Data Replication: Automated move and copy of data

Digital Preservation Advanced Practitioner Course
Glasgow, July 19th 2013

c.cacciari@cineca.it

SEVENTH FRAMEWORK
PROGRAMME

EUDAT

Outline

- "The" issue
- Starting point
- Which kind of service?
- Which kind of users?
- Flexibility
- Different transfer strategies
- Policies
- Performances
- Different federation strategies
- PID and registered data

Data trends: the challenges

Zettabytes
Exabytes
Petabytes
Terabytes
Gigabytes

Exponential growth

Increasing complexity and variety

- Where to store it?
- How to find it?
- How to make the most of it?

- **How to ensure service interoperability?**

Collaborative Data Infrastructure

- To move data across heterogenous systems and organizations, the keys are

  - **interoperability** at organisational, semantic and technical levels,

  - **scalability** and reduction of the complexity of the data management
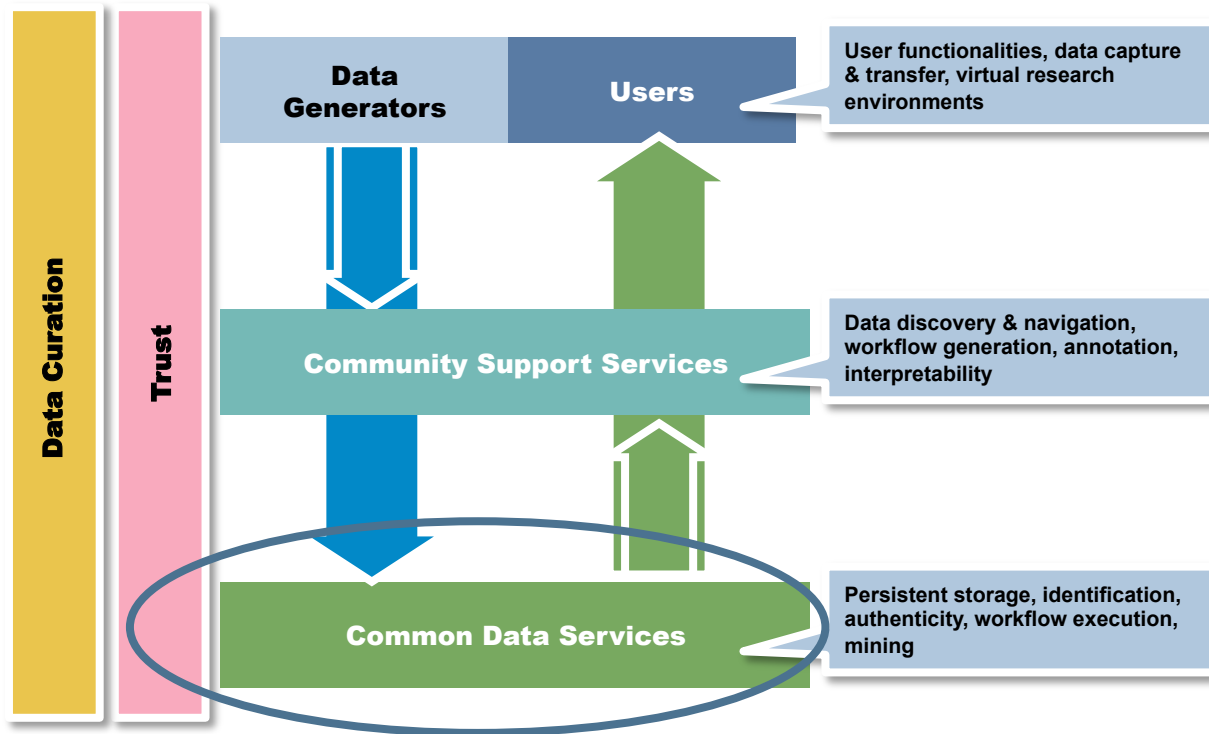
Just another infrastructure?

