THE UNIVERSITY of EDINBURGH
informatics

cisa | Centre for Intelligent Systems and their Applications

Scrying the next generation of data-intensive research infrastructure

Research at the Data-Intensive Research Group of the University of Edinburgh

Paul Martin

OSDC-PIRE 2014, University of Amsterdam

Edinburgh

THE UNIVERSITY of EDINBURGH
informatics

cisa | Centre for Intelligent Systems and their Applications

School of Informatics

www.inf.ed.ac.uk

Future research infrastructures…

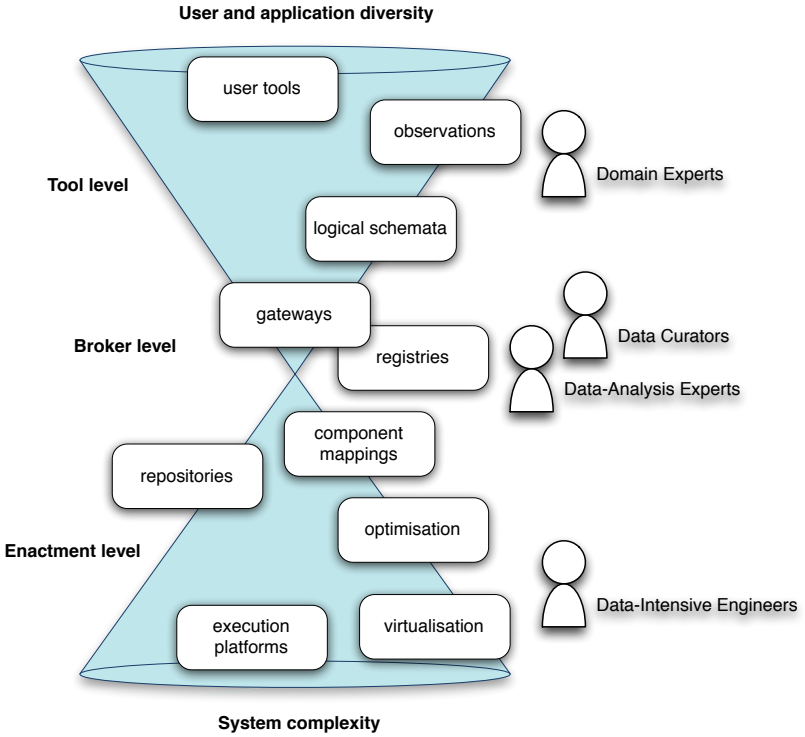
- …must support a large range of different **research interactions**.
  - Data collection, curation, processing and publication.
  - Curation of **models** and **methods**.
  - Community networks and cross-infrastructure interactions.
- …must support a diverse cast of **research actors**.
  - Investigators, empiricists, theorists, librarians, engineers, *etc*.
- …must balance conflicting issues:
  - Openness and accountability.
  - Preservation and accessibility.
  - Interoperability and efficacy.
  - Oversight and autonomy.

The Data-Intensive Research Group

- Part of the Centre for Intelligent Systems and their Applications in the School of Informatics at the University of Edinburgh.
- Research agenda focuses on of how best to address current and future data-intensive research problems:
  - How to manage large volumes of data;
  - How to process distributed data in different environments;
  - How to manage the *code and tools* used to handle data.
- Recent emphasis has been on workflow-based systems: languages and tools for workflow composition, services for deploying workflows, workflow optimisation and provenance gathering, *etc*…
- …but also, infrastructure modelling, scientific gateways, commodity supercomputing and anything else that catches our interest.

