eScience gateway stimulating collaboration in rock physics and volcanology

eScience: Innovative platform to promote persistent collaboration reesearch in Rock Physics and Volcanology

Rosa Filgueira and Malcolm Atkinson
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB, UK
{rfilguei,mpa}@inf.ed.ac.uk

Andrew Bell and Ian Main
School of GeoSciences
University of Edinburgh
Edinburgh EH8 9XP, UK
{a.bell, ian.main}@ed.ac.uk

Steve Boon and Christopher Kilburn
and Philip Meredith
Department of Earth Sciences
University College London
London WC1E 6BT, UK
{s.boon, c.kilburn,
p.meredith}@ucl.ac.uk

***Abstract*—Earth scientist observe many facets of the planet's crust and seek to integrate their resulting data to better understand the processes at work. We report on a new data-intensive science gateway designed to bring rock physicists and volcanologists into a collaborative framework that will enable then to accelerate their research and integrate well with other Earth scientist. The science gateway support three major functions:**

1) **sharing data from laboratories and observatories, experimental facilities and computational model runs;**
2) **sharing computational models and methods for analysing their experimental data; and**
3) **supporting recurrent operational tasks, such as data collection and code application in real time.**

Our prototype gateway has worked with two demanding exemplar projects giving experience of data gathering, model sharing and data analysis. The participants in those two projects found that the gateway accelerated their work, enabled new practices and formed a good platform for long-term collaboration.

***Keywords*-Science Gateway, High Performance Computing, Data sharing, Code Testing, Metadata and Storage, Earth sciences.**

I. Introduction

Scientists have always shared data and mathematical methods, recently in the form of code, pertinent to the phenomena they study. Over the last two decades rock physics and Volcanology, as well as other solid-Earth sciences, have increasingly used Internet communications for this purpose. Typically this has been on an *ad hoc* basis, for specific projects and campaigns[1].

Here we consider how to organise data from rock physics experiments and volcano monitoring to open up opportunities for sharing and comparing data, observations and model runs, and analytical interpretation methods. We focus on these discipline areas because of our experience in recent research projects, shared data types and analytical methodologies, and a growing overlap of research in the two communities. The two areas also provide informative contrasts in factors such as the nature of the datasets (duration, data rates, volumes), the level of existing collaboration and cohesion, and the needs of the community. We hypothesise that if we facilitate productive information sharing across those communities via a new science gateway, that will benefit the science. Refinement of methods and data will be more rapid through sharing and advances will be facilitated by easier access to data and computational tools. We present the functionalities of the proposed science gateway for rock physics and volcanology (*RP & V*) and its prototype implementation.

This will benefit from and stimulate concomitant progress towards adopted data standards. We feel the rock physics community would be stronger from adopting consistent standards for data interchange and for metadata describing instruments, experiments and so on. This is ineluctably a slow and continuous process. Once agreements are reached progress towards their adoption is invariably incremental, as it is easier to adopt them when new experiments and projects are set up. Until their adoption, the proposed science gateway has to balance between tolerating diversity (always necessary in research) and introducing translators to migrate towards the agreed common targets. It can also provoke progress in these areas by judicious introduction of modest initial examples, designed as a core around which standards can grow.

This paper is triggered by our collaborative work with the exemplar projects EFFORT[2] and CREEP-2 [3], which provided two forms of data, real-time and archived, as well as initial users and uses cases.

The paper is organised as follows. Section II summarises the context of the work. Section III presents the functionalities of the required science gateway. Section IV introduces our experience with the prototype. Sections V to VII explain

[1] A notable exception is the global collaboration on sharing seismic data—The International Federation of Digital Seismograph Networks (FDSN) http://www.fdsn.org.

[2] http://www.effort.is.ed.ac.uk

[3] http://www.ucl.ac.uk/es/research/ripl/research/rock-physics

each of the functionalities, their requirements and status. We conclude with a summary and proposed future work.

II. Context for data and code sharing

The present context is complex, with a growing wealth of digital data and pressing global challenges. For example, there are many groups working contemporaneously on potentially relevant and rapidly evolving standards, and on technology that is itself rapidly evolving. We cannot provide a comprehensive survey. Rather, we illustrate with selected examples.

1) The Global and Regional Frameworks for Agreeing Standards: Traditionally sharing depends on agreeing standards, choices of units, data interchange protocols, the governance of data use, persistent identifiers (PIDs), and *ontologies* (controlled vocabularies) for describing concepts. The challenge here is to develop intercommunication without inhibiting innovation. A few examples follow.

- *The Committee on Data for Science and Technology*: (CODATA)[4], is an interdisciplinary Scientific Committee of the International Council for Science Unions (ICSU).
- *The Global Earth Observation System of Systems*: (GEOSS)[5] is a effort to bring together Earth observation systems, facilitating sharing of data and information at minimal cost.
- *The Infrastructure for Spatial Information in the European Community*: (INSPIRE)[6] aims to align Spatial information in all government and administrative systems across Europe. It mandates many international standards.
- *Intergovernmental Panel on Climate Change*: (IPCC)[7] is a scientific body, set up by the United National Environment Program. It has aligned climate data representations and procedures, leading to five definitive reports.
- *Research Data Alliance*: (RDA) builds social and technical bridges to enable data sharing, internationally and across disciplines.

These large-scale standardisation efforts take a long time to take effect and scientists need to find niches of stability that meet their requirements. Two factors drive and limit change:

- Most scientists remain focused on their own goals and their work should not be disrupted.
- Many scientists spend 80% of their time data wrangling, and automation should reduce this distraction.

2) Large-Scale Projects and Initiatives: The practical development and adoption of sharing arrangements depends on leadership and investment to overcome the hurdles that individuals, countries and organisations face. These are often achieved through collaborative projects. We enumerate some of them:

- *European Plate Observing System*: (EPOS)[8] is the integrated solid Earth Sciences research infrastructure approved by the European Strategy Forum on Research Infrastructures (ESFRI). It includes rock physics, volcanology and seismology.
- *Virtual Earthquake and Seismology Research Community in Europe*: (VERCE)[9] aims to develop a data-intensive e-science environment to enable innovative data analysis and coding methods that fully exploit the wealth of data generated by the global seismology community.
- *Common Operations of Environmental Research Infrastructures*: (ENVRI) project[10] is a collaboration of the ESFRI Environment Cluster, with support from ICT experts, to develop common e-science components and services for their facilities. Their reference model based on an industry standard (ODP) will improve the design and construction of distributed ICT infrastructures for research and increase sharing [1].
- *Supersites*: It[11] is an initiative of the geohazard scientific community. The Supersites have data for the study of natural hazards in geologically active regions, including information from Synthetic Aperture Radar (SAR), GPS crustal deformation measurements, and earthquakes.