If there are hundreds of Research Infrastructures, how many different data management systems can we sustain?
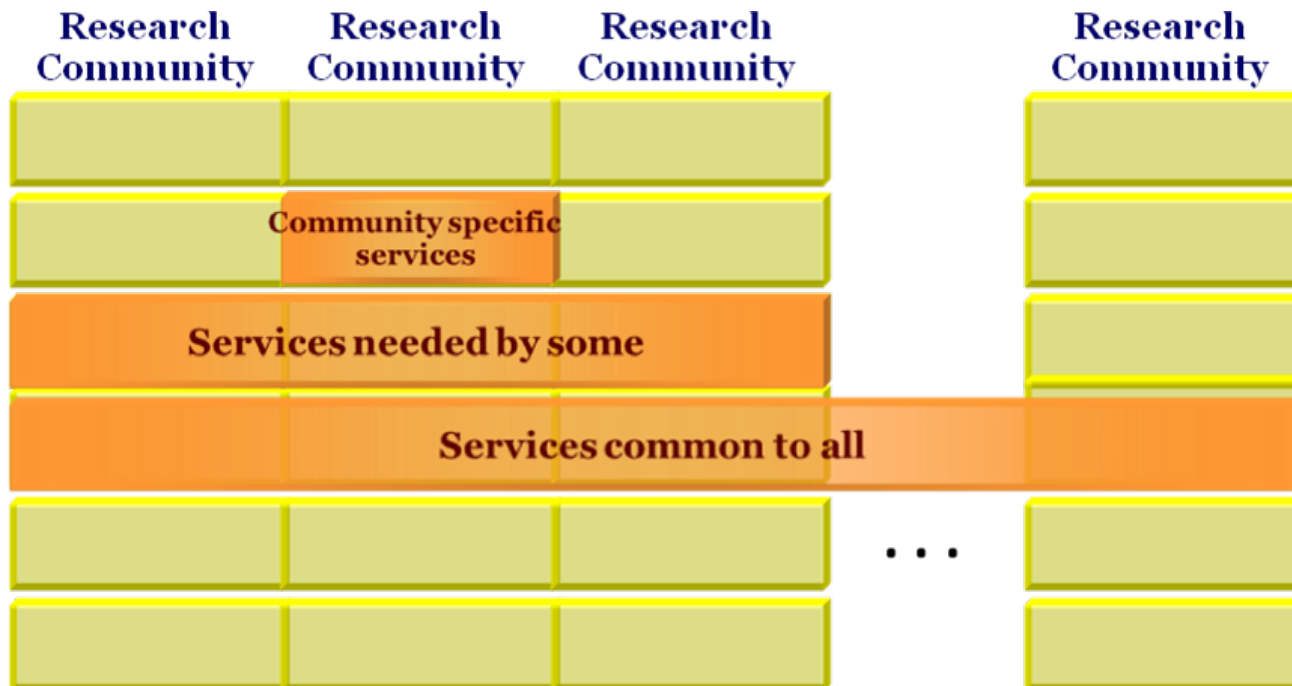
EUDAT project

- EUDAT is a three-year project that will deliver a **Collaborative Data Infrastructure (CDI)**
- Co-funded within the 7th Framework Programme
- Launched on 1 October 2011

25 European partners

Principles – where we want to be

1. Data deposited will be preserved in perpetuity

2. Data are best curated in their own communities

3. Access to data in the Collaborative Data Infrastructure is free at the point of use

4. The Collaborative Data Infrastructure will not assert ownership of any data it holds

EUDAT

Data movement: staging or replication?

Safe Replication
Data preservation and access optimization

Data Staging
Dynamic replication to HPC workspace for processing

Data

EUDAT

Replication

- **Safe Replication** to enable communities easily create replicas of their scientific datasets in multiple data centres for improving data curation and accessibility
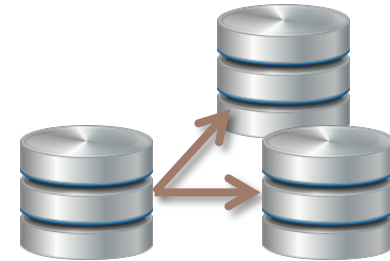
- **Data Staging** to facilitate communities to stage stored data onto external computational facilities, such as HPC resources

Once upon a time …

replicate my collection X to three data centres and store the collection safely for 10 years

EUDAT

Are you talking about clouds?

