EUDAT

Towards a pan-European Collaborative Data Infrastructure

Giuseppe Fiameni (g.fiameni@cineca.it) – Claudio Cacciari

SuperComputing, Application and Innovation – CINECA

Johannes Reatz

RZG, Germany

Damien Lecarpentier

CSC-IT Center for Science, Finland

Napoli, 16 May 2012

EUDAT Key facts

Project Name	EUDAT – European Data
Start date	1st October 2011
Duration	36 months
Budget	16,3 M€ (including 9,3 M€ from the EC)
EC call	Call 9 (INFRA-2011-1.2.2): Data infrastructure for e-Science (11.2010)
Participants	25 partners from 13 countries (national data centers, technology providers, research communities, and funding agencies)
Objectives	“To deliver cost-efficient and high quality Collaborative Data Infrastructure (CDI) with the capacity and capability for meeting researchers’ needs in a flexible and sustainable way, across geographical and disciplinary boundaries.”

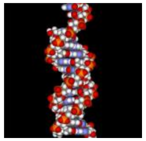
The current data infrastructure landscape: challenges and opportunities

- Long history of data management in Europe: several existing data infrastructures dealing with established and growing user communities (e.g., ESO, ESA, EBI, CERN)

- New Research Infrastructures are emerging and are also trying to build data infrastructure solutions to meet their needs (CLARIN, EPOS, ELIXIR, ESS, etc.)

European Strategy Forum on Research Infrastructures
ESFRI
EUROPEAN ROADMAP FOR RESEARCH INFRASTRUCTURES
Report 2006

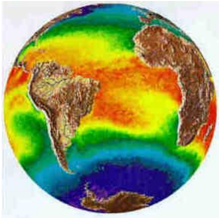
➢**However, most of these infrastructures and initiatives address primarily the needs of a specific discipline and user community**



Challenges

- Compatibility, interoperability, and cross-disciplinary research
  - ➢ how to re-use and recombine data in new scientific contexts (i.e. across disciplinary domains)

- Data growth in volume and complexity (the so-called “data tsunami”)
  - ➢ strong impact on costs threatening the sustainability of the infrastructure

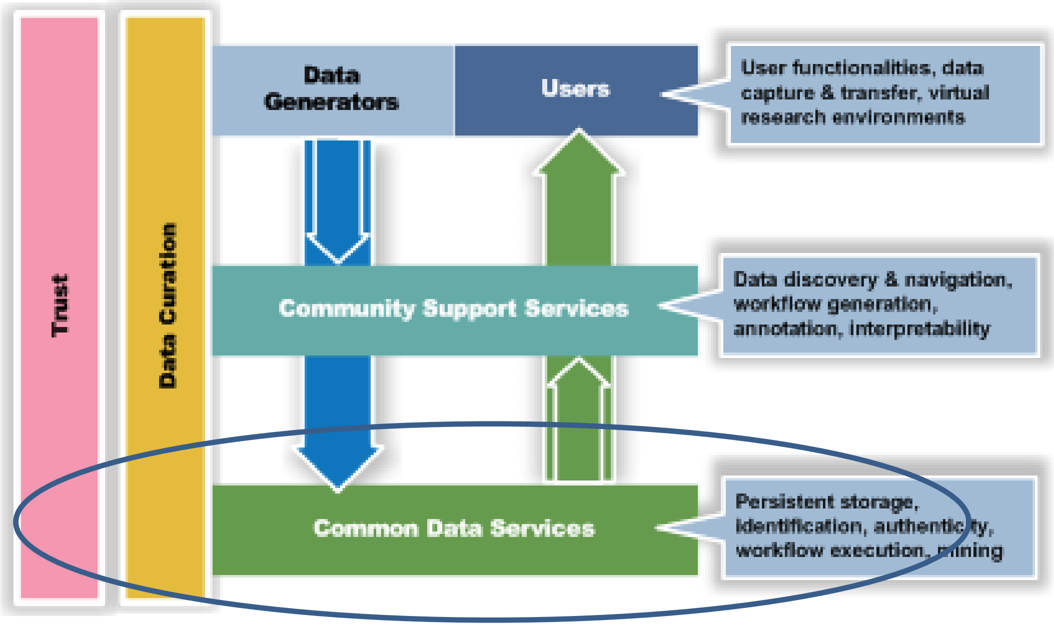


Opportunities

- Potential synergies do exist: although disciplines have different ambitions, they have common basic needs and service requirements that can be matched with generic pan-European services supporting multiple communities, thus ensuring at the same time greater interoperability.

