Views on data services

Towards a pan-European Collaborative Data Infrastructure

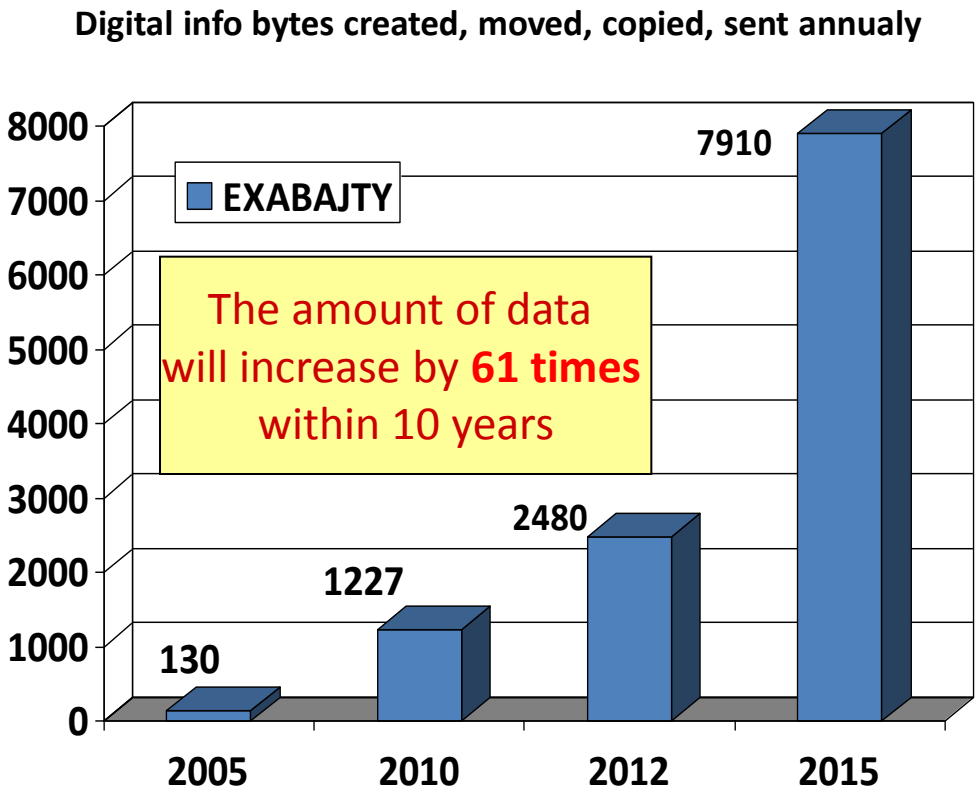
Norbert Meyer
Poznan Supercomputing and Networking Center

EUDAT

e-IRG
e-Infrastructure
Reflection Group

The Digital Universe

- 1,987 ZBytes generated since 01.01.2011
- 1,987 Zeta bytes = 1.987.000.000.000.000.000.000 bytes ….
- 2012 - 2,5 ZB (doubled within 12 months)
- 60+ % data lost due to missing hardware capacity …..



*) source : „The 2011 IDC DIGITAL UNIVERSE STUDY sponsored by ECM2"

EUDAT Key facts

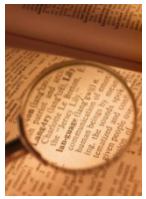
Project Name	EUDAT – European Data
Start date	1st October 2011
Duration	36 months
EC call	Call 9 (INFRA-2011-1.2.2): Data infrastructure for e-Science (11.2010)
Participants	25 partners from 13 countries (national data centers, technology providers, research communities, and funding agencies)
Objectives	"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"

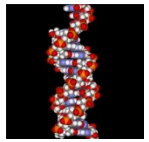
The current data infrastructure landscape: challenges and opportunities

- 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.)