I already do it every day in my cloud space

We are talking about a …

robust

safe

highly available

Replication Service

… which is not a personal cloud space

EUDAT

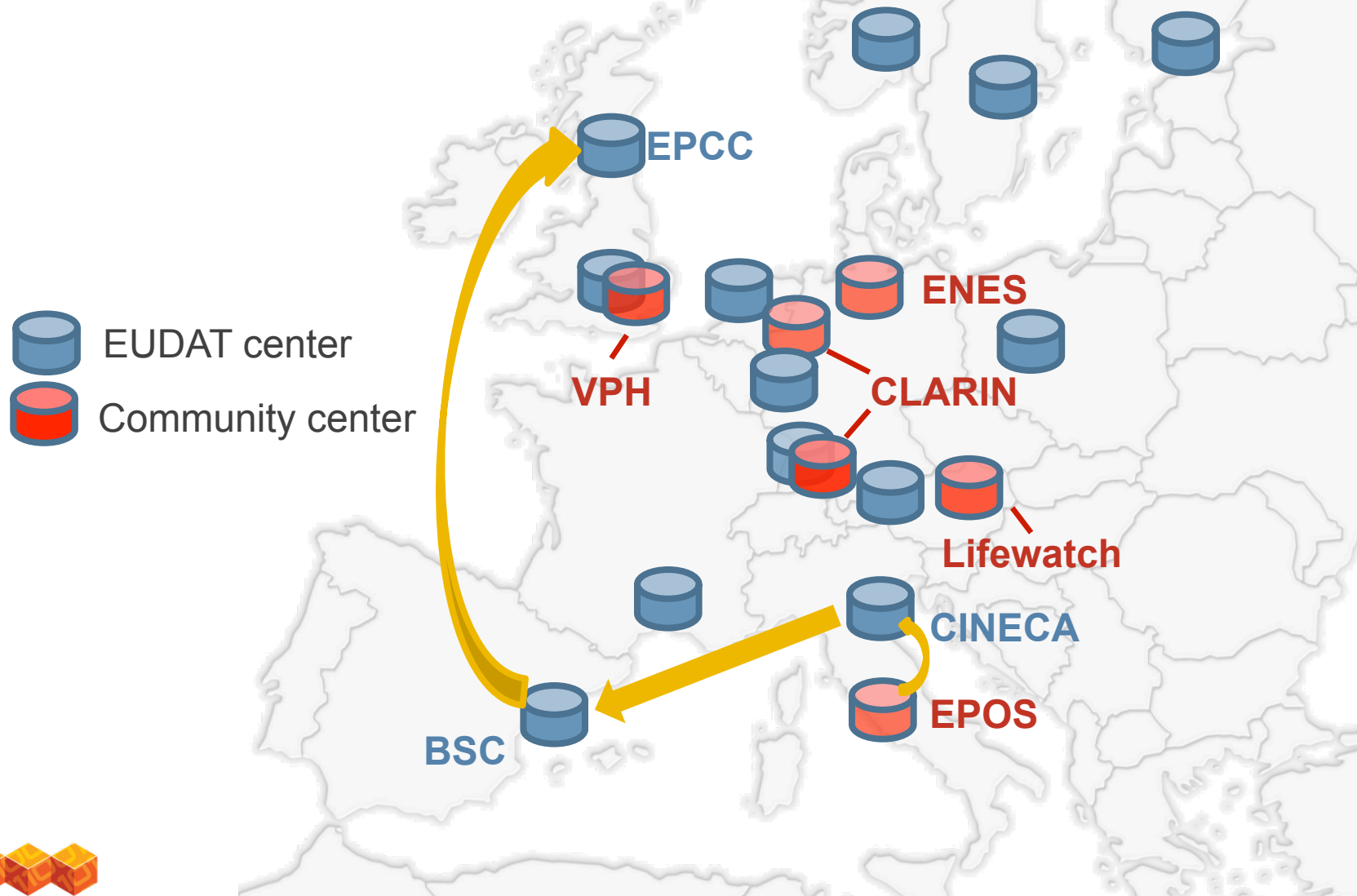# What about trust?