3) Examples from other fields: Sharing workflows and code among researchers is a powerful way of sharing expertise, so that researchers can reuse and repurpose research techniques within and across domains. There are several projects that work to develop those sharing opportunities, for example:

- *myExperiment*: The myExperiment science gateway is a collaborative environment where scientists can safely publish their workflows and experiments, share them with groups and find those of other's [2].
- *WF4ever*: The WF4ever project is pioneering strategies which aid preservation of scientific objects, such as experiments defined using a workflow notation [3].
- *The integrated Rule-Oriented Data-management System*: (iRODS), is a community-driven, open source, data grid software solution [4] to help researchers, archivists and others manage (organise, share, protect, and preserve) large sets of data.
- *SCI-BUS*: The SCI-BUS project [5] has been established to support the developers of science gateways. It uses ER-Flow to handle multiple workflow systems and has a component *DCI-Bridge* that can be used independently to ship jobs to a wide variety of computational infrastructures handling data movement and

[4]http://www.codata.org/
[5]http://www.epa.gov/geoss
[6]http://inspire.jrc.ec.europa.eu/
[7]http://www.ipcc.ch/
[8]http://www.epos-eu.org/
[9]www.verce.eu
[10]https://www.egi.eu/community/projects/ENVRI.html
[11]http://supersites.earthobservations.org/

access controls automatically.

- *EUDAT*: The EUDAT project incorporates data movement in conjunction with data replication and data staging services[12]. EUDAT manages interaction with the data transportation, computational and storage services This reduces significantly the effort and cost of setting up *and running* such facilities.

III. Rock Physics and Volcanology Science Gateway functionalities

The purpose of a science gateway is to provide an intellectual and technical focus, so that collaboration and sharing are the overall goal of the science gateway:

> *To make existing collaborative research practices much easier, faster and more efficient, and to enable and stimulate new research.*

These benefits are achieved by science gateways through:

- sustained operation allows that scientists to become expert and to depend on their facilities and technical and organisational improvements to accumulate;
- co-location integration and archiving of a collection of data, code, tools and interfaces, in one logical framework to make it easier to apply methods that combine the facilities; and
- the focus provided stimulates communication among researchers and provides a forum for locally and pragmatically agreeing interchange, access and attribution standards.

These benefits only become manifest if the relevant researchers engage with the science gateway in sufficient numbers. The *raison d'être* of the prototype science gateway is to ascertain whether it attracts them in sufficient numbers. To do this it has to reach a minimum level of functionality and content. The prototype rock physics and volcanology science gateway will need the following functionality:

1) A well-organised and easily accessed store of contributed and sharable data:
   a) live data from on going experiments and monitoring.
   b) archival collections of data suitable for long-term studies and benchmarking, and
   c) model data, e.g., from simulations and test cases.
2) A shared collection of methods and code that encourages data processing and modelling:
   a) code to conduct analyses, e.g., to forecast brittle rock failure,
   b) simulation code, e.g., to synthesise test data or model physical phenomena such as acoustic emissions and strain, and
   c) methods, represented as scripts or workflows, that capture effective ways of using the science gateway's resources, e.g., to run an analysis periodically and to visualise its results together with the analysed data.
3) Convenient support for recurrent tasks, including:
   a) automated and reliable data upload (from running experiments, monitoring),
   b) periodic activation of codes and methods, e.g., applying a list of analyses to a stream of uploaded data every 15 minutes, and
   c) triggered activation of codes and methods, e.g., when a periodic analysis detects energy above a specified threshold, send messages to a list of researchers and start more intensive analyses.

IV. EFFORT: Earthquake and Failure Forecasting in Real Time

This project has built a prototype of the gateway. Its a multidisciplinary collaboration between geoscientists (School of GeoSciences, University of Edinburgh), rock physicists (Department of Earth Sciences, University College London (UCL)), and informaticians (School of Informatics, University of Edinburgh). Brittle rock failure plays a significant role in the timing of a range of geophysical hazards, such as volcanic eruptions; Yet the predictability of brittle failure is unknown. EFFORT aims is to provide a facility for developing and testing codes to forecast brittle failure for experimental and natural data. The code is tested in real-time, *verifiably prospective mode*, to avoid selection biases that are possible in retrospective analyses. This requires rapid data assimilation and continuous uploading of forecast and uncertaintis.

The project will quantify the predictability of brittle failure, and how this predictability scales from simple, controlled laboratory conditions to the complex, uncontrolled real world. In EFFORT we collect experimental data from controlled experiments:rock physics experiments from the UCL laboratory, sub-sea deformation experiments in the CREEP-2 project [6], and volcanic monitoring data from INGV observatory Etna[13] and IGN observatory Hierro[14]. We also generate synthetic rock physics data using python programs to test the algorithms. The tasks of the projects can be presented as:

- *UCL Rock Physics*: To run triaxial rock deformation laboratory experiments, and CREEP-2 experiments.
- *Edinburgh Informaticians*: To design and to implement *EFFORT gateway*.
- *Edinburgh Geosciences*: To applicate forecasting methods by using the UCL experiments and observatory data.

Figure 1 shows how the experimental data from observatories and laboratories are transferred to the *EFFORT gateway*. Geoscientists are able to view the data, contribute codes, analyse data or run models.

[12] http://www.eudat.eu/safe-replication and http://www.eudat.eu/data-staging respectively.