➢**Strategy needed at pan-European level**

The CDI concept



User functionalities, data capture & transfer, virtual research environments

Data discovery & navigation, workflow generation, annotation, interpretability

Persistent storage, identification, authenticity, workflow execution, mining

EUDAT Core Service Areas

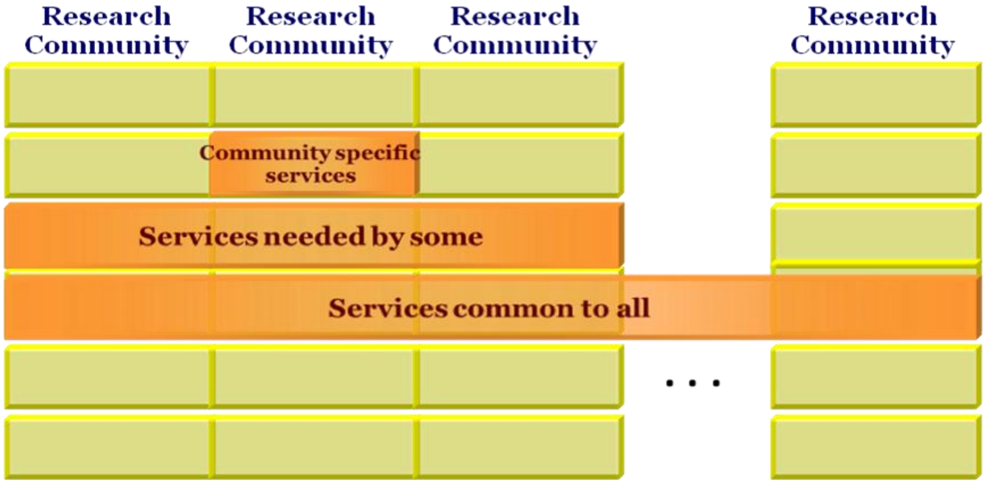
Community-oriented services

- Simple Data Acces and upload
- Long term preservation
- Shared workspaces
- Execution and workflow (data mining, etc.)
- Joint metadata and data visibility

Enabling services (making use of existing services where possible

- Persistent identifier service (EPIC, DataCite)
- Federated AAI service
- Network Services
- Monitoring and accounting

Core services are building blocks of EUDAT‘s Common Data Infrastructure
mainly included on bottom layer of data services



Data centers and Communities

CINES

epcc

UNINETT sigma

Science & Technology Facilities Council

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

sara
Computing & Networking Services

SNIC

DKRZ

CERFACS

JÜLICH
FORSCHUNGSZENTRUM

CINECA

umweltbundesamt
PERSPEKTIVEN FÜR UMWELT & GESELLSCHAFT

Max-Planck-Institut für Meteorologie

POZNAN SUPERCOMPUTING NETWORKING

CSC

Max Planck Institute for Psycholinguistics

KIT
Karlsruhe Institute of Technology

CERN

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

maatG

INGV

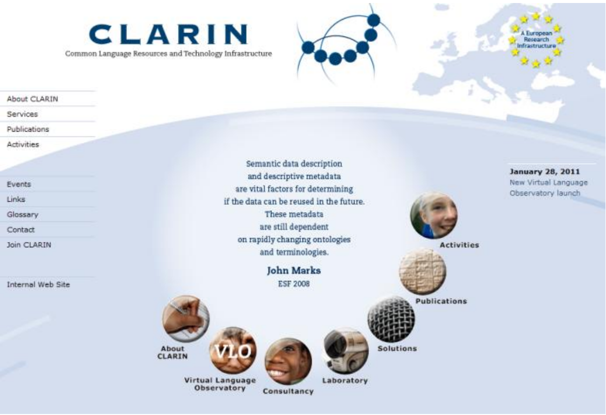
RedIRIS

UCL

rzg
IPP
RECHEN-ZENTRUM GARCHING

EUDAT

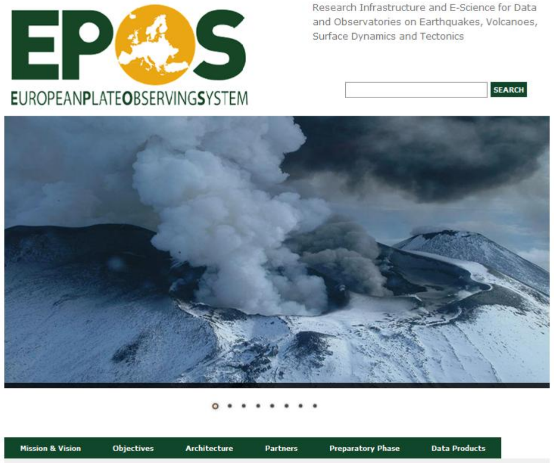
First EUDAT Communities



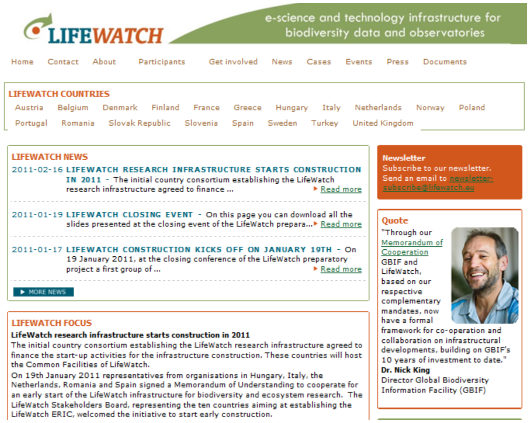
Semantic data description
and descriptive metadata
are vital factors for determining
if the data can be reused in the future.
These metadata
are still dependent
on rapidly changing ontologies
and terminologies.

