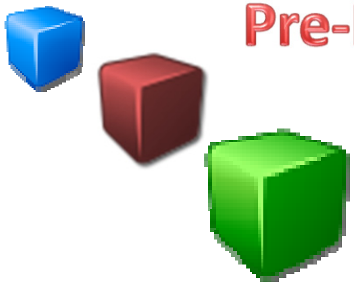
Safe Data Replication Service

A Simple and Reliable EUDAT Service for Scientific Communities



Pre-Production State

Morris Riedel et al.
Juelich Supercomputing Centre
EUDAT Conference, Barcelona

SEVENTH FRAMEWORK PROGRAMME

EUDAT

Date: 23th October 2012

EUDAT Portal

Integrated APIs and harmonized access to EUDAT facilities

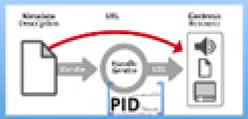
Metadata Catalog

Aggregated EUDAT metadata domain.
Data inventory

Requirement
Provide an inventory of metadata across disciplines

Function
- Joint metadata domain
- Catalog indexing stored data

eudat-metadata@postit.csc.fi

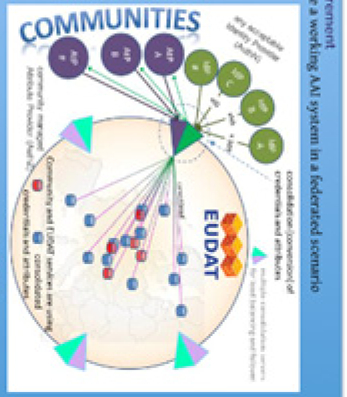


PID-based index

AAI

Network of trust among authentication and authorization actors

Requirement
Provide a working AAI system in a federated scenario



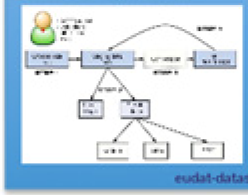
Function
- Integrate existing identification systems
- Establish a network of trust among AA, IdP and SP providers, attribute authorities and federations attribute harmonization

eudat-AAI@postit.csc.fi

Data Staging

Dynamic replication to HPC workspace for processing

Requirement
Provide a service to stage data between EUDAT infrastructure and HPC/HTC resources



Function
Dynamically replicate subset of data stored in EUDAT to HPC workspace

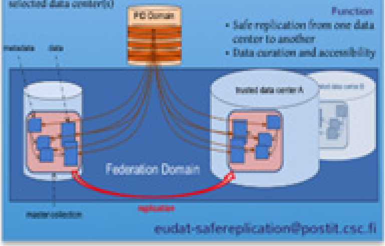
eudat-datastaging@postit.csc.fi

Safe Replication

Data curation and access optimization

Requirement
Provide a service to replicate and curate data to selected data center(s)

Function
- Safe replication from one data center to another
- Data curation and accessibility

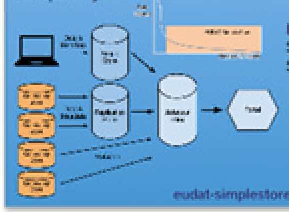


eudat-safereplication@postit.csc.fi

Simple Store

Researcher data store (simple upload, share and access)