[13] http://www.ct.ingv.it/en/

[14] http://www.02.ign.es/ign/main/index.do

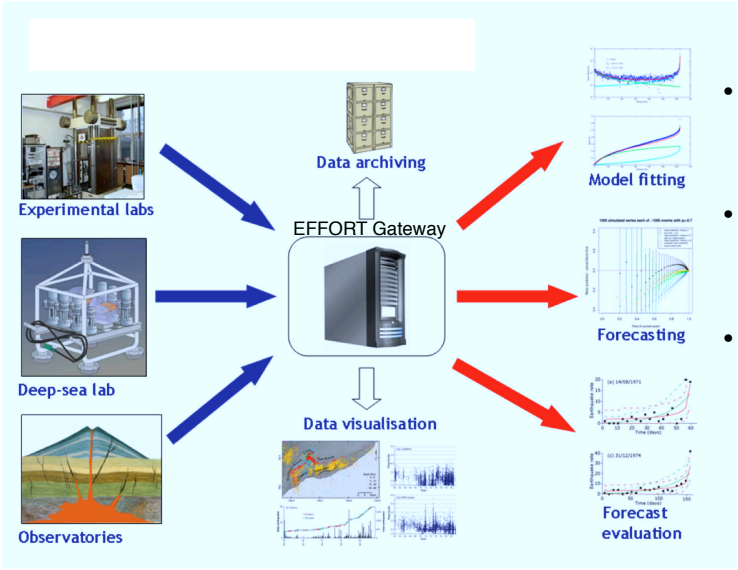

Figure 1. Overview of EFFORT project.

The EFFORT gateway is a Liferay 6.1[15] web portal with several Rapid portlets [7] for accessing data and code, generating synthetic data, running code and displaying results. The EFFORT gateway is located in a VMware virtual machine provided by University of Edinburgh. The virtual machine runs Debian/Linux 7 as O.S and it has 8GB of memory. It has an NAS repository mounted to reliably store all data.

V. Sharing Data

The gateway provides a shared repository where rock physics and volcanic monitoring data from different laboratories, observatories, experiments and syntheses can be uploaded easily. An initial choice of data and metadata conventions will seed revision and extension by geoscientists. How this interworks with other disciplines, institutional and national repositories development require policies for data and code publication and access control.

A. Design and motivation

The following data-management facilities are needed:

- *Upload data*: Mechanisms to reliably transfer multi-channel streaming data and metadata periodically from experiments and observatories. These transfer mechanisms must be easy to set up at the data-source site and require little operational oversight.
- *Download*: Download of selected data to researchers' local resources for any permitted purpose.
- *Computation*: It should be possible to arrange for computation against the cost of data held in the science gateway. This avoids potentially expensive data transfers to run on a researcher's own computers, but this has to be balanced against computational provision underpinning the science gateway.
- *Visualisation*: The science gateway provides researchers with controlled visualisation of selected data. This allows researchers to monitor the progress of experiments that are remotely located.
- *Metadata and catalogues*: As commonalities are recognised by the rock physics community they will adopt consistent standards for data interchange and for metadata describing data, instruments, and experiments.
- *Working storage*: Researchers benefit if they have an allocation of personal storage arranged close to the rock physics gateway. They frequently need to store intermediate results, their selections of data and their derived data being prepared for publication. This personal storage improves productivity. Generally, researchers wish to return to it repeatedly, like a "*persistent shopping basket*", access it from multiple locations and trust in its privacy.
- *Archiving*: Once researchers have deposited data in the rock physics science gateway, they would like to rely on its preservation for agreed periods with understood and specified risks of loss. This should also ensure the long-term interpretation of persistent identifiers (PID) used to refer to data in publications previous.
- *Citation and attribution*: For its continued support the science gateway needs citation and attribution in papers, so that others learn about it and its utility is recognised.

B. Implementation and Experience

Figure 2 shows the structure of the *EFFORT gateway*[16]. Currently rock-loading experiments are being conducted in the UCL rock physics laboratory and the data are being streamed to the prototype gateway. This gateway also contains uploaded archive data from previous rock loading experiments and some modest accumulations of volcanic observatory data. It is being prepared for handling long-running rock-loading experiment data streaming from a cage of experiments 2km below the surface of the Mediterranean in the CREEP-2 project. Most of the experiments that we collect contain time-series data that belong to one of two different classes: *Event catalogue data* (ECD) and *Sampled continuous data* (SCD). Volcanic observatories and rock physics laboratories can produce data of both classes in a single experiment. ECD consist of a series of events (e.g. acoustic emissions, earthquakes, volcanic eruptions) that occur at discrete times and have a specific attributes (e.g. location, depth, magnitude, duration, ...). SCD consist of a series of times at which a continuous variable has been measured, and the value of that variable. The sample times are defined by the instruments' operator, and may (or may not) be evenly spaced (e.g. daily, every second). Consequently the ECD and SCD from the rock physics

[15]http://www.liferay.com

[16]www.effort.is.ed.ac.uk

laboratories are different from volcanology observatories. We represent those datasets using four different structures.

- *Event catalogue laboratory data* (ECLD)
- *Event catalogue volcanology data* (ECVD)
- *Sample continuous laboratory data* (SCLD)
- *Sample continuous volcanology data* (SCVD)

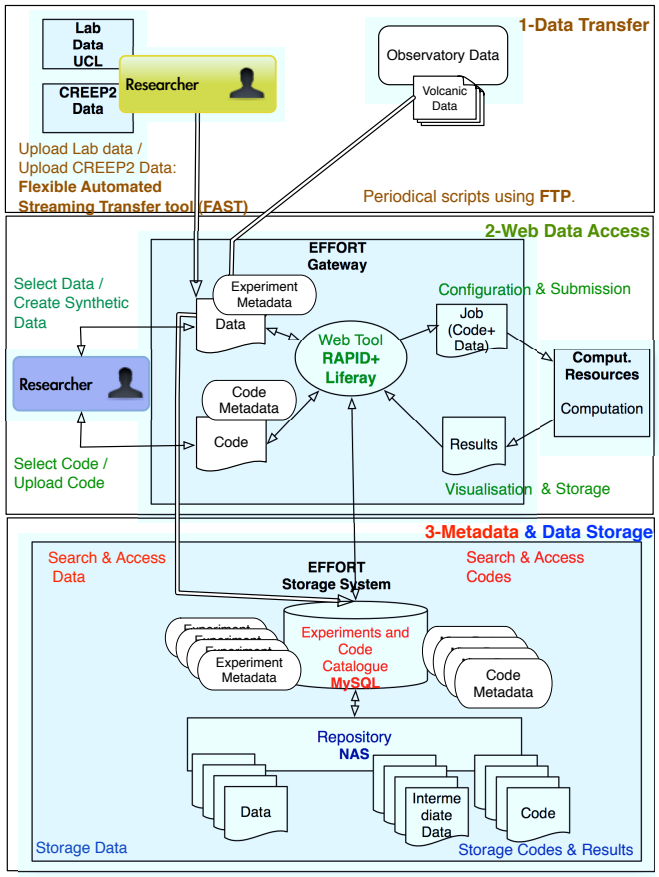

Figure 2. Structure of the EFFORT gateway.