Supporting Research Interactions

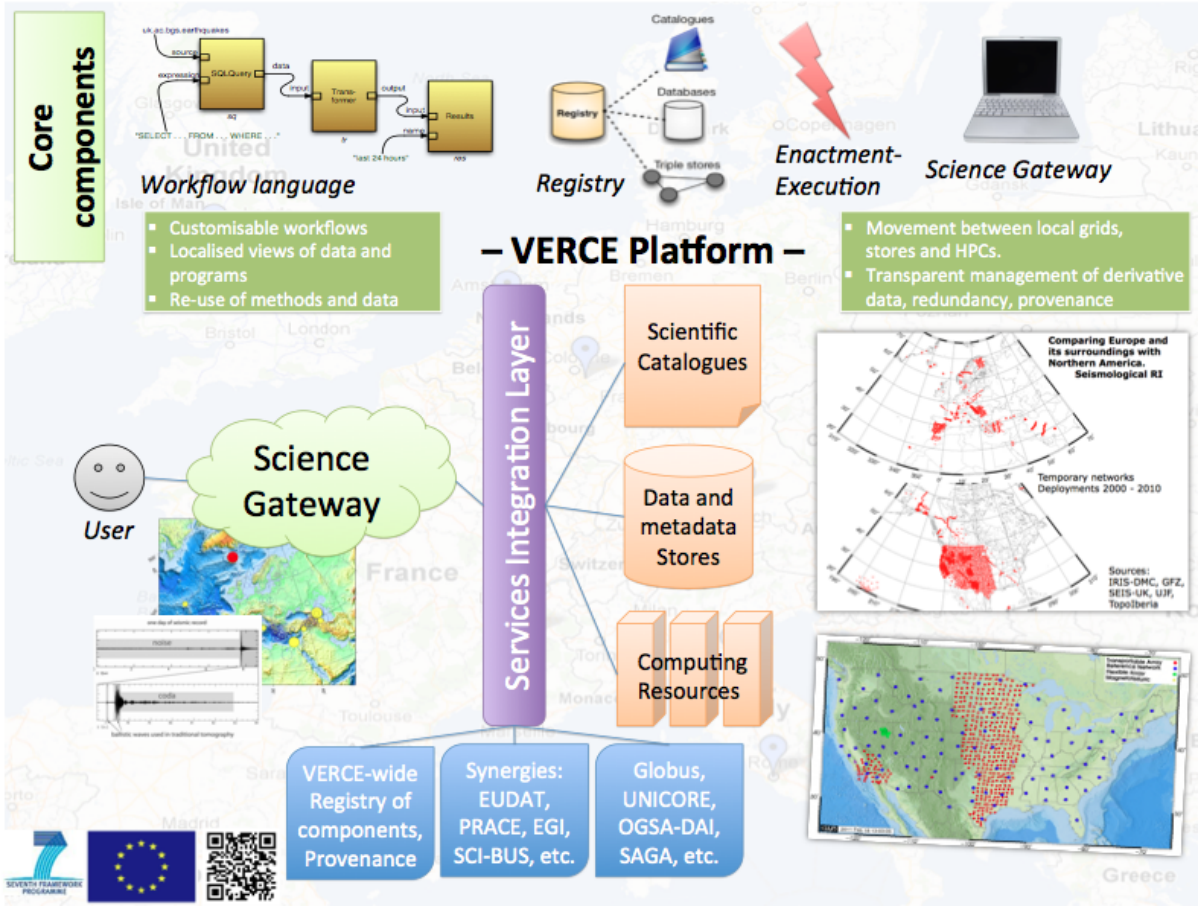
- Support a diverse range of interactions by domain experts at the high level…
- …by providing standard interchange formats…
- …that sit atop a heterogeneous array of execution platforms.