John Marks
ESF 2008

January 28, 2011
New Virtual Language Observatory launch



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



LIFEWATCH COUNTRIES

Austria Belgium Denmark Finland France Greece Hungary Italy Netherlands Norway Poland Portugal Romania Slovak Republic Slovenia Spain Sweden Turkey United Kingdom

LIFEWATCH NEWS

2011-02-16 LIFEWATCH RESEARCH INFRASTRUCTURE STARTS CONSTRUCTION IN 2011 - The initial country consortium establishing the LifeWatch research infrastructure agreed to finance ... ▸ Read more

2011-01-19 LIFEWATCH CLOSING EVENT - On this page you can download all the slides presented at the closing event of the LifeWatch prepara...▸ Read more

2011-01-17 LIFEWATCH CONSTRUCTION KICKS OFF ON JANUARY 19TH - On 19 January 2011, at the closing conference of the LifeWatch preparatory project a first group of ... ▸ Read more

Newsletter
Subscribe to our newsletter.
Send an email to newsletter-subscribe@lifewatch.eu

LIFEWATCH FOCUS

LifeWatch research infrastructure starts construction in 2011

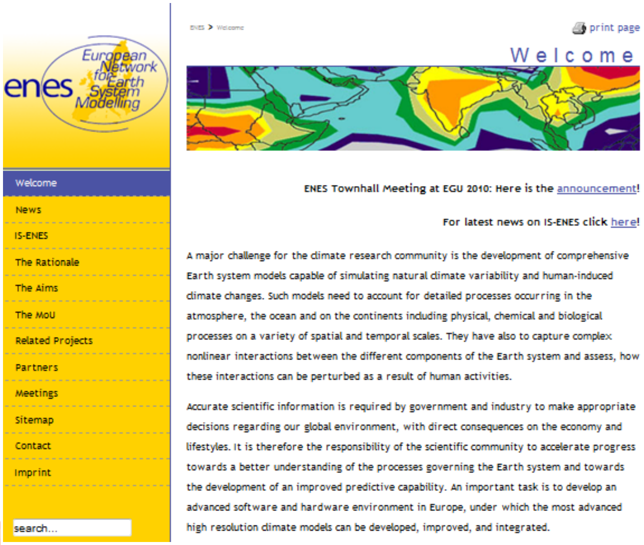
The initial country consortium establishing the LifeWatch research infrastructure agreed to finance the start-up activities for the infrastructure construction. These countries will host the Common Facilities of LifeWatch.

On 19th January 2011 representatives from organisations in Hungary, Italy, the Netherlands, Romania and Spain signed a Memorandum of Understanding to cooperate for an early start of the LifeWatch infrastructure for biodiversity and ecosystem research. The LifeWatch Stakeholders Board, representing the ten countries aiming at establishing the LifeWatch ERIC, welcomed the initiative to start early construction.

Quote

"Through our Memorandum of Cooperation GBIF and LifeWatch, based on our respective complementary mandates, now have a formal framework for co-operation and collaboration on infrastructural developments, building on GBIF's 10 years of investment to date."

Dr. Nick King
Director Global Biodiversity Information Facility (GBIF)



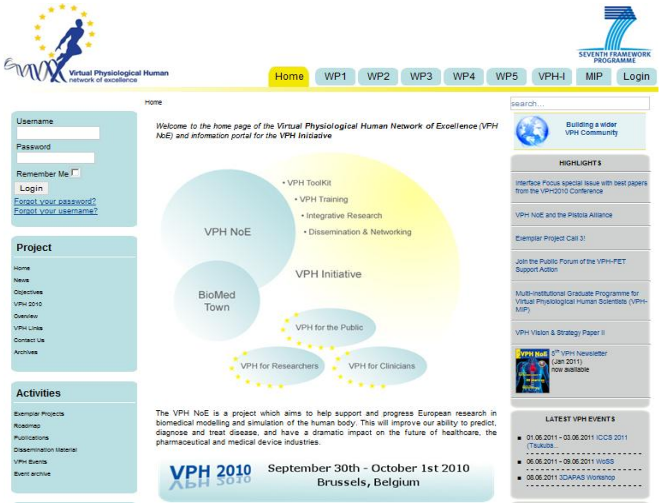
Welcome

ENES Townhall Meeting at EGU 2010: Here is the announcement!

For latest news on IS-ENES click here!

A major challenge for the climate research community is the development of comprehensive Earth system models capable of simulating natural climate variability and human-induced climate changes. Such models need to account for detailed processes occurring in the atmosphere, the ocean and on the continents including physical, chemical and biological processes on a variety of spatial and temporal scales. They have also to capture complex nonlinear interactions between the different components of the Earth system and assess, how these interactions can be perturbed as a result of human activities.

Accurate scientific information is required by government and industry to make appropriate decisions regarding our global environment, with direct consequences on the economy and lifestyles. It is therefore the responsibility of the scientific community to accelerate progress towards a better understanding of the processes governing the Earth system and towards the development of an improved predictive capability. An important task is to develop an advanced software and hardware environment in Europe, under which the most advanced high resolution climate models can be developed, improved, and integrated.



Welcome to the home page of the ***Virtual Physiological Human Network of Excellence*** *(VPH NoE) and information portal for the* ***VPH Initiative***

The VPH NoE is a project which aims to help support and progress European research in biomedical modelling and simulation of the human body. This will improve our ability to predict, diagnose and treat disease, and have a dramatic impact on the future of healthcare, the pharmaceutical and medical device industries.