In the *EFFORT gateway*, we are working on several facilities for sharing data :

- *Upload data*: We have designed and developed a new adaptive data transfer java tool called *FAST* (Flexible Automated Streaming Transfer) to upload experimental data and metadata periodically in real time from the UCL Laboratory to our repository–see Figure 2. *FAST* selects an appropriate protocol spending on the type of data transferred. *FAST* is easy to set up and requires little oversight during operation. It copes automatically with interruptions to local operations and communications. It is compatible with all operating systems, and suitable for use in other rock physics laboratories.The deposited experimental data (composed of one or more datasets) are automatically assigned a unique identifier in the *repository*. *FAST* also generates minimum metadata for the *catalogue* that appears as soon as a researcher has initiated the data transfer. Bulk upload of previous result data is available via *FAST*. On the other hand, to obtain volcanology data we have automated download from the observatories' sites by using periodic scripts activating an *FTP* protocol [8].
- *Access data*: The *gateway* allows researchers to search the metadata using criteria specific to rock physics and volcanology via wrapped *sql-queries*. The gateway has implementing these searches and presenting their results by using *jython* plugins with different *sql-queries*. Those queries are executed in the virtual machine against the *catalogue*, and the results are presented dynamically to the users in the gateway. Metadata browsing is supported allowing the data about data to be presented and accessed via the Web portal. Both data and metadata will be made computationally accessible via a RESTful Web service. Figure 3 displays an example about how a researcher visualizes the metadata of experimental data and selects it. The gateway allows a user to filter the selected data, (see Figure 3), by choosing an interval of data and the Completeness Magnitude threshold.
- *Visualisation*: As data are deposited in the repository, a visualisation of the accumulated data is made available for display in the Web portal – see Figure 4.
- *Metadata and catalogues*: The *gateway* uses a *repository* to hold all of the data and code, and a *catalogue* (MYSLQ database [9]) to hold all the corresponding metadata as shown in Figure 2. Figure 5 represents the Entity Relationship (ER) diagram of the *catalogue* representing the structure of experimental data.
- *Data Traffic*: In the current *CREEP-2 experiment*, every day a new text file (SCLD-type) is generated for storing strain, stress and porosity of the current rock sample. New data are appended to this file every minute. As soon as Acoustic Emissions (AE)[17] start, a new text file (ECLD-type) containing the number of AE events and their energy, peak amplitude, duration, load, displacement and pore volume is generated. In this case, data are appended to this file every microsecond. The first type of file is synchronised with their copies in our *repository* every minute from the beginning to the end of the day. The AE files are also synchronised with their copies in the *repository* every minute, but only during periods when AE are being generated. All the transfer/synchronization and cataloguing operations of the CREEP-2 experiment are made automatically by *FAST*. Regarding the volcanic experiments, INGV Observatory Etna and IGN Observatory Hierro, publish volcano catalogues with several files (ECVD-type) on their websites. Those files are downloaded daily from

[17]AE is the phenomenon of emitting elastic waves as a result of irreversible or partially reversible changes in the structure of a solid under the action of various external and internal physical factors. The duration of the AE could be short or long. It is a phenomenon that can not be predicted or controlled. Several AE periods can be recorded for the same experiment.

the websites and stored in the *repository*, and metadata are automatically inserted in the *catalogue*. In the future, we expect additional datasets from Etna; Other earthquake catalogues (ECVD-type), and strain or gas measurements (SCVD-type). Finally, the EFFORT gateway allows the generation of synthetic datasets (SCLD and ECVD types) which are stored in the *repository* with the corresponding metadata in the *catalogue*. The metadata includes which researcher has started the experiment, team and organisation is a member the researchers to help with future searches. One experiment can be composed of several sub-experiments, as shown in Figure 5, e.g a CREEP-2 experiment can be composed of several rock samples. Any element in the *repository* is represented in the *catalogue* by an entry in the *Repository_Item* table.

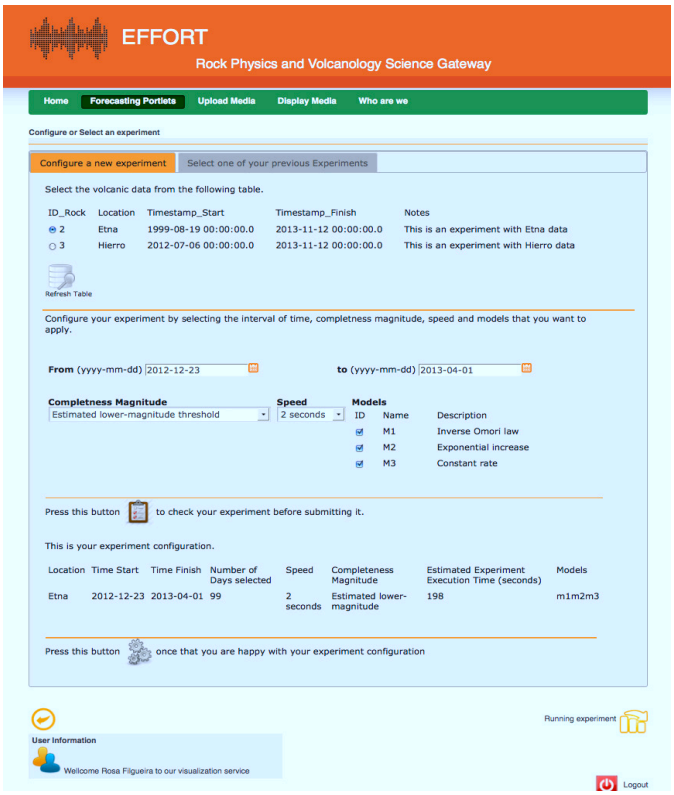

Figure 3. Selecting, filtering and checking a job before submitting it.