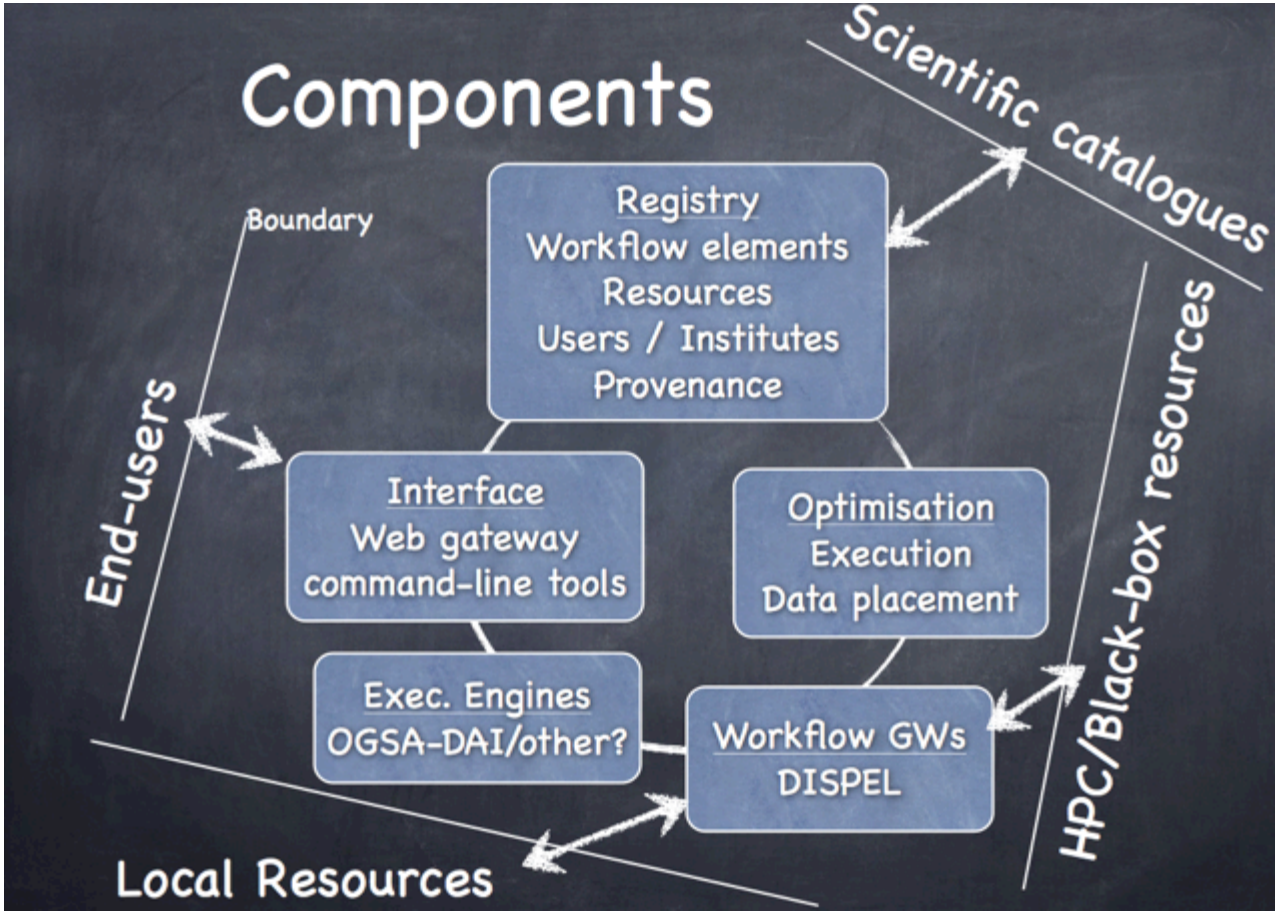
VERCE

- Virtual Earthquake and Seismology Research Community e-Science Environment in Europe.
- Design, build and integrate components for data processing in the seismology domain.
  - Streamline the process of configuring and conducting several standard types of computational task.
  - Open facilities for the broader community.
  - Focus on particular ‘data-intensive’ and ‘HPC’ use-cases.
- ‘Satellite’ project of EPOS (European Plate Observing System).
  - Contribute to EPOS Core Services.