VPH 2010 September 30th - October 1st 2010
Brussels, Belgium

HIGHLIGHTS

- Interface Focus special issue with best papers from the VPH2010 Conference
- VPH NoE and the Pistoia Alliance
- Exemplar Project Call 3!
- Join the Public Forum of the VPH-FET Support Action
- Multi-institutional Graduate Programme for Virtual Physiological Human Scientists (VPH-MIP)
- VPH Vision & Strategy Paper II
- 5th VPH Newsletter (Jan 2011) now available

LATEST VPH EVENTS

- 01.06.2011 - 03.06.2011 ICCS 2011 (Tsukuba...
- 06.06.2011 - 09.06.2011 WoSS
- 08.06.2011 3DAPAS Workshop

Building the services

6 service/use cases identified

Safe replication: Allow communities to safely replicate data to selected data centers for storage and do this in a robust, reliable and highly available way.

Dynamic replication: Perform (HPC) computations on the replicated data. Move (part of) the safely replicated data to a workspace close to powerful machines and move the results back into the archives.

Metadata: A joint metadata domain for all data that is stored by EUDAT data centers by harvesting metadata records for all data objects from the communities.

Simple store : A function that will help researchers mediated by the participating communities to upload and store data which is not part of the officially handled data sets of the community.

PID: a robust, highly available and effective PID system that can be used within the communities and by EUDAT.

AAI: A solution for a working AAI system in a federation scenario.

SAFE_REPLICATION@EUDAT

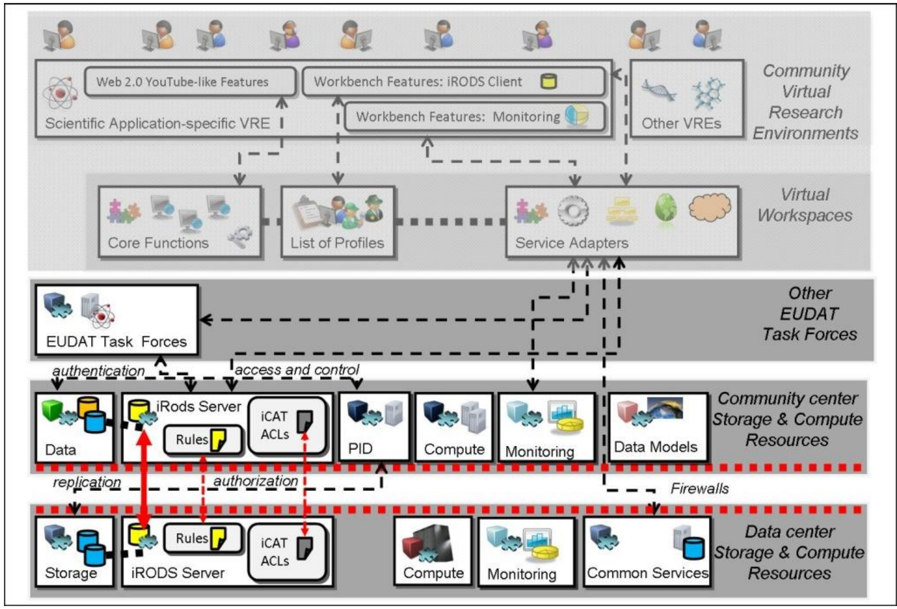
Safe Replication

Objective: Allow communities to replicate data to selected data centers for storage and do this in a robust, reliable and highly available manner.

Description The ability to safely and simply replicate data from one data center to another is essential to EUDAT's task of improving data curation and accessibility.
Several EUDAT user communities (CLARIN, ENES, EPOS, and VPH) have identified safe replication as a common need, and are working to design a blueprint for managing data replication based on users' requirements and constraints

Data replication solutions and services are embedded into critical security policies, including firewall setups and user accounting procedures.

More info: eudat-safereplication@postit.csc.fi

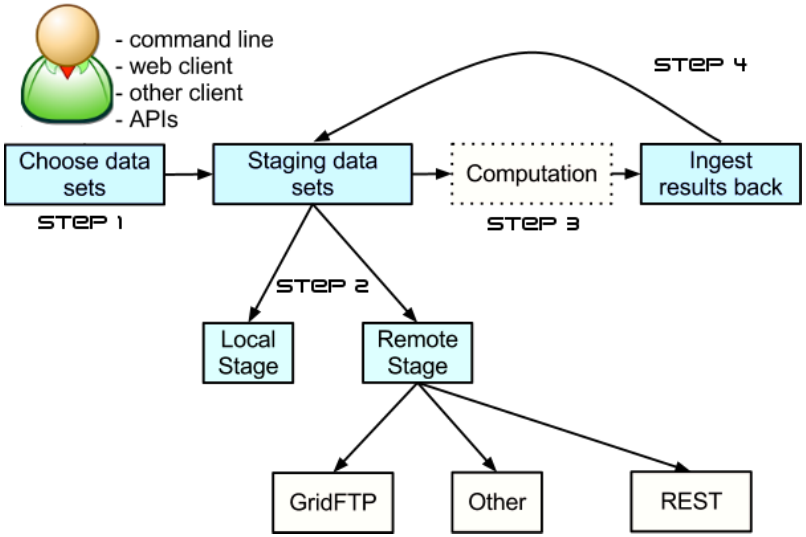


DATA_STAGING@EUDAT

Data Staging

Objective: Allow communities to stage data between EUDAT resources and HPC/HTC resources for computational purposes.

