for seismic data	Science gateway using Liferay and gUSE
Misfit analysis library	Seismology PE libraries	dispel4py Python system	Job submission via DCI-Bridge
Pre- and post-processing pipelines	Seismic trace processing pipelines	Linking seismic data and metadata	Gateway deployed on clouds
Mesh and model uploads and catalogues	Pipeline editing and creation	dispel4py on local machines, HPC and specialised clusters and cloud	Computations on HPC and institutional clusters
Integration of moment tensor sources	Production and capture of metadata	Provenance handling and exploitation	Security requirements compliance
Production and capture of metadata	Data ingest tools	iRODS and MongoDB for metadata	Scientific workflows
		GridFTP for bulk data transport	Provenance management

silico experiment to a more flexible and comprehensive way of supporting users during the data preparation and result evaluation, where data and methodologies are merged into a single exploratory space.

Context specific helpers improve the uptake of such tools. For instance, the current search interface over the experiments, adopting metadata relevant to the community's context, would in future make use of a tag cloud of terms. Their relevance can be weighted according to the user's scope, without hiding foreign metadata vocabularies which might have been used to describe results obtained by different applications, using the same workflow components. This contextual information can be extrapolated by the analysis of provenance data, particularly with the dynamic participation of users, which should enable consistent and effective acquisition of the information which they seek.

These adaptive and general interfaces, aimed at exploring the computational processes, make it easier for researchers to exploit their results. For instance, combining exploratory actions with data-shipment to other facilities, exposes them to a much bigger space of data services; still within the framework of a federated authentication infrastructure. This is achieved by the automated generation of helpers that handle security and transfer protocols. Such provenance-empowered 'awareness tools', provide users with the opportunity to learn about the computational process behind the tool they use; even when processes fail they yield rapid feedback. This does not detract from their primary focus and the principle uses of results.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

A key and sustained asset of VERCE has been very close collaboration between geoscientists and computer scientists. This was a crucial requirement for success; without it the task forces would not have worked. But it is also a key long-term goal, VERCE has to deliver a VRE where a growing number of seismologists and solid-Earth scientists will continue to advance their science and increase the power of their platform for doing their science. This has required that the payoff to seismologists from students to professors had to be manifest, but there also had to be benefits for the teams of ICT experts who run the dedicated and shared resources

seismologists need. Their payoff, is systems that interwork without their intervention, and templates, workflows and tools that automate many of their recurrent tasks. VERCE has achieved this positive feedback loop and demonstrated it for its computational and data-intensive communities.

This has been achieved by making progress on all four research tracks identified by Mattmann as critically important for future data science, based on many years of experience at NASA and at the Apache Software Foundation [35].

- *Rapid scientific algorithm integration:* The platform can easily support new simulation codes, thanks to a flexible design of configuration schemas and interactive components, which evolve around interoperable formats and functional abstractions. In that respect, the next target will be SPECFEM3D_GLOBE[29] as an alternative simulation code to improve support for global-scale waveform propagation studies. In the data-intensive contexts, new processing algorithms can be quickly put in practice, thanks to the portability of dispel4py which allows for their flexible evaluation across environments. We will improve the integration of the platform with registry and software packaging mechanisms, to better support the interactive creation of data-intensive workflows, which can be then deployed and executed under the full control of users, including provenance recording and metadata customisation.

- *Intelligent data movement:* This is key to meeting operational constraints and to achieving performance. We automatically handle getting the data to the right place at the right time, releasing resources in a timely manner while automatically collecting metadata and provenance records, so users can find and navigate their data. Performance includes user requirements for immediate feedback and control of an experiment. We can instruct the platform to produce and selectively make available intermediate results at runtime, fostering the rapid reaction of humans, as well as software agents external to the DCI. This is possible as we address one of the most important limitations Marrmann identified—the lack of effective metadata management. We do not considered HPC jobs as closed systems, but

---

[29]geodynamics.org/cig/software/specfem3d_globe/

rather active entities which can trigger behaviours and state changes in external services while they run. This results from combining runtime provenance analysis with selective data movements. We will develop more dynamism and semantics into this framework, in order to exploit diverse and reactive agents.

- *Use of Cloud Computing:* Clouds are used in VERCE for the operation and sustainability of the platform itself, rather then for providing elastic resources and software stacks typical in *Cloud Marketplaces* that provide virtual appliances. We use VM deployments where we can with mappings onto advanced data-intensive technologies. This provides the system administration gains even where resource allocation takes different forms. Many HPC providers are evaluating the impact of virtualisation on their architectures. A key achievement delivered is to recognise both the potential and drawbacks of different virtualisation models and to allow research activities to move fluently between them.