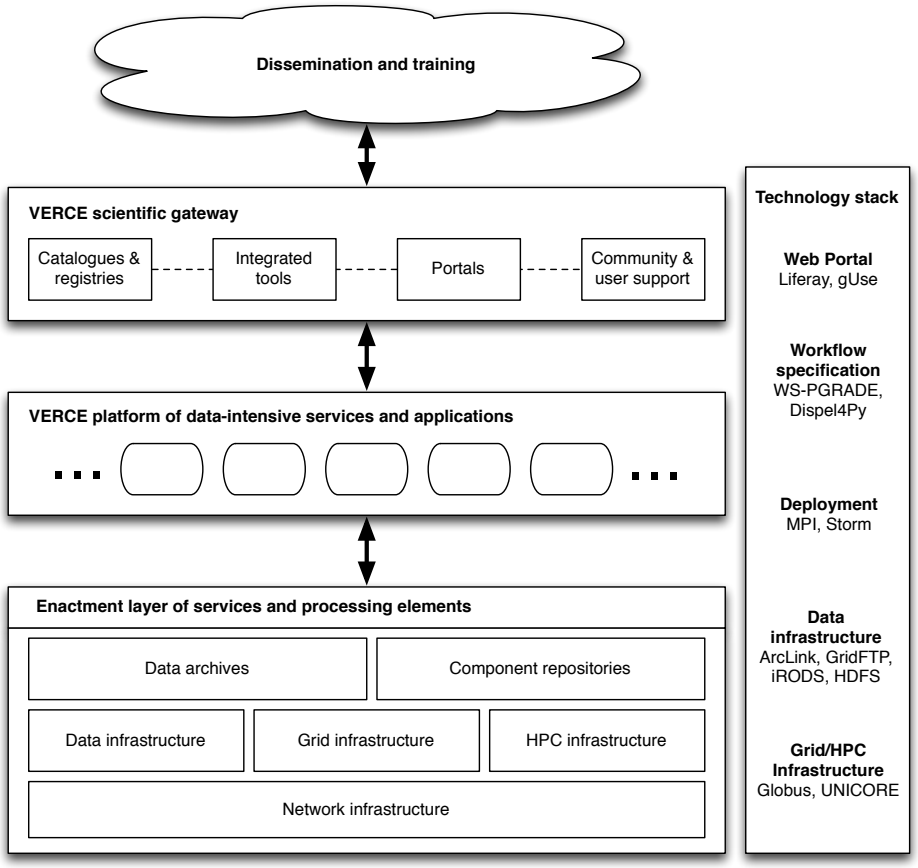
VERCE Overview



VERCE Principles



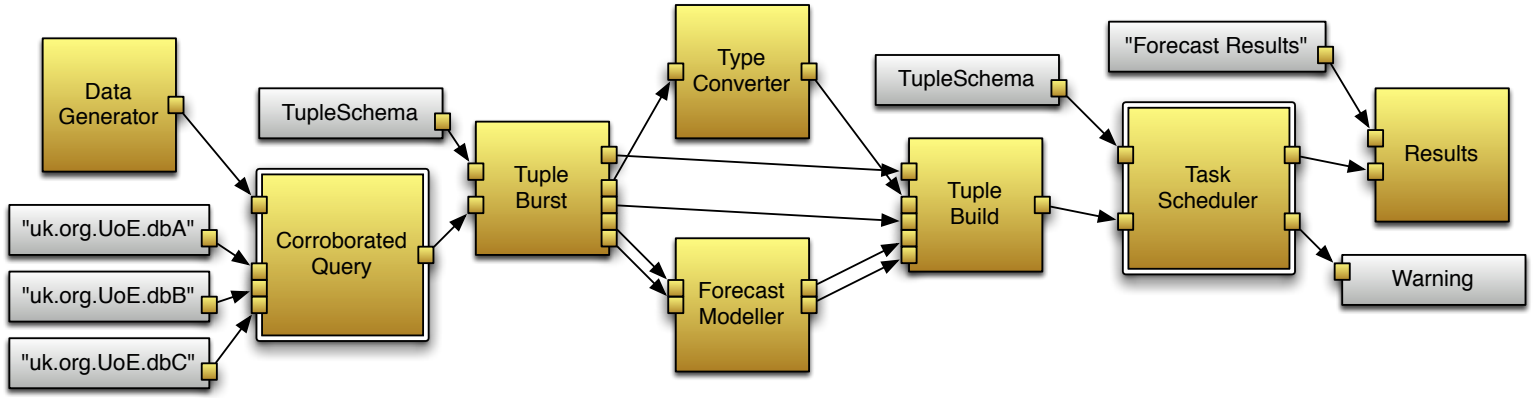
VERCE Technology Stack (c.2014)



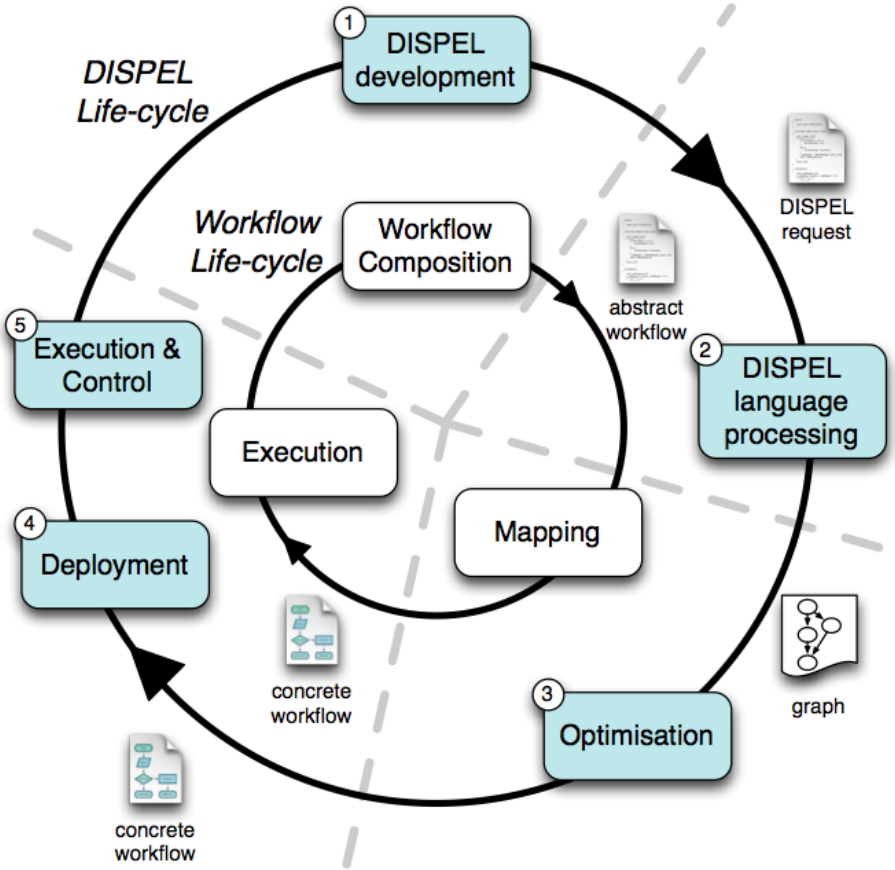
Dispel4Py

- Python-based implementation of DISPEL (Data-Intensive Systems Process Engineering Language).
  - Used to describe distributed data-streaming workflows at a logical level.
  - Wraps Python code into Processing Elements (PEs; initial focus on seismology applications).
  - Workflow graph can be deployed on various platforms (currently Storm and MPI).
- Principles of Dispel:
  - Inline specification of new PEs as compositions of existing PEs.
  - Strong typing for both language and dataflow with additional semantic (domain) annotation.
  - Work in progress…