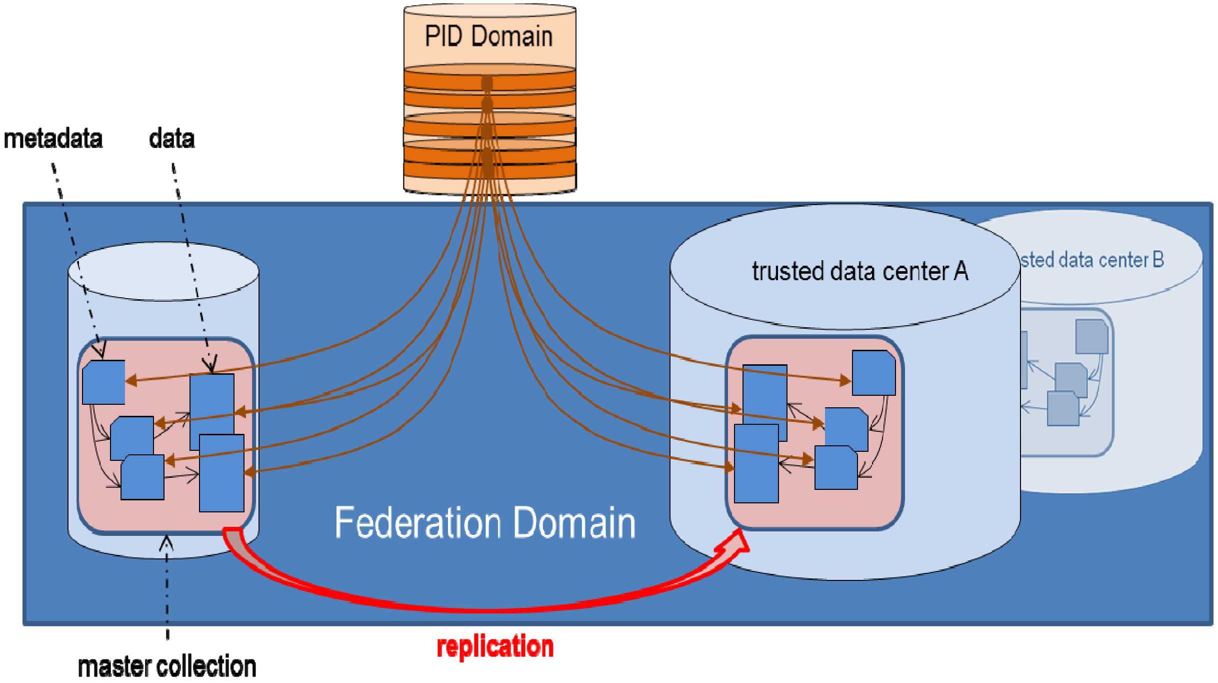
Requirement
Provide a simple service to store user data temporarily

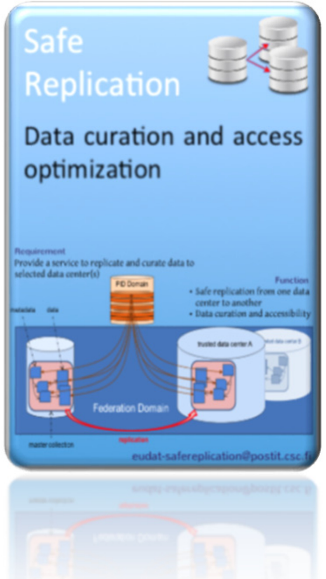


Function
Simple upload
Store data

eudat-simplestore@postit.csc.fi

Safe Replication Service in a Nutshell




Safe Replication

Data curation and access optimization

Requirement
Provide a service to replicate and curate data to selected data center(s)

Function
- Safe replication from one data center to another
- Data curation and accessibility