VI. Sharing Codes and Methods

A shared collection of code for rock fracture prediction can allow code builders to compare predictive performance, and to refine each other's implementations. In the same way as for sharing data, we propose one shared repository where rock physics codes and methods can be easily stored with consistent control policies and attribution mechanisms. Metadata describing each code stored in the repository is needed for instance, to build a catalogue of codes that

Figure 4. Visualisation of Experimental data in EFFORT gateway.

allows codes to be discovered and used. We use the term "*code*" inclusively, to cover code used to model brittle-rock fracture and the physical phenomena encountered, code to prepare data arriving from experiments for analysis, code to perform analyses and code to synthesise test data. For the moment, methods, i.e. sequences of analytic operations, are also represented by such codes, but in the future they may be represented by scientific workflows [10]–[13]. For the present the only form of code handled is python scripts as this has satisfied the first users[18]. We describe the facilities that the proposed science gateway must provide to make this possible.

A. Design and motivation

- *Upload and store codes and methods*: The science gateway provides an easy mechanism to upload codes and methods with control policies. In each case the corresponding metadata describing the code is supplied or constructed.
- *Description of codes*: In order to promote the sharing of code among the rock physics community a good description of every method is required. This increases the number of users of each method, and allows researchers to build on each other's implementations, or progressively enhance their own. The metadata describing the code or method and the sharing controls are stored in

[18]When a multilingual repository of computational methods is needed a general purpose approach, such as myExperiment [14], and WF4ever should be embedded in the gateway

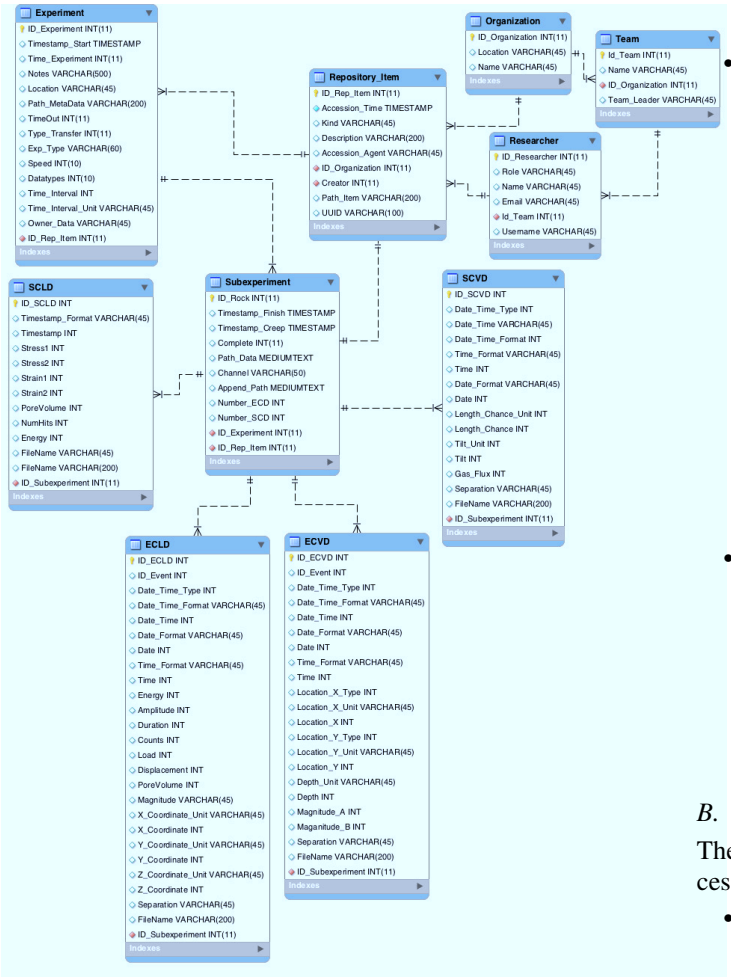

Figure 5. Structure of metadata about experimental data.

the catalogue.

- *Search for codes and methods*: Science gateway allows researchers to perform searches over the codes by using different queries such as: searches by task, by researcher, by experiment type or by result type. Using the result of each search, a researcher will be able to access directly codes or methods, or contact its originator, for example, to request permission to use or make a derivative of a code.
- *Access to codes*: The science gateway allows researchers with the correct rights to access code and methods. Once accessed they may be applied to appropriate experimental, synthetic or derivative data stored in the rock physics science gateway or downloaded for use on a researcher's own computers.
- *Services to test codes*: An easily used mechanism enabling researchers to run codes against selected data is offered with the science gateway mediating the choice of computational resources. Results can be visualised or saved using the mechanisms introduced in Section V. This will be extended, in due course, with services for *benchmarking and validating analytic code*.
- *Services to run codes*: These may be similar to the services for testing, but they may offer fewer arrangements for diagnosis and monitoring in exchange for improved ease of use, particularly for repeat runs with the same parameterisation. A researcher will be able to *set up* a run specifying parameters, input data and result handling. The result handling may include visualisation during and after the run, ingest of the results into the storage and catalogue systems and generation of associated metadata. The researcher will be able to *run* this prepared package easily and repeatedly, possibly supplying a smaller set of parameters or modifying some of the previous settings. The implementation may draw on workflow technology which may be exposed to the researchers to allow them to compose more sophisticated runs.
- *Citation and attribution*: To encouraging the contribution, use and sharing of codes and methods, the science gateway will support citation and attribution of the codes and methods used; this also facilitates accurate recording of the derivation of data for recalculation and scrutiny. It requires suitable metadata, effective provenance tracking and straightforward generation of persistent identifiers for use in citations.