Description*:* This service will allow the communities to dynamically replicate a subset of their data stored in EUDAT to an HPC machine workspace in order to be processed.



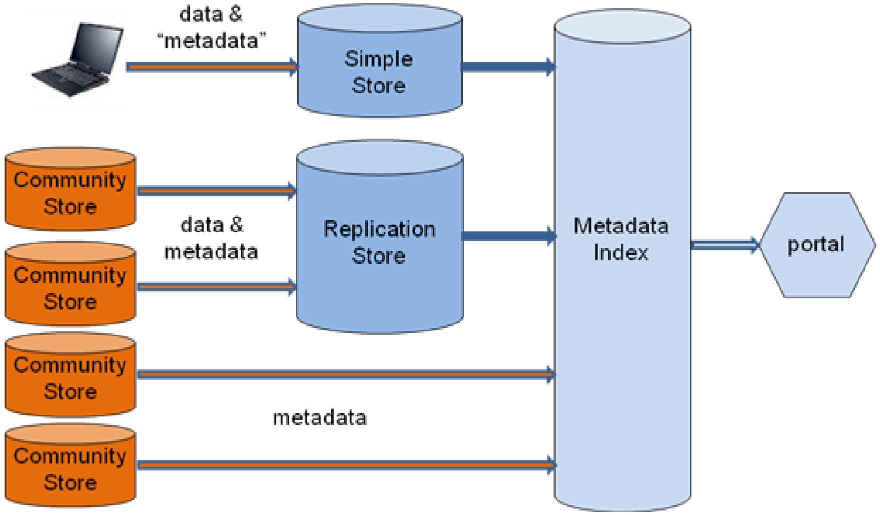
More info: eudat-datastaging@postit.csc.fi

METADATA@EUDAT

Metadata

Objective: Create a joint metadata domain for all data stored by EUDAT data centers and a catalogue which exposes the data stored within EUDAT, allowing data searches.

Description: The EUDAT repository should provide an inventory of metadata from different communities



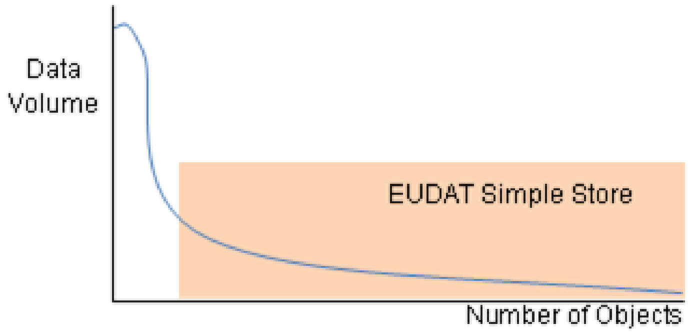
More info: eudat-metadata@postit.csc.fi

SIMPLE_STORE@EUDAT

Simple Store

Objective: Create an easy to use service that will help researchers mediated by the participating communities to upload and store data which is not part of the officially handled data sets of the community.

Description: This service will address the long tail of "small" data and the researchers/citizen scientists creating/manipulating them and NOT the short head of big data.



More info: eudat-simplestore@postit.csc.fi

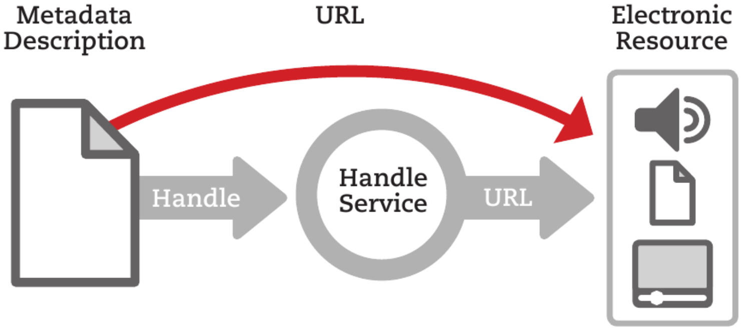
PIDS@EUDAT

Persistent Identifiers

Objective: Deploy a robust, highly available and effective PID service that can be used within the communities and by EUDAT.

Description: Keeping track of the “names” of data sets or other digital artefacts deposited with the CDI requires more robust mechanisms than “noting down the filename”. The PID service will be required by many other CDI services, from Data Movement to Search and Query.

Currently considering use of both EPIC for data objects, and DataCite to register DOIs (Digital Object Identifiers for published collections.