eudat-safereplication@postit.csc.fi

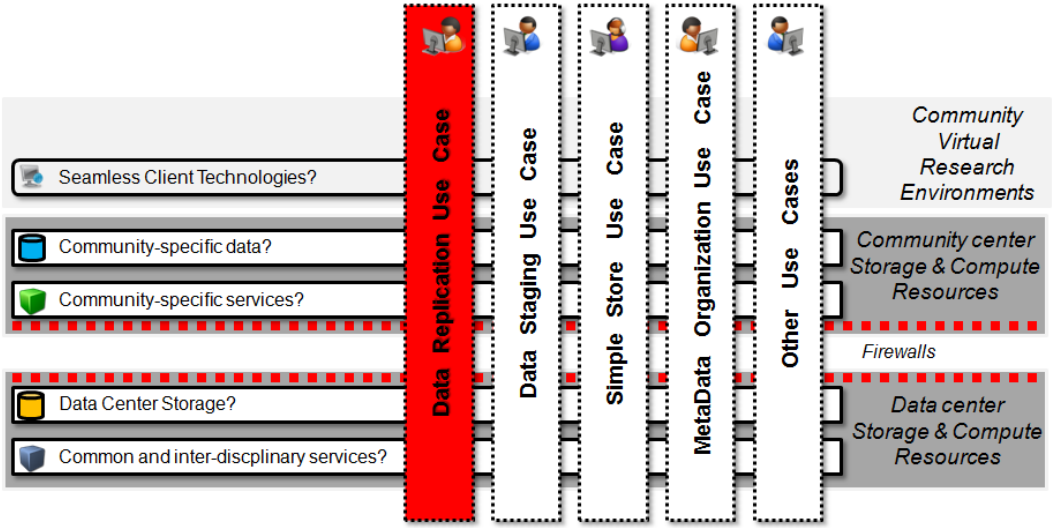
Make data referencable

Better accessibility of scientific data

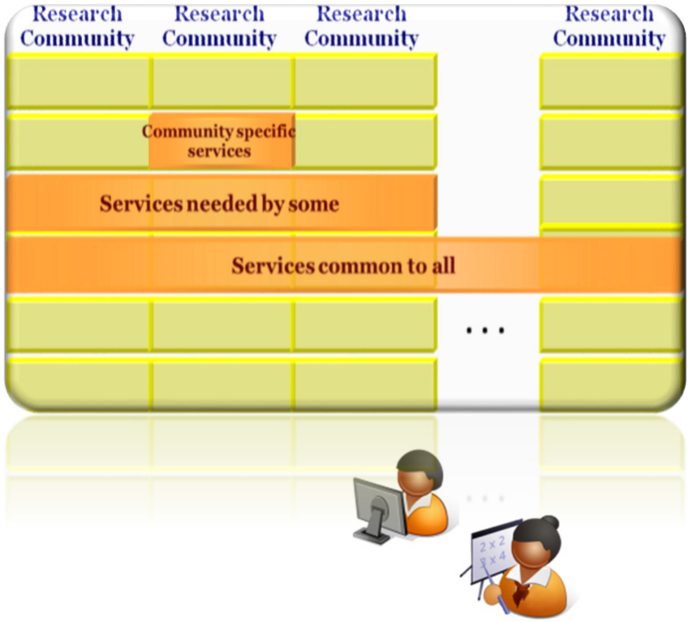
More optimal data curation

High degrees of reliability and trust

Federated Approach for Use Cases

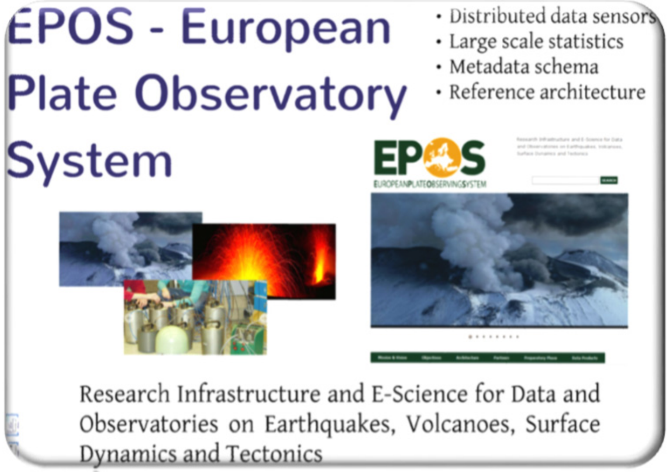


Data Replication Use Case
Data Staging Use Case
Simple Store Use Case
MetaData Organization Use Case
Other Use Cases
Seamless Client Technologies?
Community-specific data?
Community-specific services?
Data Center Storage?
Common and inter-discplinary services?
Community Virtual Research Environments
Community center Storage & Compute Resources
Firewalls
Data center Storage & Compute Resources



Create M replications at different data centers for N years,
exclude data centers X to data centers Z from the replication scheme
and make them all accessible by maintaining the given access permissions.