B. Implementation and Experience

The science gateway provides facilities for uploading, accessing and testing codes:

- *Upload and store codes*: Researchers can upload as many codes to the *repository* as they want. As soon as a code has been uploaded, the researcher triggering the upload has the opportunity to designate it as shareable with all users, or to leave it in the default designation of "*only accessible to the person who performed the upload*". The choice of such defaults and the selectivity of the options available is a typical example of the kind of policy decision a science gateway governance process would consider. When the need arises, codes in other languages, such as C or Fortran, will be accommodated.
- *Description of codes*: The *gateway* solicits and creates metadata for every code uploaded. Figure 6 shows the ER diagram for representing the codes in the *catalogue*. This metadata is used by the *gateway* to validate and automate tasks requested. In the *catalogue* the codes are represented by the *Code* table, and the results of applying codes to data by the *Derivative* table. As we introduced before, any element stored in the *repository* is represented in the *catalogue* as an entry in the *Repository_Item* table. So, codes and results are also represented like this.
- *Search for codes*: Researchers can search the catalogue for codes by using pre-packaged *sql-queries* based on the criteria identified by rock physicists and volcanologists. In the current version the available searches are

by task and by researcher. As the user community and number of codes grow mechanisms for encouraging consistent use of (an agreed and prepared) vocabulary covering the established aspects of the codes will be needed. In Figure 3 we can see how the gateway has presented to the researcher the option to use three codes (called models in the Figure): M1, M2 and M3. In this case only the codes suitable for volcanology data are presented.

- *Access to codes*: Once a researcher has selected the code(s) that is going to be used for analysing an experiment, it can be obtained from the gateway, i.e., downloaded as a file. If the researcher is authorised it can be modified and uploaded again as a new version of the same code.
- *Services to test and run codes*: Once a researcher selects a code (as shown Figure 3) and the experimental data to which it should be applied, the EFFORT gateway submits the corresponding computational job to a high-performance computational (HPC) resource hiding technical details. This is possible thanks to Rapid technology, which enables the configuration of web portlets that connect to the computational resources. We define a computational job as a run of a code with a set of experimental data as an input parameter.
- *Varpy library*: A new open-source toolbox, called *Varpy*, has been designed [15]. It was inspired by the Obspy library [16], and provides a Python framework for analysing volcanology and rock-physics data. It provides several functions, which allows users to define their own workflows to develop models, analyses and visualisations. The library also makes it easier to arrange that code is in a form suitable for running in the *gateway*. Care has been taken to ensure that the library can also be used outside of the *gateway*, e.g., on a researcher's own laptop. The goal of the Varpy library is to accelerate the uptake of computational codes by researches in volcanology and rock physics. It does this via two mechanisms:
  1) supplying a library of ready-made functions that are generally useful; and
  2) providing a context for creating, sharing and comparing further functions.

  We anticipate two groups of readers and Varpy users:
  - the majority who use functions already written and in the library; they will predominantly arrange to use sequences of these functions with their own parameterisations; and
  - contributors, who, as well as using provided functions, also want to write additional functions for their own use or to add to the library.

  The computational resource that we have used is THE Edinburgh Compute and Data Facility (ECDF) cluster. ECDF consists of 130 IBM dx360M2 iDataPlex servers with two IntelWestmere E5620 quad core processors and 24 GB of RAM, all connected through Gigabit ethernet. Researchers can check the configuration of the job before submitting it (see Figure 3). We have started preliminary work using the Open Science Data Cloud (OSDC)[19] for submitting jobs.

  Once a job is submitted to the cluster, the results are displayed in the gateway in real time, as shown Figure 7, catalogued and stored in the data repository, allowing further researcher-instigated operations to retrieve, inspect and aggregate results. Usually, a job executes one or more codes on data that grows periodically. This means, that the codes are applied to the data several times. The *gateway* refreshes the visualisation of the results when the codes are reapplied to the data.

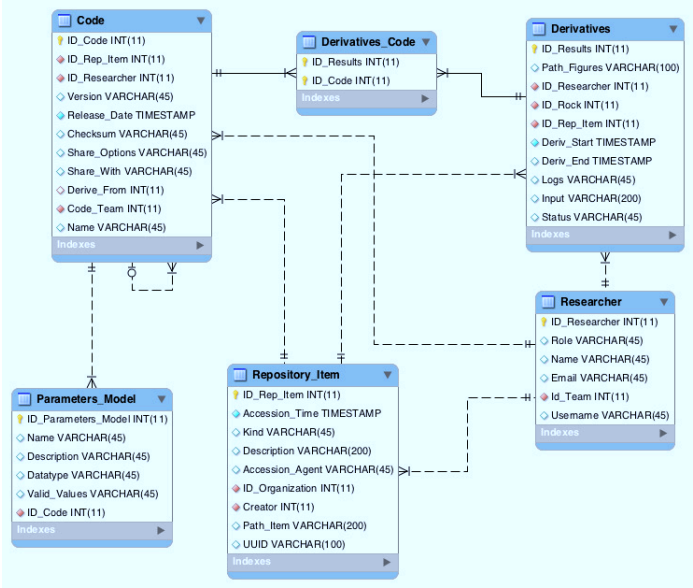

Figure 6. ER Diagram showing structure of metadata about codes.

VII. Recurrent Operations

Support for recurring tasks, such as data collection and code application to the data will expedite rock physics and volcanology research. Once the scientists (experimentalists or code builders) have set up their initial requirements about any recurrent operation, the science gateway will perform them periodically or when predefined conditions fire a *trigger*. This saves researchers time directly initiating repeated runs and permits quasi-continuous monitoring. As we defined in the Section VI, we use the term "*code*". The minimum recurrent operations are introduced below.

A. Design and motivation

- *Automated data upload*: The automated data upload mechanism (already introduced in Section V) is needed

[19]https://www.opensciencedatacloud.org/

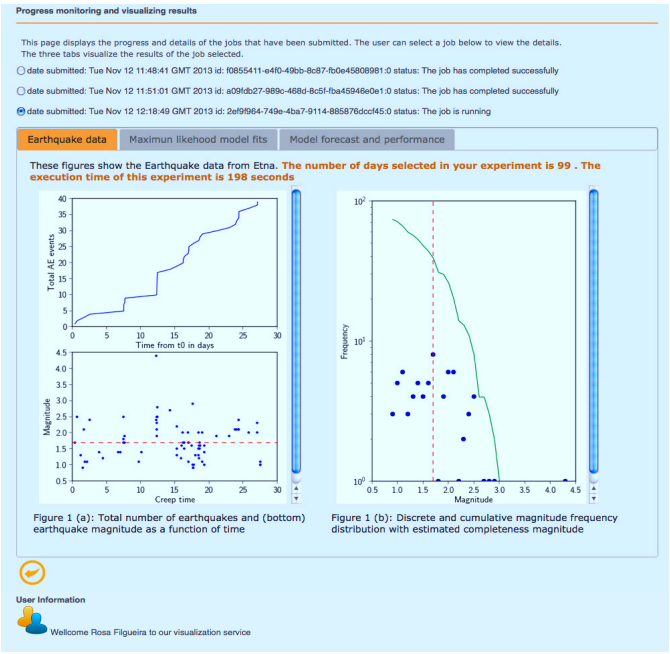

Figure 7. Displaying the results of a job in real time.