More info: eudat-persistentidentifiers@postit.csc.fi

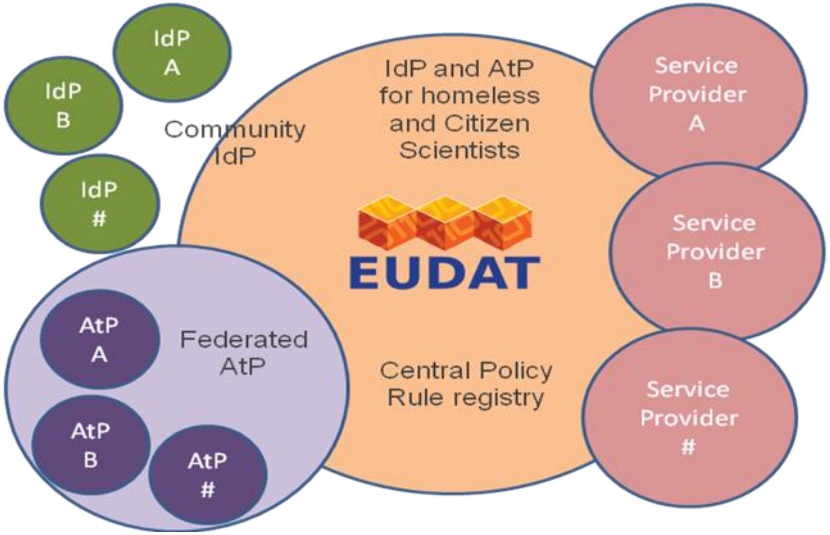
AAI@EUDAT

AAI – Distributed Authentication

Objective: Provide a solution for a working AAI system in a federated scenario.

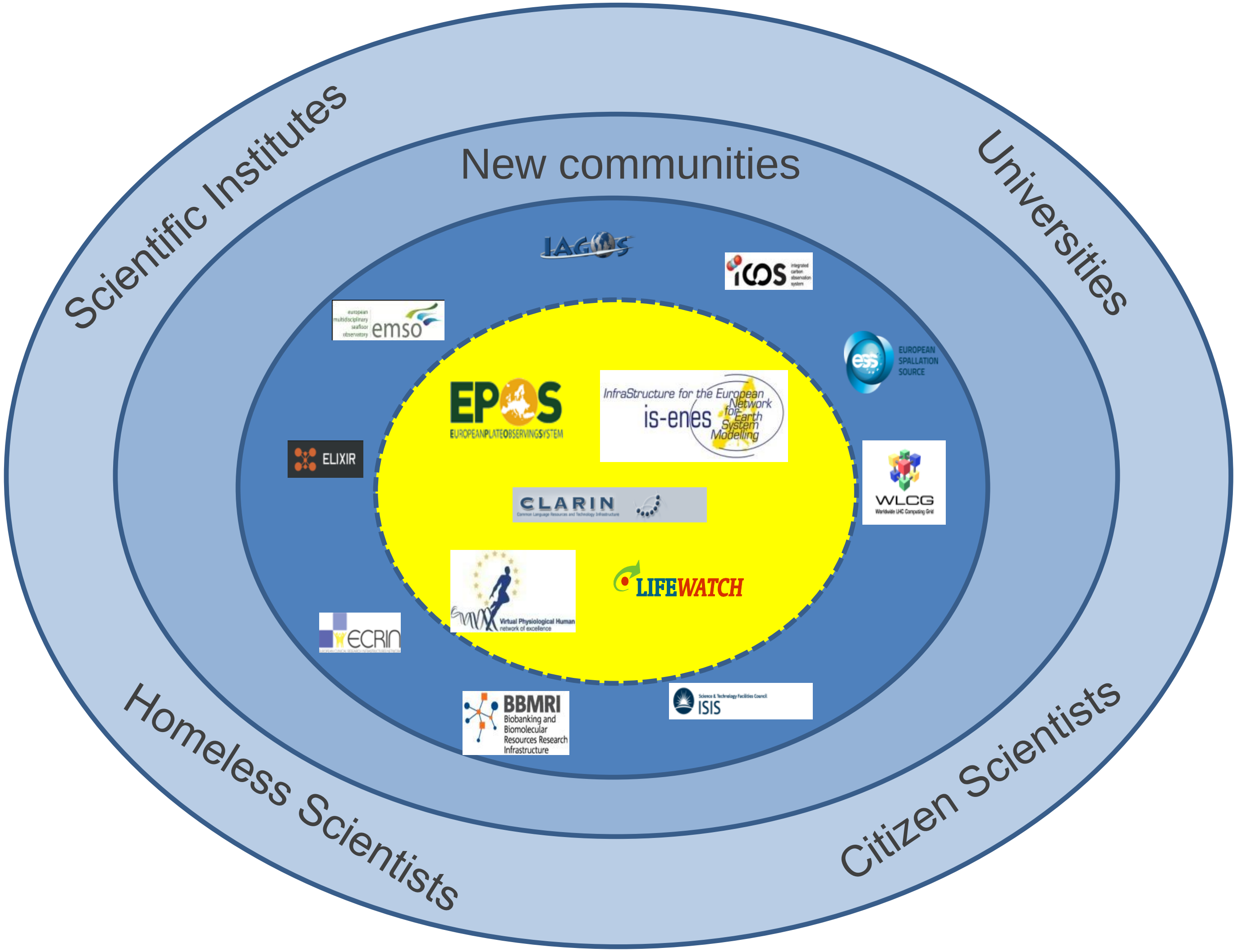
Description: Design the AA infrastructure to be used during the EUDAT project and beyond.