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

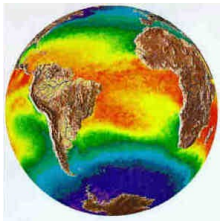




Challenges

- 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**

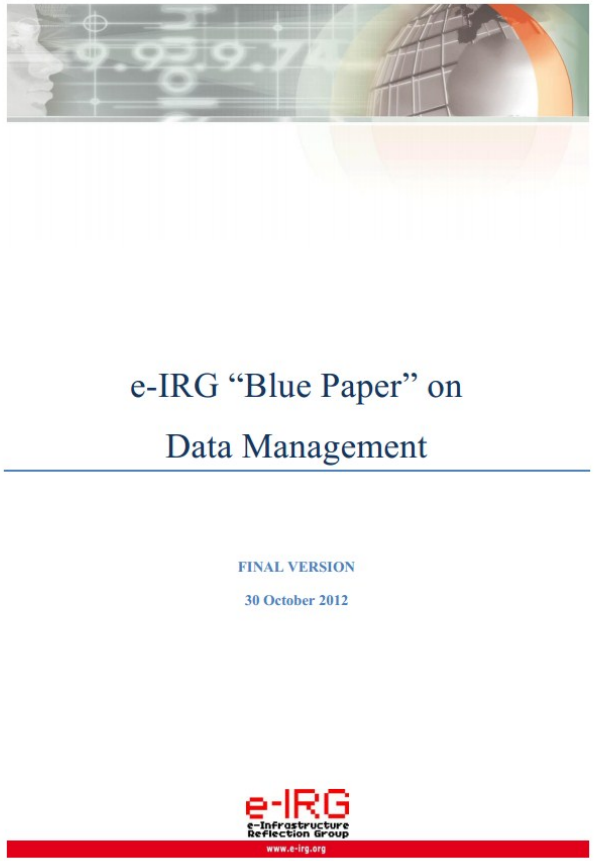
EUDAT

Blue Paper

Invited ESFRI cluster projects

BioMedBridges, DASISH, ENVRI, CRISP
(pilot projects)

also DC-NET, PaNdata, ITER



Importance of requirements – e-irg Blue Paper

Large datasets
Restricted access
Metadata structure
Federated AAI
Accounting
Data provenance
Integration
Interoperation
High trust and security
Reliability

Access
Advanced search functionality
Data preservation
Interpretation of data
Unified access
High quality
Simplified access
Searching information
Important stakeholders

EUDAT Core Service Areas

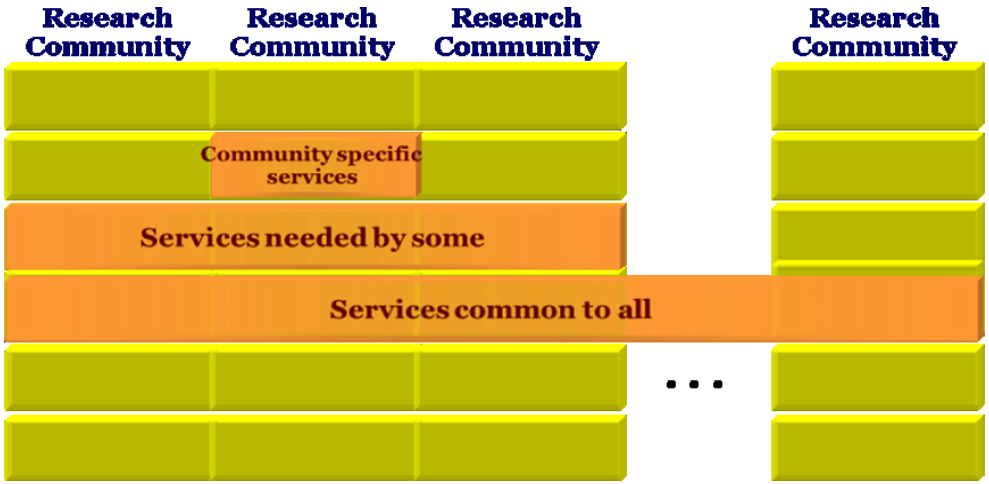
Community-oriented services

- 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



Building the services

6 services/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.