to obtain periodic increments of experimental or observational data. It allows researchers to specify the data source and which channels of data should be uploaded, and the frequency of collecting increments. It also permits researchers to specify *timeouts*, which are durations that cause a named trigger to be fired if no data has been received from the source (and channel) during the specified period. They may also set the criteria for recognising the end of an experiment and associate a trigger with that event.

- *Trigger management* A *trigger* is a *named* entity which can *fire* when a condition associated with the trigger becomes true. For example, a trigger called *midnight* might be set up, which fires every time it is midnight (00:00:00 UTC) (Coordinated Universal Time). There are two main classes of trigger, those concerned with time and those concerned with states of (incoming) data. In order to manage the lifetime of triggers, the researchers require the following five operations:
  1) *Creation:* That creates a new trigger that has a unique identifier and definition that specifies when that trigger fires.
  2) *Edit:* For a trigger that already exists.
  3) *Delete:* For an existing trigger the owner should be able to remove the trigger from the system.
  4) *List:* Obtain a list of the available triggers and their definitions. Optionally, it should be possible to list selectively, and to see the dependent codes that are set up to be run when a trigger fires.
  5) *Manipulate dependents:* This enables a researcher, to add, remove or re-order the set of codes that are run when a trigger fires.

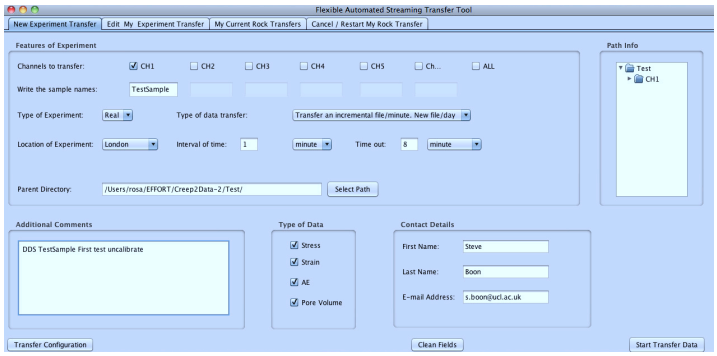

Figure 8. Automated data upload provided by *FAST*.

- *Activation of codes*: A code, analysis or standard action (such as sending an email or SMS message) can be associated with one or more triggers, so that it is applied automatically by the gateway, i.e., its associated code or workflow is run, whenever those triggers fire.

B. Implementation and Experience

We have started to introduce some of the described recurrent operations.

- *Automated data upload*: *FAST* provides a mechanism to automate the data upload, as shown Figure 8. Before starting to upload experimental data, the rock physics researcher provides the following information to *FAST*:
  - how many channels (rock samples) the research wants to transfer, and their names. (e.g Channel 1 named TestSample).
  - how often the researcher wants to upload datasets (or increments to datasets) to the repository,
  - how the datasets are to be written in local file(s) (create a new file or appended to an existing file),
  - the estimated experiment duration (hours, days, months, etc),
  - the details of the sub-experiment(s), experiment(s), equipment, laboratory generating the data, and the rock physics researcher who initiated the transfer.

  With all of this information, *FAST* is able to upload data periodically and automatically to the repository. If any problem happens during the data transfer, *FAST* sends a notification to the researcher who initiated the data transfer. In the CREEP-2 preparation experiment data was uploaded automatically from the UCL laboratory to the *repository* by using *FAST* for a total of 43 days.
- *Periodic activation of codes*: The EFFORT gateway allows researchers to run different codes periodically against the experimental data that are being or have been uploaded. This does not currently use the concept of triggers. They can set the interval between runs of their codes (by default codes are run every minute). This assumes that the execution time of the codes is less than the specified interval. Eventually, that will need to be verified by the gateway and their activation

suspended with notification to the relevant researcher(s) of the detected overload. New results for interval of time are displayed through the gateway.

VIII. Lessons learnt and FutureWork

In this paper we have presented the design of a new rock physic and volcanology science gateway based on Internet technologies to encourage the collaboration between different rock physics laboratories and volcanology observatories worldwide. The benefits are as follows:

- Promotes innovation and potential new *RP & V* data uses
- Provides fast testing and propagation of ideas
- Leads to new collaborations between *RP & V* data users and *RP & V* data creators
- Encourages improvement and validation of *RP & V* research methods
- Reduces the cost of duplicating *RP & V* data collection
- Increases the impact and visibility of *RP & V* research
- Promotes the *RP & V* research that created the data and its outcomes
- Provides important resources for education and training
- Provides an archiving service and platform for data sharing where sponsors (such as a U.K research council) require it

Our experience leads us to seek the following properties for the ITC environments that support scientific research.

- *Independence from technology:* If scientists develop a method, it runs unchanged on their machine (even in the field), on their institution's cluster, on shared resources, and on HPC systems.
- *Independence from change:* Their scientific focus can proceed unchanged without any gratuitous disruption.
- *Scalability and flexibility:* If many scientists want to use the data or computation, or if one or more needs a great deal of data or computation, this should be possible without reformulation, it only requires the authority to use resources.
- *Innovation potential:* Whilst those, typically the majority, who want to retain their current focus are protected, the critical minority of pioneers and innovators and those spanning disciplines or sub-disciplines, should find this easier because experiment and exploration are encouraged without limitations imposed by the ITC systems. In those cases where the new methods prove valuable it should be easy to share them and to propagate the necessary changes to data, software tools and services.
- *Freedom from trivial tasks:* Both the scientists and the those building new technology to support the science should find that many, hitherto tedious, tasks such as data wrangling, plugging in software and data, and handling changes in the digital context, should be handled reliably by automation with reference to experts only when necessary but with inspection always possible.
- *Efficiency:* as measured by agreed cost measures, such as weighted sums of carbon footprint, operational budget and response time, will be achieved by dynamic adjustment of the operational processes.
- *Reliability* The communities who use any research infrastructure should be able to trust the availability and correct operation of the ITC systems.
- *Repeatability* Any preserved process when run on preserved data should give results that agree with those of previous comparable runs.
- *Credit and blame:* The records kept by the system should be sufficient to fairly attribute credit and blame.