Can you find where your data are physically stored on the cloud?

Or who can access them?

No, because clouds are opaque

While a collaborative data infrastructure is transparent

http://www.flickr.com/photos/andercismo/2349098787

EUDAT

replicate my collection X to three data centres

EUDAT center

Community center

EPCC

ENES

VPH

CLARIN

Lifewatch

CINECA

BSC

EPOS

EUDAT

Then is it a complex mechanism suited only for expert data managers?
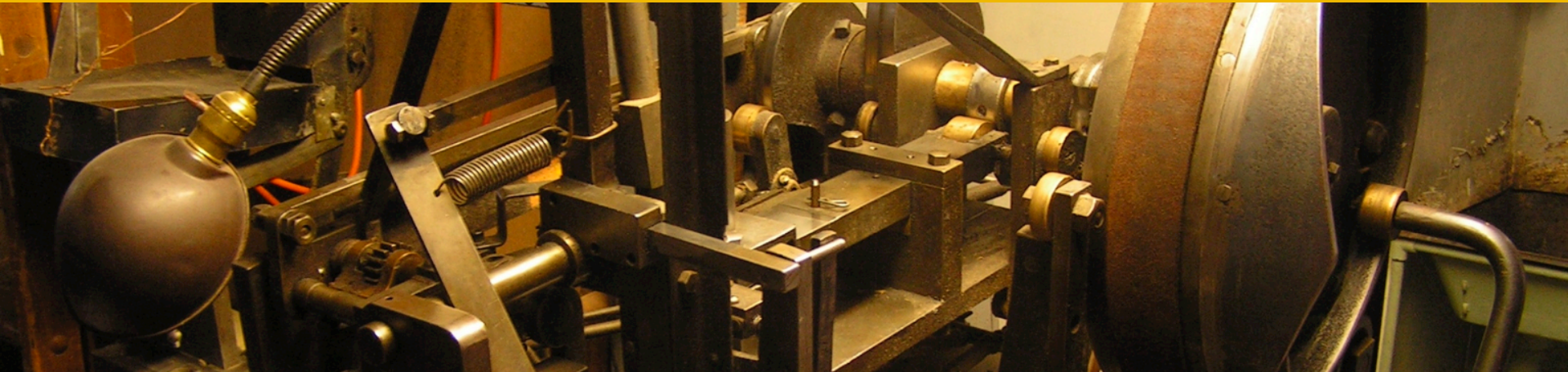
No.

It is a flexible approach able to encompass quite different usage scenarios.