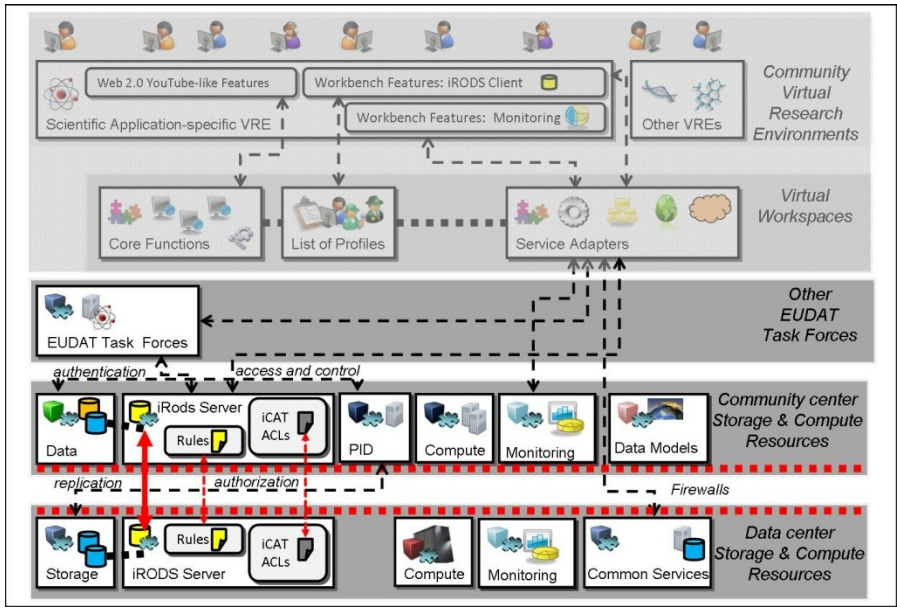
AAI: A solution for a working AAI system in a federation scenario.

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.

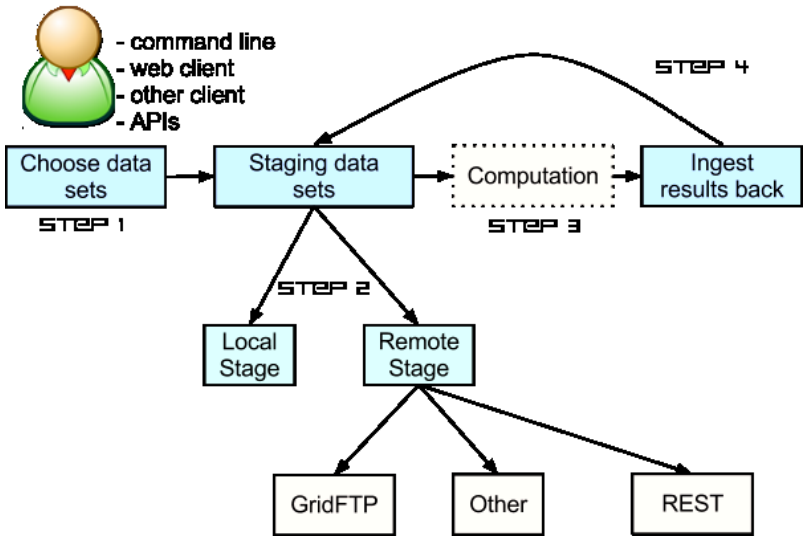
More info: eudat-safereplication@postit.csc.fi

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.



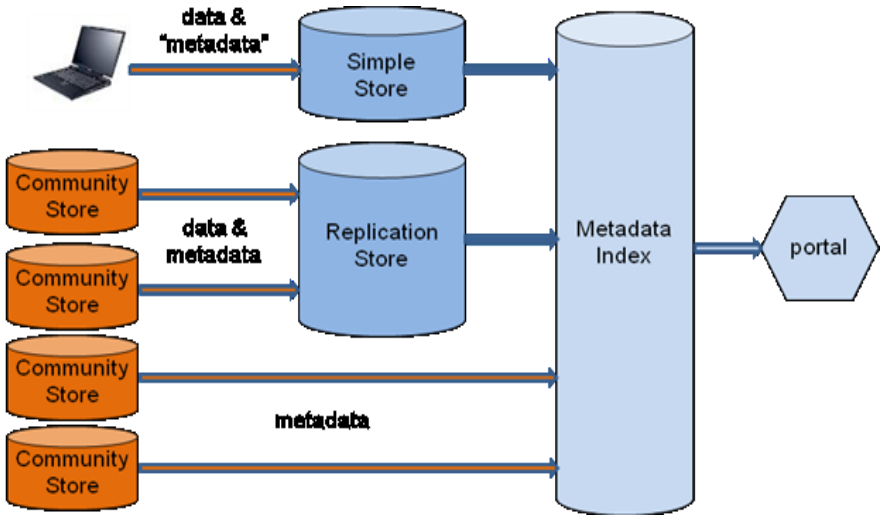
More info: eudat-datastaging@postit.csc.fi