Key tasks:
Leveraging existing identification systems within communities and/or data centers
Establishing a network of trust among the AA actors: Identifty Providers (IdPs), Service Providers (SPs), Attribute Authorities and Federations
Attribute harmonization



More info: eudat-AAI@postit.csc.fi

Background



Scientific Institutes

New communities

Universities

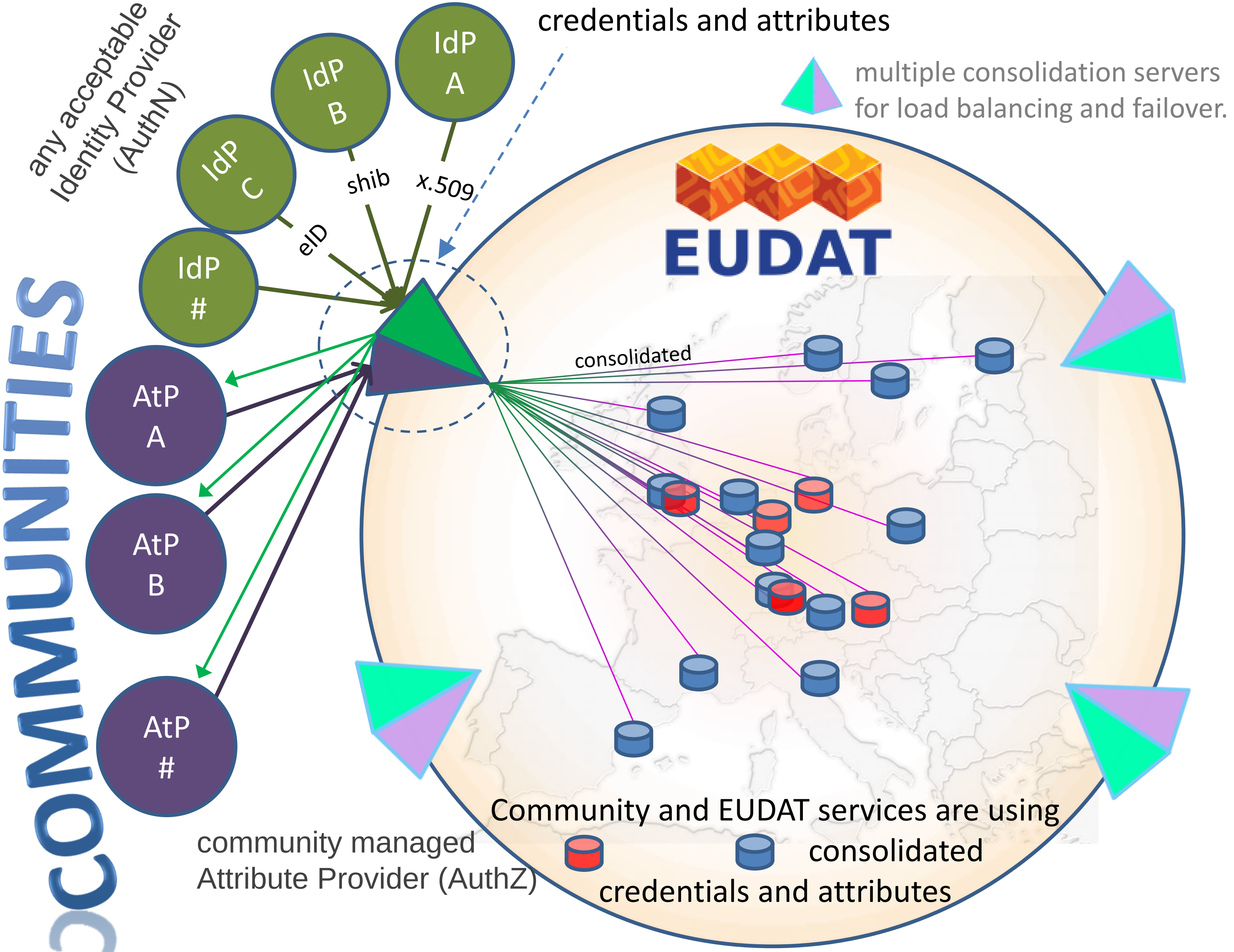
Homeless Scientists

Citizen Scientists

IAGOS
ICOS integrated carbon observation system
european multidisciplinary seafloor observatory emso
EUROPEAN SPALLATION SOURCE
EPOS EUROPEAN PLATE OBSERVING SYSTEM
InfraStructure for the European Network for Earth System Modelling is-enes
ELIXIR
WLCG Worldwide LHC Computing Grid
CLARIN Common Language Resources and Technology Infrastructure
Virtual Physiological Human network of excellence
LIFEWATCH
ECRIN
BBMRI Biobanking and Biomolecular Resources Research Infrastructure
Science & Technology Facilities Council ISIS

AA process’actors

1. Federations
2. Multiple IdPs (e.g. home institute IdP)
   - Provision for supporting “homeless” users, cf SWITCH
   - Attributes from home institute
   - Technology – IdPs should use the same technology
3. Attribute authorities
   - Attributes relating to collaborations/communities (e.g. roles, memberships)
   - Each community should be prepared to manage and publish the user attributes
4. Multiple service providers
   - All consuming the *same* identities and attributes
   - Single Sign on: single IdP



consolidation (conversion) of credentials and attributes

multiple consolidation servers for load balancing and failover.