EUDAT

Horizontal flexibility

Single researcher

Team

Community

Upload and forget

Upload, add metadata, share
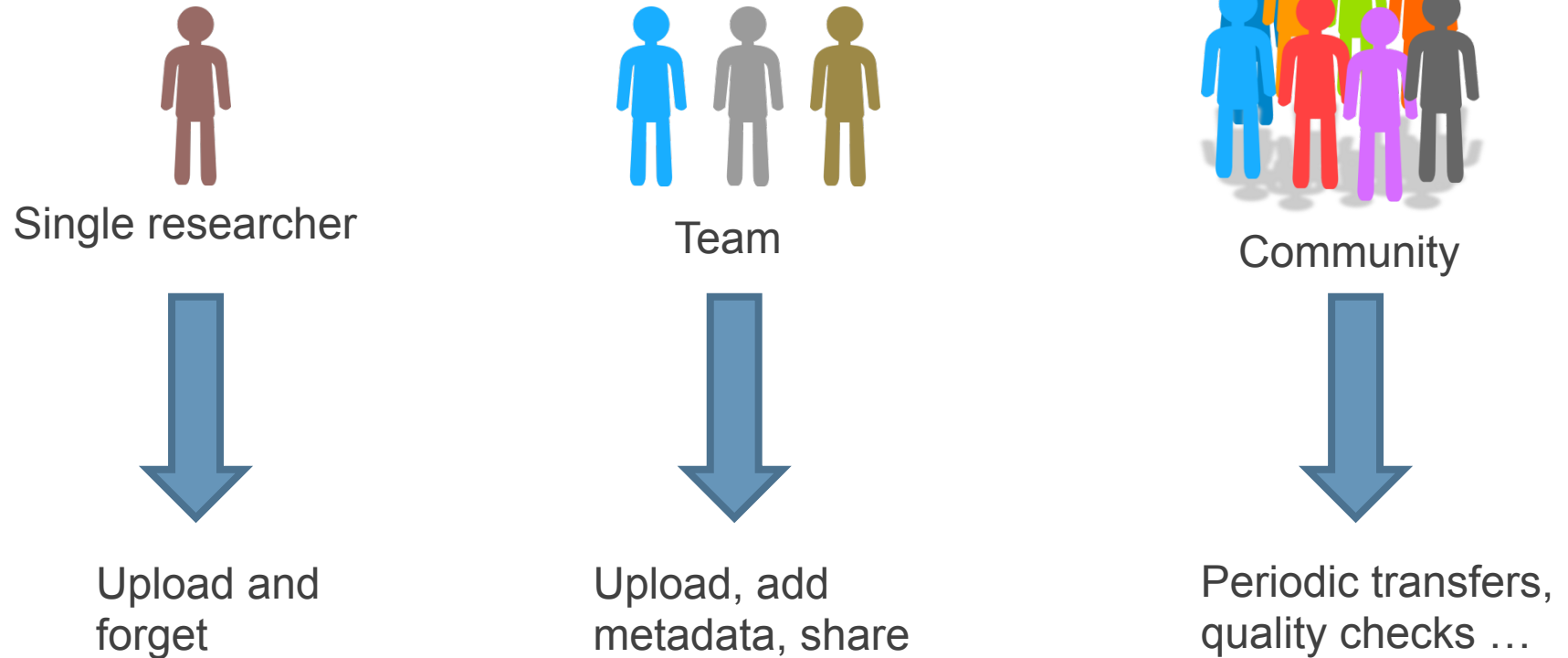
Periodic transfers, quality checks …

Different strategies for different usage scenarios

EUDAT service overview

Metadata Catalogue
Aggregated EUDAT metadata domain. Data inventory

Safe Replication
Data preservation and access optimization

Data Staging
Dynamic replication to HPC workspace for processing

Simple Store
Researcher data store (simple upload, share and access)