- *Harnessing the power of open source in software development for science:* VERCE itself is a demanding user of open source projects, such as gUSE, iRODS, Liferay. Several of its developments are becoming open source projects as a policy for the sustainability of its subsystems, see IV. *e.g.* the data-intensive framework dispel4py, will be packaged to reach a wider audience of users and developers.

These technical directions must be aligned with organisational strategies. The VERCE community will continue to grow and investment in training for established and new seismologists to alert them to its potential will need commensurate growth. Seismology continuously strives to refine the resolution of local, regional and global Earth models, as current resolutions are very often one-or-two orders of magnitude short of imaging the fundamental geological processes. Taking advantage of rapid advances in geodesy and geodynamic modelling is another challenge. The breadth of solid-Earth sciences will be extended with an increase in the diversity of data, simulations, and interconnected resources to enable this, in part in the context of EPOS. To achieve such extensions will require resources, but far less than starting again in each domain and merging later. There is a challenge to raise the *activation energy* to make that transition, which requires clarity about the benefits. This paper offers a glimpse of much future potential.

Lessons learned:

1) There are many resources: computational, data and working methods formalised as code and workflows, that have many technical and organisational variations that will continue because of education, investment and commitments. Providing a means for researchers to use these in combination without encountering the underlying complexity opens up new research avenues and improves their productivity dramatically. It also increases the opportunity for experts in different sub-

domains to pool their insights and skills.

2) It is difficult to find organisational and technical paths to reach that goal. Establishing short-lived, agile multi-disciplinary task forces to address specific issues is key to developing the depth of understanding and commitment to make progress.

3) The orders of magnitude gain in productivity have to be balanced with flexibility to enable researchers to create, refine and drive whatever innovative approach they choose. The challenges in each domain are so great that flexibility must be retained and exploited through multiple methods.

4) The independent paths to improvement of each contributing organisation, sub-discipline and technology are necessary for sustained vitality. Therefore, their composition must depend on mappings that are easily revised.

5) *Jam tomorrow* is not enough. Vision and strategic thinking are needed to achieve convergence, but immediate paths from current contexts to working examples with evident payoffs are necessary to maintain engagement, avoid unacceptable disruption and refine the target vision.

Open issues:

1) It is not possible to completely hide different usage policies such as identification, authentication, authorisation and accounting, because light-weight mechanisms to help large numbers of students and citizen scientists don't offer the trustworthy protection needed for major facilities used by moderate numbers of specialists. This balance between convenience and verified security requires treatment for many application domains as they embrace multi-disciplinary, multi-orgainisational and multi-national collaboration. However, this fundamental issue is exacerbated by a legacy of diverse IAAA mechanisms in use. The full breadth of institutions involved must at least circumvent these legacy impediments to collaboration.

2) The sustained delivery of e-Infrastructures is needed to warrant researchers becoming dependent on them. This will involve the maintenance of many software elements, some of which are introduced above. Just to keep them running with no enhancements, software has to be revised to meet changes in its operational context. Normally there are also pressures to address bugs and to introduce new features. Most components should be widely used and open source, so that many communities can contribute to necessary maintenance. As far as possible, the e-Infrastructure should use such components and replace them if they do not have a thriving community. This leaves three forms of software maintenance that every e-Infrastructure community needs to be responsible for: *a*) their fair contribution to the multi-community software elements, *b*) the mappings to and integration between the common

software elements to meet their specific needs, and *c*) on hopefully the rare occasions when a major element needs to be replaced by a thriving alternative, the integration of that. Today this maintenance investment is only available for novelty items and recognised simulation codes. Many other software elements need maintenance for the investment in e-Infrastructure to survive and for the improved research environment to be sustained.

3) The tools and methods for setting up and maintaining flexible mappings need to be improved substantially.

REFERENCES

[1] P. S. Earle and *et al*., "Prompt Assessment of Global Earthquakes for Response (PAGER)," USGS, Tech. Rep., 2009.

[2] A. M. Dziewonski and *et al*., "Determination of earthquake source parameters from waveform data for studies of global and regional seismicity," *J. Geophys. Res.*, vol. 86, pp. 2825–2852, 1981.

[3] G. Ekström and *et al*., "The global CMT project 2004-2010: Centroid-moment tensors for 13,017 earthquakes," *Phys. Earth Planet. Inter.*, vol. 200-201, pp. 1–9, 2012.

[4] E. Casarotti and *et al*., "CUBIT and seismic wave propagation based upon the spectral-element method: An advanced unstructured mesher for complex 3D geological media," in *Proc. 16th Int. Meshing Roundtable*, vol. 5B.4, 2008, pp. 579–597.

[5] C. Tape and *et al*., "Adjoint tomography of the southern California crust," *Science*, vol. 325, pp. 988–992, 2009.

[6] A. Fichtner and *et al*., "Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods," *Geophys. J. Int.*, vol. 179, pp. 1703–1725, 2009.