Dispel4Py workflow illustration



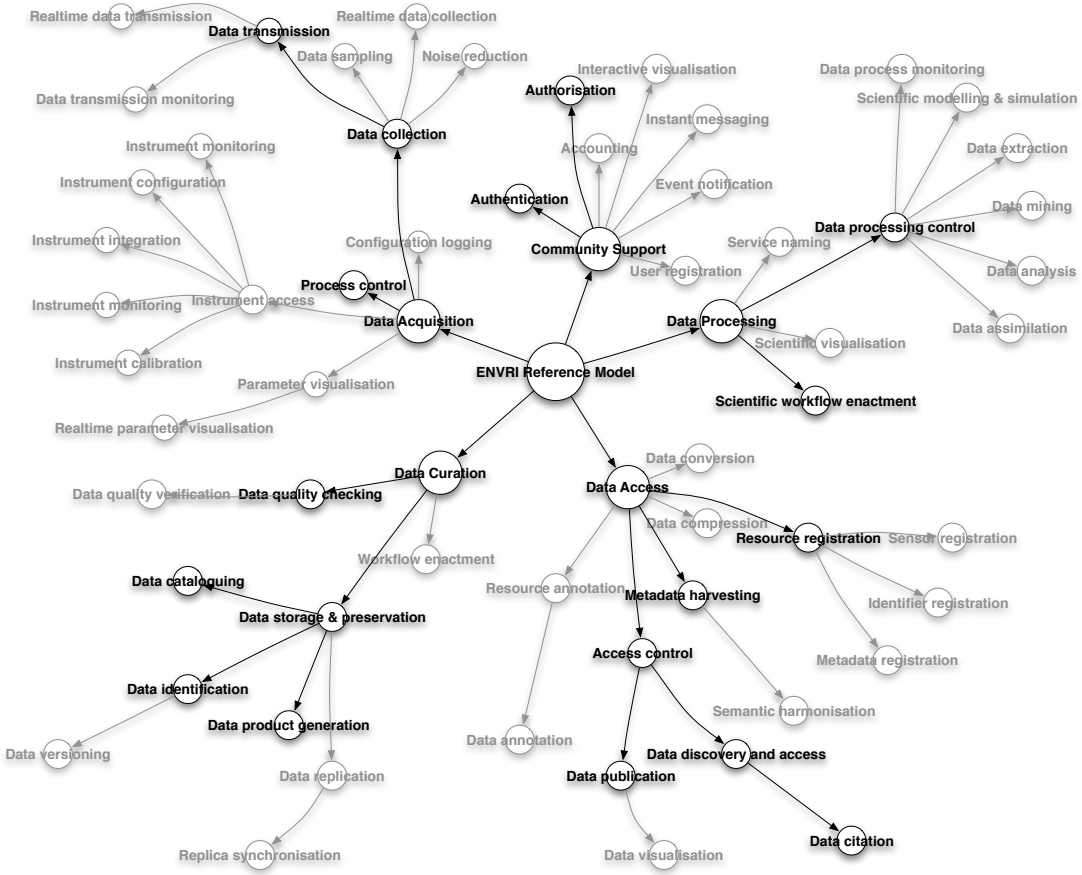
Dispel4Py lifecycle



ENVRI

- Common Operations of Environmental Research Infrastructures.
- Initiative to promote interoperability between ESFRI projects in the Environmental Cluster.
  - Model characteristics of environmental research infrastructures to identify commonalities and gaps.
  - Provide tools and services for data discovery and integration.
  - Improve social links between ESFRI and affiliated projects.
- Part of a general strategic effort to simplify the construction of bespoke infrastructure by pooling expertise and resources.