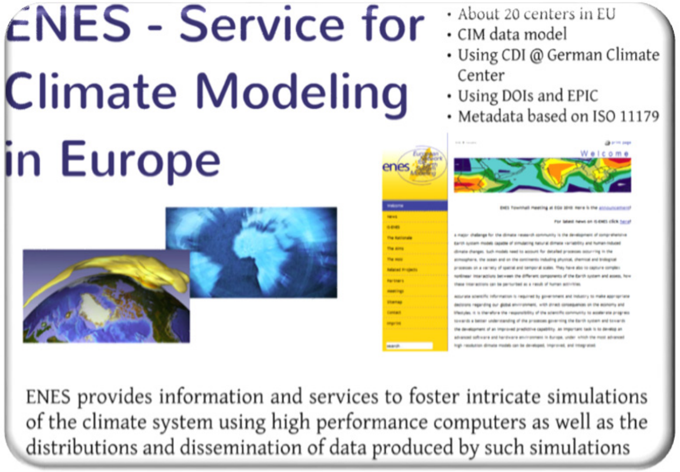
Forming Strong EUDAT Collaborations



EPOS - European Plate Observatory System

- Distributed data sensors
- Large scale statistics
- Metadata schema
- Reference architecture

Research Infrastructure and E-Science for Data and Observatories on Earthquakes, Volcanoes, Surface Dynamics and Tectonics



ENES - Service for Climate Modeling in Europe

- About 20 centers in EU
- CIM data model
- Using CDI @ German Climate Center
- Using DOIs and EPIC
- Metadata based on ISO 11179

ENES provides information and services to foster intricate simulations of the climate system using high performance computers as well as the distributions and dissemination of data produced by such simulations

CLARIN - Common Language Resources and Technology Infrastructure

- About 200 centers in EU
- Require PIDs, CMDI
- ISOcat, SCHEMcat
- Virtual Language Obs.

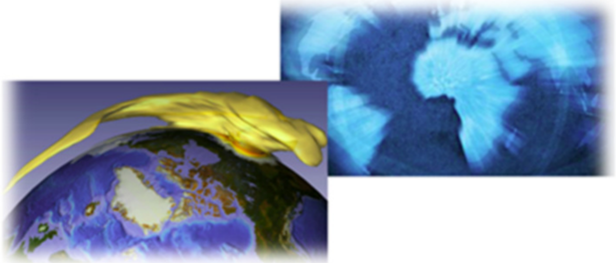
http://www.clarin.eu/vlo/

The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable

Scientific Community	Community Centers	Data Centers
CLARIN	MPI-PL	RZG, SARA
ENES	DKRZ	JSC, CSC
EPOS	INGV	CINECA, SARA

Use Case Example: Climate Science Data

- ENES: Service for Climate Modelling in Europe
  - Provides services to foster intricate simulations of the climate system using high-performance computers
  - **Enables the distribution and dissemination of data produced by such simulations**
  - Other Facts: *about 20 EU centres; CIM data model; uses DOIs and EPIC handles; metadata in ISO 11179;*



InfraStructure for the European Network for Earth System Modelling

is-enes

Slide content kindly provided by Hannes Thielmann, DKRZ, a climate scientist in earth system modelling