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



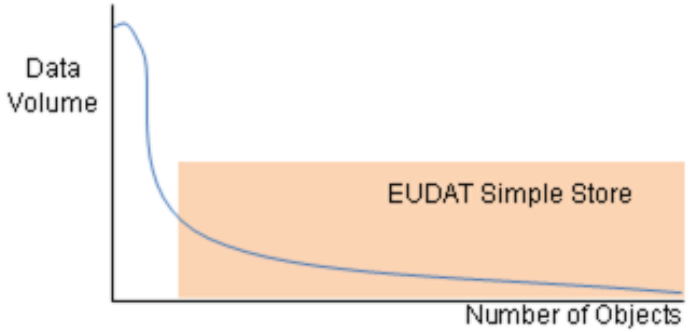
More info: eudat-metadata@postit.csc.fi

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

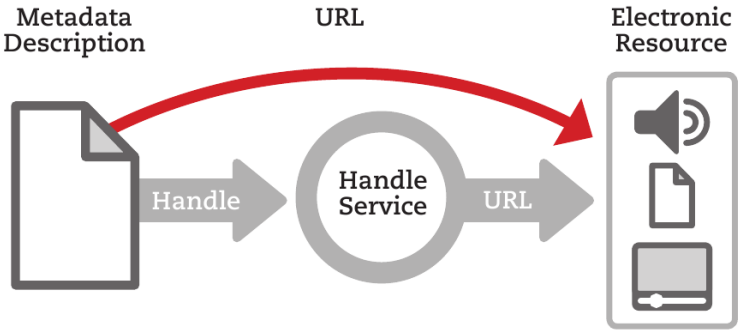
PIDS@EUDAT

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

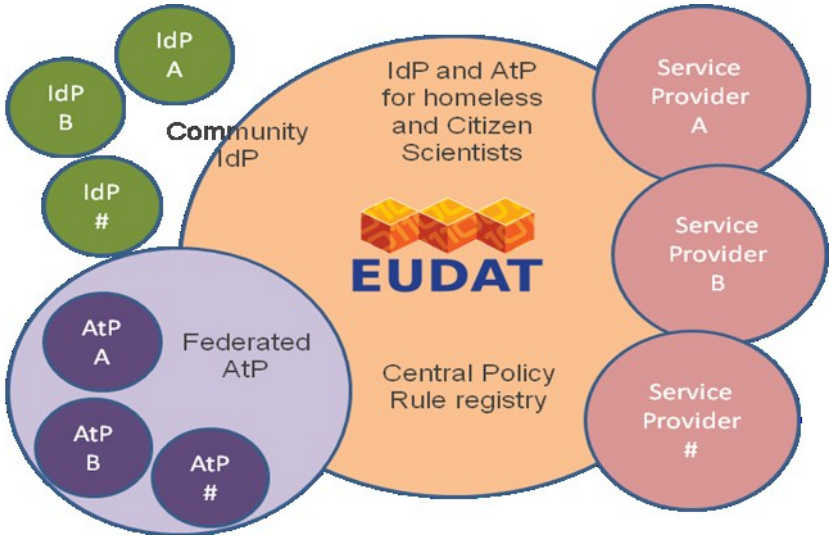
AAI@EUDAT

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

Expected benefits of the CDI

- **Cost-efficiency through shared resources and economies of scale**
  - Better exploitation of synergies between communities and service providers
  - Support to existing scientific communities' infrastructures and smaller communities
- **Trans-disciplinarity**
  - Inter-disciplinary collaboration
    - Communities from different disciplines working together to build services
    - Data sharing between disciplines – re-use and re-purposing
    - Each discipline can solve only part of a problem
- **Cross-border services**
  - Data nowadays distributed across states, countries, continents, research groups are international
- **Sustainability**
  - Ensuring wide access to and preservation of data
    - Greater access to existing data and better management of data for the future
    - Increased security by managing multiple copies in geographically distant locations
  - Put Europe in a competitive position for important data repositories of world-wide relevance