ENVRI Requirements

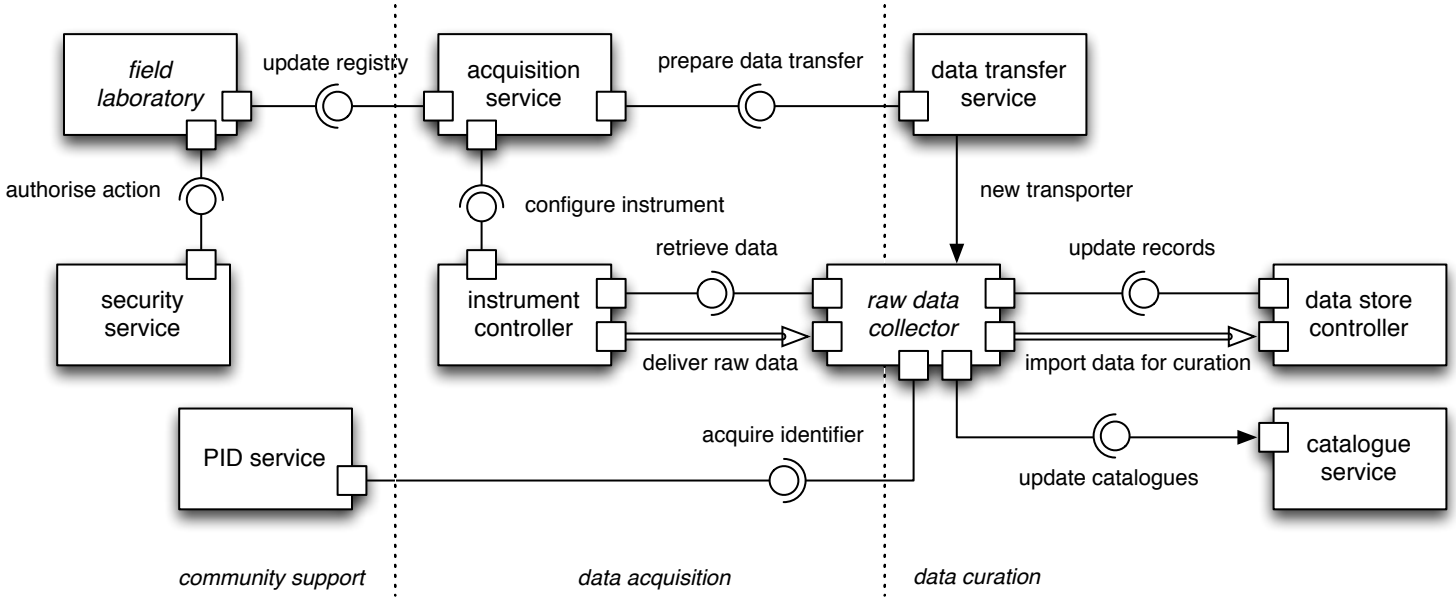


ENVRI Reference Model

- A standard abstract model for environmental research infrastructures.
- Founded on RM-ODP (Reference Model for Open Distributed Processing).
  - Standard for modelling distributed systems.
  - Viewpoint based: Enterprise, Information, Computation, Engineering and Technology.
  - Support for UML-style design.
- Current model iteration based on core ‘data pipeline’ (acquisition, curation, access).
  - Lightweight modelling of Enterprise (Science), Information and Computational Viewpoints.
  - Main study cases: EISCAT_3D, EPOS and ICOS.

ENVRI Reference Model Example

- Example of raw data collection from the computational viewpoint:



EFFORT

- Earthquake and Failure Forecasting in Real Time.
- Project to monitor rock failure experiments in real time.
  - Rock samples are subjected to continued pressure in laboratory conditions.
  - Stress leads to deformation, leading to sudden failure.
  - Models of rock failure may apply to plate deformation and volcanic events.
- Need ability to continuously and reliably collect data from remote experiments, relate to proposed models and provide visualisations on demand.
- Project expanded to build a standard library for volcanology and rock physics analyses (VarPy).

EFFORT system overview

Technologies

Data Transfer

- Creep-2: **FAST**: **Adaptive JAVA tool** by using **RSYNC protocol**.

Data Storage

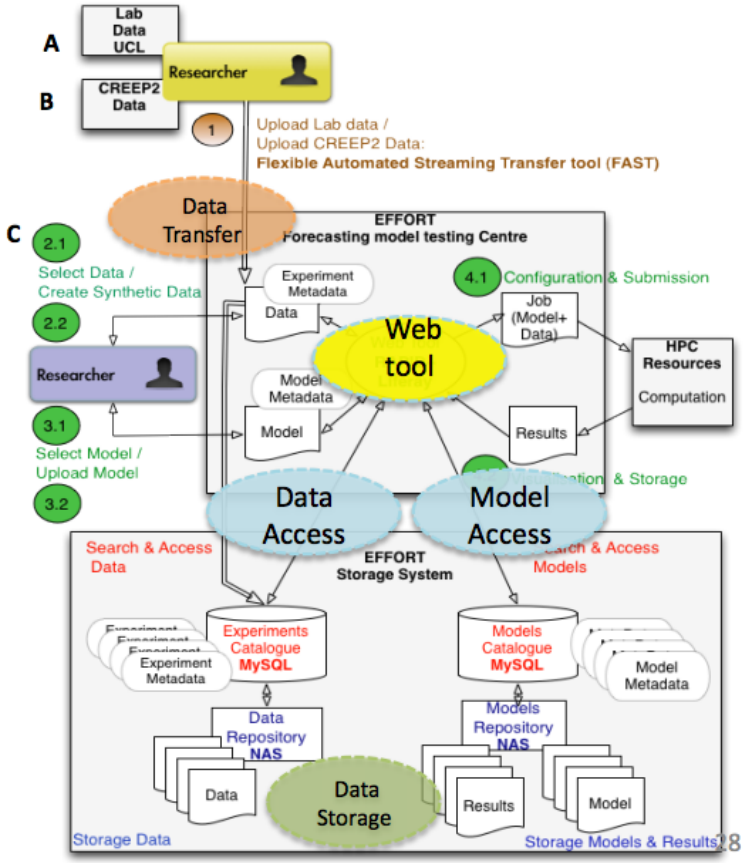
- Creep-2: **Metadata** storage in **MYSQL** database. **Data** storage in **NAS File System**.