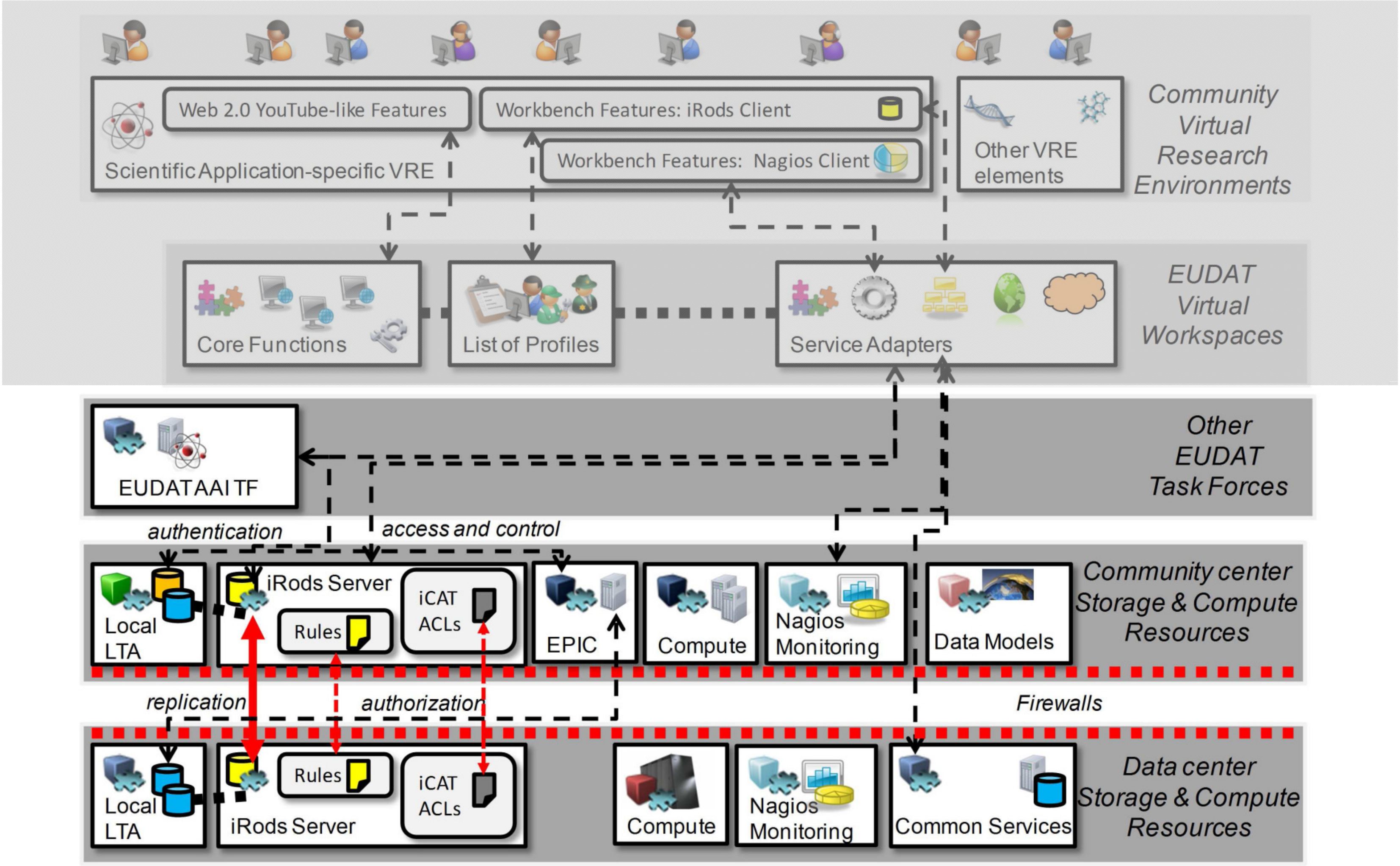
EUDAT

Concrete Replicated Climate Scientific Data

- Complexity: ENES & CMIP5 & IPCC AR5
  - ENES contributes to the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5).
  - Coupled Model Intercomparison Project Phase 5 (CMIP5) model data that will serve as the basis for IPCC AR5.
  - This data prepared will be made available to the international climate community.
  - The Earth System Grid Federation (ESGF) is a partnership of climate modeling centers created to provide secure, web-based, distributed access to CMIP5 model data.