AAI
Network of trust among authentication and authorization actors

PID
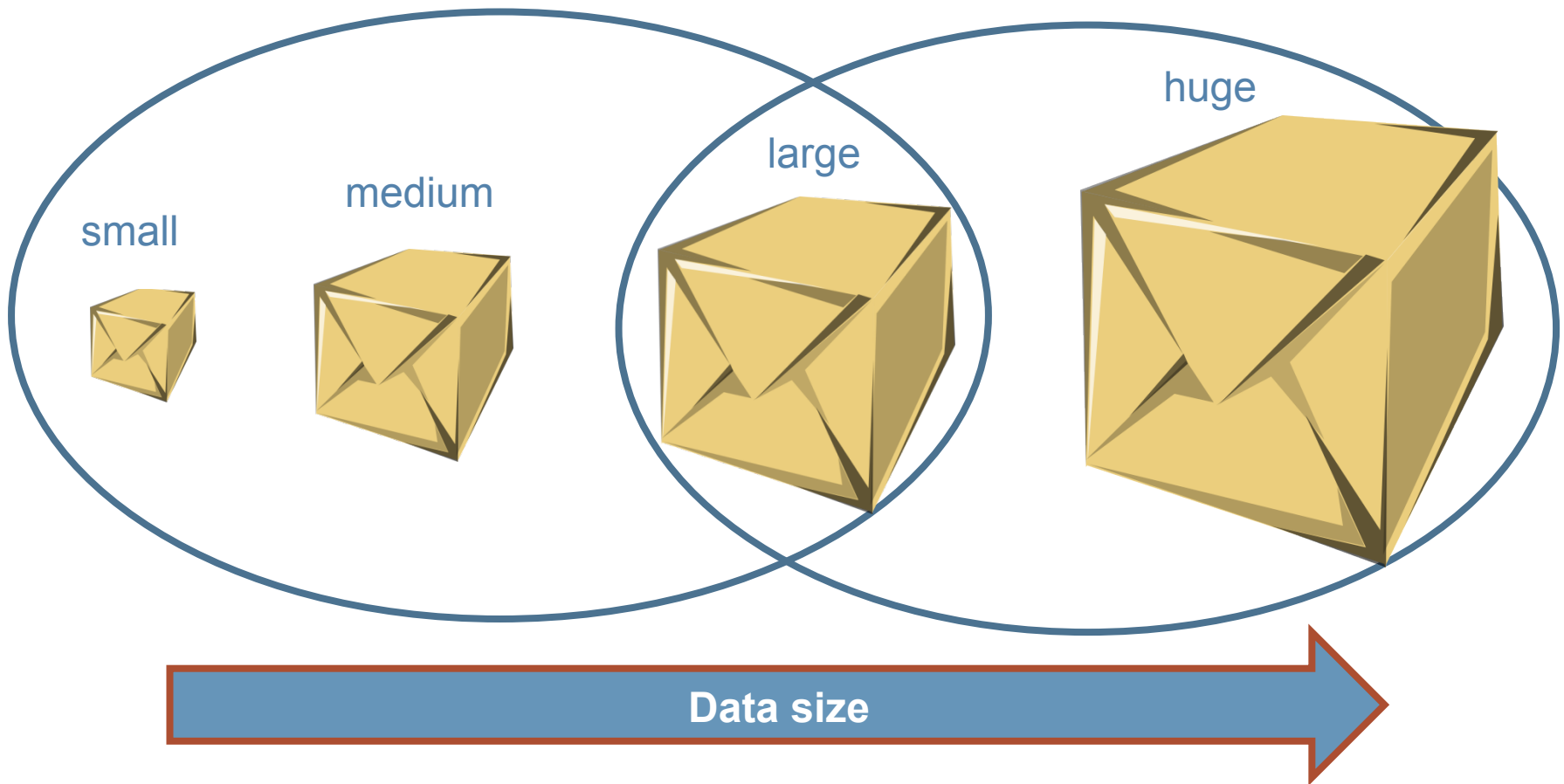Anchor for identifica-tion and integrity

Data   PID   Metadata

Service availability, does not mean user awareness

Transfer approaches

Coming back …

replicate my collection X to three data centres and store the collection safely for 10 years

Apparently a simple statement

But you need to plan it, then you need …

Policies !