Individually, most of these goals are achievable, but much research is needed to deliver them in combination for the diverse elements of an Earth science research infrastructure. As future work we would like to convert the new rock physics and volcanology science gateway prototype into design that integrates well with EPOS. By using EFFORT investment we can test that design. However, this requires that the *RP & V* community reaches an agreement on common metadata, data formats, collaboration protocols and builds a community of integrated *RP & V* laboratories, observatories and researchers. The implementation will need to involve due to advances in the Earth sciences, due to advances to IT technology provision and in response to policy and practice decisions by the Earth scientist. The next phase would therefore be engineered in accord with the ENVRI reference model [17], would re-use developments in major e-Infraesture platforms, such as EUDAT, EGI[20] and ER-FLOW[21] [18] and would take account of emerging metadata standards such as WF4ever and W3C[22] provenance [19]

Acknowledgment

The research described here has been supported by the NERC UK Grant (NE/H02297X/1).

References

[1] P. F. Linington, Z. Milosevic, A. Tanaka, and A. Vallecillo, *Building Enterprise Systems with ODP: An Introduction to Open Distributed Processing*, 1st ed. Chapman & Hall/CRC, 2011.

[2] D. De Roure and C. Goble, "Software design for empowering scientists," *IEEE Software*, vol. 26, no. 1, pp. 88–95, 2009.

[3] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garcia-Cuesta, J. Gómez-Pérez, G. Klyne, K. Page, M. Roos, J. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C. Goble, "Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse," in *Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012), Heraklion, Greece*, May 2012.

[20] https://www.egi.eu/

[21] https://www.erflow.eu/

[22] http://www.w3.org/2011/prov/wiki/Main_Page

[4] M. Conway, R. Moore, A. Rajasekar, and J.-Y. Nief, "Demonstration of Policy-Guided Data Preservation Using iRODS," in *POLICY*, 2011, pp. 173–174.

[5] P. Kacsuk, Z. Farkas, M. Kozlovszky, G. Hermann, A. Balasko, K. Karoczkai, and I. Marton, "Ws-pgrade/guse generic dci gateway framework for a large variety of user communities," *J. Grid Comput.*, vol. 10, no. 4, pp. 601–630, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1007/s10723-012-9240-5

[6] B. Steve, M. Philip, H. Michael, B. Laura, and F. Paolo, "Creep-2: Long-term time-dependent rock deformation in a deep-sea observatory," 2010, http://adsabs.harvard.edu/abs/2010EGUGA..12.7874B.

[7] Data-Intensive Research Group, University of Edinburgh, "Rapid: Giving computational science a friendly face," 2012, http://research.nesc.ac.uk/rapid.

[8] A. Bhushan, "File Transfer Protocol (FTP) status and further comments," RFC 414, Internet Engineering Task Force, December 1972. [Online]. Available: http://www.ietf.org/rfc/rfc414.txt

[9] M. Widenius, D. Axmark, and A. B. Mysql, *MySQL Reference Manual*, 1st ed. O'Reilly Media, Inc., 2002. [Online]. Available: http://www.worldcat.org/isbn/0596002653

[10] D. Hull, K. Wolstencroft, R. Stevens, C. A. Goble, M. R. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services." *Nucleic Acids Research*, vol. 34, pp. 729–732, 2006.

[11] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, "Scientific workflow management and the Kepler system," *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, pp. 1039–1065, 2006.

[12] I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, *Workflows for e-Science: Scientific Workflows for Grids*. Springer-Verlag, 2007.

[13] E. Deelman, D. Gannon, M. Shields, and I. Taylor, "Workflows and e-Science: An overview of workflow system features and capabilities," *Future Generation Computer Systems*, vol. 25, no. 5, pp. 528–540, 2009.

[14] D. De Roure, C. Goble, and R. Stevens, "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows," *Future Generation Computer Systems*, vol. 25, pp. 561–567, 2009. [Online]. Available: doi:10.1016/j.future.2008.06.010

[15] R. Filgueira, M. Atkinson, A. Bell, B. Snelling, and I. Main, "Varpy: A python library for volcanology and rock physics data analysis. egu2014-3699," 2014.

[16] T. Megies, M. Beyreuther, R. Barsch, L. Krischer, and J. Wassermann, "ObsPy - What Can It Do for Data Centers and Observatories?" *Annals Of Geophysics*, vol. 54, no. 1, pp. 47–58, apr 2011. [Online]. Available: http://www.annalsofgeophysics.eu/index.php/annals/article/view/4838/5039

[17] Y. Chen, P. Martin, B. Magagna, H. Schentz, Z. Zhao, A. Hardisty, A. D. Preece, M. P. Atkinson, R. Huber, and Y. Legr, "A common reference model for environmental science research infrastructures." in *EnviroInfo*, ser. Berichte aus der Umweltinformatik, B. Page, A. G. Fleischer, J. Gbel, and V. Wohlgemuth, Eds. Shaker, 2013, pp. 665–673. [Online]. Available: http://dblp.uni-trier.de/db/conf/environfo/environfo2013.html#ChenMMSZHPAHL13

[18] G. Terstyanszky, T. Kukla, T. Kiss, P. Kacsuk, Ákos Balaskó, and Z. Farkas, "Enabling Scientific Workflow Sharing through Coarse-Grained Interoperability," *Journal of Future Generation Computing Systems*, submitted 2013 (under review) expected 2014.

[19] A. Spinuso, J. Cheney, and M. Atkinson, "Provenance for seismological processing pipelines in a distributed streaming workflow," in *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, ser. EDBT '13. New York, NY, USA: ACM, 2013, pp. 307–312. [Online]. Available: http://doi.acm.org/10.1145/2457317.2457369