Slide content kindly provided by Hannes Thielmann, DKRZ, a climate scientist in earth system modelling

Use Cases Derived Reference Architecture



Selected References

- Documentation on iRODS and EPIC/Handle system available on the Web
- 1st EUDAT Conference Training Day - Many training sessions yesterday!
  - PID handling & services, iRODS policies, rules, micro-services, etc.
- EUDAT Newsletter April 2012
  - Check the EUDAT WebSite
- M. Riedel and P. Wittenburg et al.
  'A Data Infrastructure Reference Model with Applications - Towards Realization of a ScienceTube Vision with a Data Replication Service',
  Journal of Internet Applications,
  to be published early 2013
- Contact to specialists:
  eudat-safereplication@postit.csc.fi

A Data Infrastructure Reference Model with Applications

Towards Realization of a ScienceTube Vision with a Data Replication Service

Morris Riedel · Peter Wittenburg · Johannes Reetz · Mark van de Sanden · Jedrzej Rybicki · Benedikt von St. Vieth · Giuseppe Fiameni · Giacomo Mariani · Alberto Michelini · Claudio Cacciari · Willem Elbers · Daan Broeder · Robert Verkerk · Elena Erastova · Michael Lautenschlaeger · Reinhard Budig · Hannes Thielmann · Peter Coveney · Stefan Zasada · Ali Haidar · Otto Buechner · Cristina Manzano · Shiraz Memon · Shahbaz Memon · Heikki Helin · Jari Suhonen · Damien Lecarpentier · Kimmo Koski · Thomas Lippert

Received: date / Accepted: date

Abstract The wide variety of scientific user communities work with data since many years and thus have already a wide variety of data infrastructures in production today. The aim of this paper is thus not to create one new general data architecture that would fail to be adopted by each and any individual user community. Instead this contribution aims to design a reference model with abstract entities that is able to federate existing concrete infrastructures under one umbrella. A reference model is an abstract framework for understanding significant entities and relationships between them and thus helps to understand existing data infrastructures when comparing them in terms of functionality, services, and boundary conditions. A derived architecture from such a reference model then can be used to create a federated architecture that builds on the existing infrastructures that could align to a major common vision. This common vision is named as 'ScienceTube' as part of this contribution that determines the high-level goal that the reference model aims to support. This paper will describe how a well-focused use case around data replication and its related activities in the EUDAT project [6] aim to provide a first step towards this vision. Concrete stakeholder requirements arising from scientific end users such as those of the European Strategy Forum on Research Infrastructure (ESFRI) projects underpin this contribution with clear evidence that the EUDAT activities are bottom-up thus providing real solutions towards the so often only described 'high-level big data challenges'. The followed federated approach taking advantage of community and data centers (with large computational resources) further describes how data replication services enable data-intensive computing of terabytes or even petabytes of data emerging from ESFRI projects.

Keywords Reference Model · ScienceTube · Data Infrastructure · Replication

M. Riedel, J. Rybicki, B. von St. Vieth, A.S. Memon, M.S. Memon, Th. Lippert, O. Buechner, C. Manzano
Juelich Supercomputing Centre
E-mail: m.riedel@fz-juelich.de

P. Wittenburg, D. Broeder, W. Elbers
Max Planck Institute for Psycholinguistic, Nijmegen, Netherlands

J. Reetz, E. Erastova
Rechenzentrum Garching, Munich, Germany

M. van de Sanden, R. Verkerk
Stichting Academisch Rekencentrum Amsterdam, Amsterdam, Netherlands

G. Fiameni, C. Cacciari, G. Mariani
CINECA, Bologna, Italy

A. Michelini
INGV, Italy

M. Lautenschlaeger, R. Budig, H. Thielmann
Deutsches Klimarechenzentrum, Hamburg, Germany

P. Coveney, S. Zasada, A. Haidar
University College London, London, UK

J. Suhonen, H. Helin, D. Lecarpentier, K. Koski
CSC - IT Center for Science, Finland