Web tool

- EFFORT website developed with **LIFERAY** web-framework. The website has several **RAPID** portlets to select data and submit the forecast models to the **HPC resources**. The results of the models are visualized in the website.

Data & Model Access

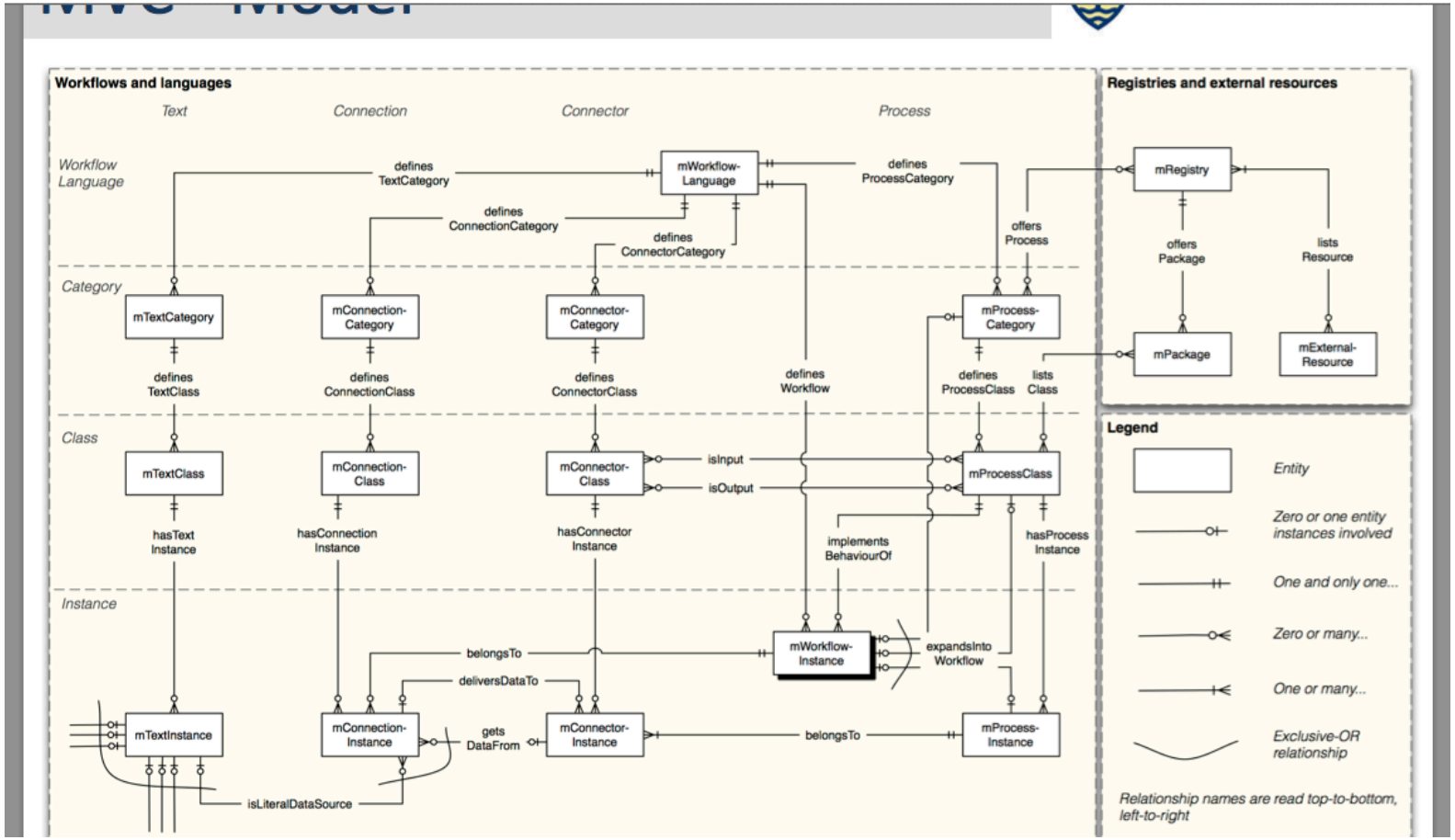
Through the EFFORT website, metadata about different experiment are displayed. The user can select the data to be use for executing the models.