[7] H. Zhu and *et al*., "Structure of the european upper mantle revealed by adjoint tomography," *Nature Geoscience*, vol. 5, pp. 493–498, 2012.

[8] L. Colli and *et al*., "Full waveform tomography of the upper mantle in the South Atlantic region: Imaging a westward fluxing shallow asthenosphere?" *Tectonophysics*, vol. 604, no. C, pp. 26–40, 2013.

[9] D. Peter and *et al*., "Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes," *Geophys. J. Int.*, vol. 186, pp. 721–739, 2011.

[10] A. Ringler and *et al*., "The data quality analyzer: A quality control program for seismic data," *Comp. & Geosci.*, vol. 76, pp. 96–111, 2015.

[11] T. Megies and *et al*., "ObsPy—What can it do for data centers and observatories?" *Annals of Geophys.*, vol. 54, no. 1, pp. 47–58, 2011.

[12] R. Filgueira and *et al*., "dispel4py: An agile framework for data-intensive escience," in *Proc. IEEE eScience*, 2015.

[13] D. Latorre and *et al*., "Man-induced low-frequency seismic events in Italy," *Geophys. Res. Lett.*, vol. 41, no. 23, pp. 8261–8268, 2014.

[14] P. Kacsuk, Ed., *Science Gateways for Distributed Computing Infrastructures: Development framework and exploitation by scientific user communities*.   Springer, 2014.

[15] P. Kacsuk and *et al*., "WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities," *J. Grid Comput.*, vol. 10, no. 4, pp. 601–630, 2012.

[16] L. Scognamiglio and *et al*., "Real-time determination of seismic moment tensor for the Italian region," *Bull. Seis. Soc. of America*, vol. 99, pp. 2223–2242, 2009.

[17] M. Kozlovszky and *et al*., "DCI Bridge: Executing WS-PGRADE Workflows in Distributed Computing Infrastructures," in *[14]*, P. Kacsuk, Ed.   Springer, 2014, ch. 4, pp. 51–67.

[18] S. WG, "Security Assertion Markup Language (SAML) V2.0 Technical Overview," OASIS, Tech. Rep., 2008.

[19] K. Benedyczak and *et al*., "Key aspects of the UNICORE 6 security model," *FGCS*, vol. 27, no. 2, pp. 195–201, 2011.

[20] D.-E. Smørgrav, "Pluggable Authentication Modules," Networks Associates Technology, Inc., Tech. Rep., 2003.

[21] R. Filguiera and *et al*., "dispel4py:A Python Framework for Data-Int. Scientific Computing," *IJHPCA*, In press 2015.

[22] R. Filgueira and *et al*., "Applying selectively parallel I/O compression to parallel storage systems," in *Euro-Par*, ser. LNCS.   Springer, 2014, vol. 8632, pp. 282–293.

[23] MPI Forum, "MPI: A message-passing interface standard," *IJ of Supercomputer Applications*, vol. 8, pp. 165–414, 1994.

[24] C. Walter, "Kryder's Law: The doubling of processor speed every 18 months is a snail's pace compared with rising hard-disk capacity, and Mark Kryder plans to squeeze in even more bits," *Scientific American*, pp. 32–33, 2005.

[25] E. Givelberg and *et al*., "An architecture for a data-intensive computer," in *Proc. NDM '11*.   ACM, 2011, pp. 57–64.

---

[30]www.orfeus-eu.org

[26] A. Hajnal and *et al.*, "Remote storage resource management in WS-PGRADE/gUSE," in *[14]*, P. Kacsuk, Ed. Springer, 2014, ch. 5, pp. 69–81.

[27] M. Conway and *et al.*, "Demo. of Policy-Guided Data Preservation Using iRODS," in *POLICY*. IEEE, 2011, pp. 173–174.

[28] *POLICY 2011, Int. Symp. on Policies for Distributed Systems and Networks*. IEEE, 2011.

[29] EU Parliament, "Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)," *Official J. EU*, vol. 50, no. L108, 2007.

[30] W. D. Pauw and *et al.*, "Visual debugging for stream processing applications," *Runtime Verification*, pp. 18–35, 2010.

[31] W3C, "PROV Model Primer," www.w3.org/TR/prov-primer.

[32] I. Foster, "The virtual data grid: a new model and arch. for data-intensive collab." in *SSDBM*. IEEE, 2003, pp. 11–11.

[33] S. Gesing and *et al.*, "Workflows in a Dashboard: A New Generation of Usability," in *Proc. WORKS '14*. IEEE, 2014, pp. 82–93.

[34] H. Igel, "Computational seismology progress," 2015, personal communication.

[35] C. A. Mattmann, "Cultivating a research agenda for data science," *Journal of Big Data*, vol. 1, 2014.