community managed Attribute Provider (AuthZ)

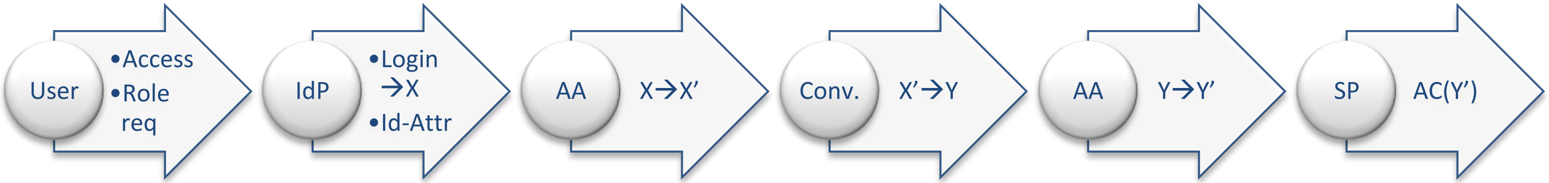
Community and EUDAT services are using 🟥 🟦 consolidated credentials and attributes

Assumptions, Statements

- The IdP is an issuer of any kind of acceptable identity credential (x.509, shibboleth/saml2, card based eID, OpenID, credentials from social networks)
- Communities are assumed to manage their AtP (but they can offer a IdP services too if needed)
- AtPs can make use of the consolidated identity credentials to map their attributes (roles) to identities (green arrows indicate the usage of consolidated credentials).
- The credential/attribute conversion service is a gateway to EUDAT services which must be high-available. Therefore this service should be distributed over more than one server (load balancing, failover).
- The conversion service must be safe and trustworthy. Domains of trust can be fragmented (although they are encourages to collaborate). As a possible solution, specific centres could offer their conversion service for „their“ affiliated communities (and service providers).
- The AtP of the „community“ of (homeless) citizen scientists can be managed by EUDAT.
- Citizen scientists should be able to use any acceptable means for AuthN (including the eID on their national ID card)
- The EUDAT services need to build a trusted connection only to these credential consolidation gateway. No need to maintain large distributions of (e.g. IGTF) CA certificates etc. at the SP side.

AA process: general overview

Some steps are of course optional



User
- Access
- Role req

IdP
- Login →X
- Id-Attr

AA X→X’

Conv. X’→Y

AA Y→Y’

SP AC(Y’)

Technology

- Within the federation
  - **Shibboleth** (Web), **Moonshot** (Non Web)
  - eduRoam (based on **RADIUS**: Remote Authentication Dial-In User Service)
  - **User certificates**, **OpenId**
  - **XACML** (eXtensible Access Control Markup Language)
  - **Oauth2** (Google, Facebook, Microsoft)
- Outside the federation
  - Credential conversion: special SP to create “external” credential

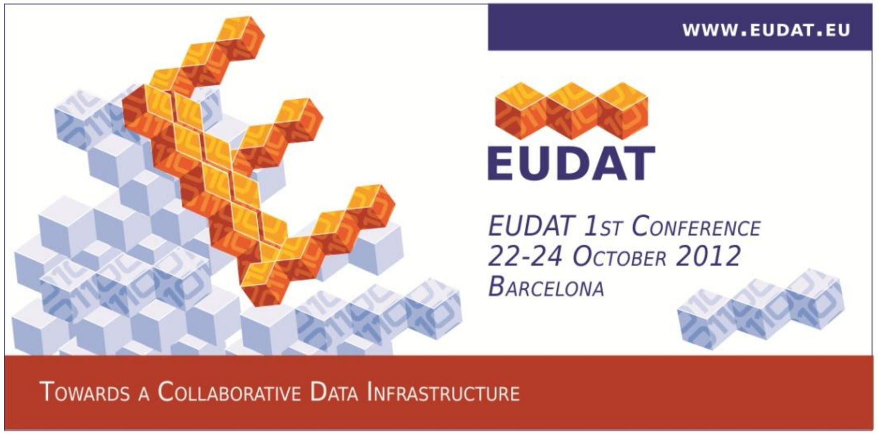
Challenges

- Leveraging existing identification systems

- Establishing a network of trust among the AA actors: IdPs, SPs, Attribute Authorities, Federations

- Attributes harmonization: it is necessary to agree on a common way to interpret different set of attributes.

Welcome to the 1st EUDAT Conference!



WWW.EUDAT.EU

EUDAT

EUDAT 1st Conference
22-24 October 2012
Barcelona

Towards a Collaborative Data Infrastructure

22-24 October 2012, Barcelona

- International event with keynotes from Europe and US
- A forum to discuss the future of data infrastructures
- Project presentations and poster sessions
- 2nd EUDAT User Forum
- Training tutorials

Welcome to the 1st EUDAT Training Days

Building Blocks of Data Infrastructures 1 , 25-26 June 2012. Amsterdam

•25 June (12pm-6pm): Policy-Rule based Data Management

•26 June (9am-11am): Use of Handles (EPIC, DataCite) for Persistent Identification

•26 June (11:30am-3pm): Distributed Authentication and Authorization

EUDAT