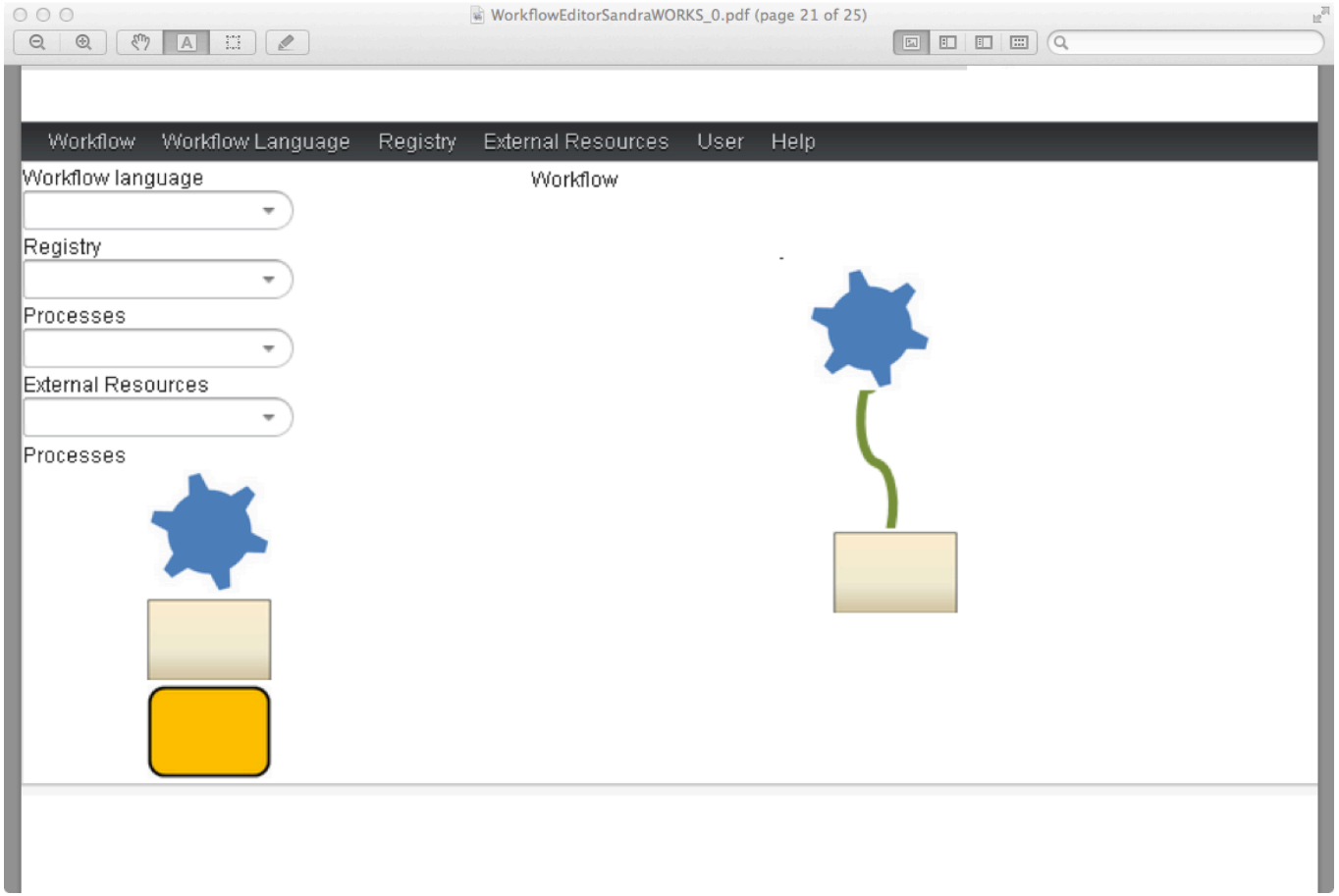
GeWWE

- Generic Web-based Workflow Editor
- Project to build a multi-target workflow editing tool.
  - Thesis is that workflows are always built from the same fundamental components (standard schema).
  - Average user is not keen to learn any specific workflow programming language (like Dispel…).
  - Can map workflows to a number of target languages / platforms.

GeWWE Schema



GeWWE screenshot



Other Projects

- EDIM1 – commodity data-brick computing.
- TerraCorrelator – doing data-intensive geoscience.
- DECIPHER – quasi-anonymous analysis of medical data.