EUDAT

Vertical flexibility

iRODS Data System Components

User Interface
Web or GUI Client to Access and Manage Data & Metadata*

iRODS Server
Data on Disk

iRODS Rule Engine
Implements Policies

iRODS Metadata Catalog
Database Tracks state of data

rule-client

EUDAT

EUDAT consortium is working on

Policies for the Collaborative Data Infrastructure management
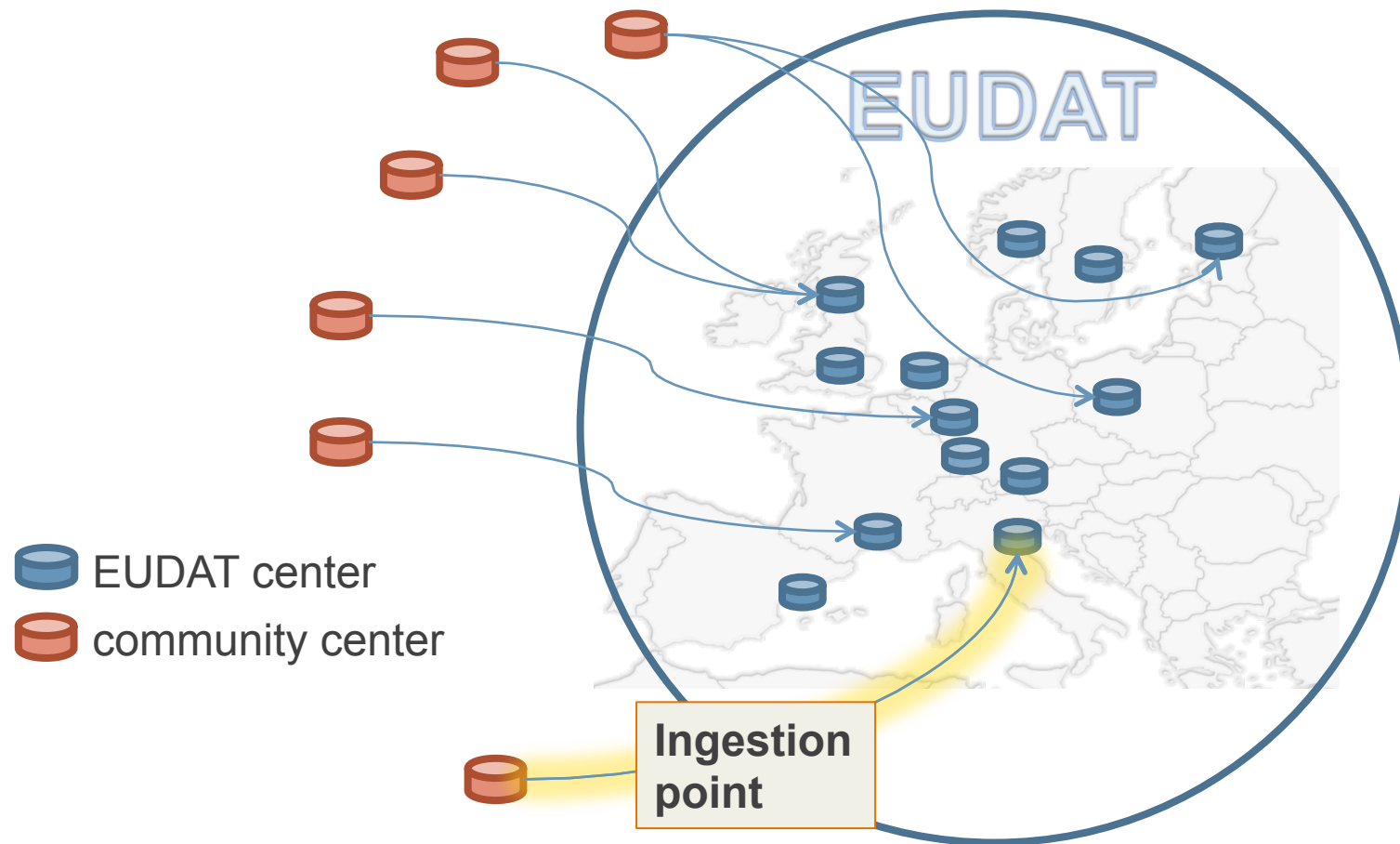
We can call them "internal policies"

EUDAT

How to manage software updates?

What about security bugs?

Which systems are monitored?

Internal shared policies

Each community has one or more doors to connect to the infrastructure

EUDAT center

community center

Ingestion point

replicate my collection X to three data centres and store the collection safely for 10 years …

Updating the sub-collection X_1 weekly

And the sub-collection X_2 hourly

And keeping on-line the data uploaded during the last six months

community policies

EUDAT

So far so good, we have our infrastructure, our policies

What is missing?

Performances?

EUDAT

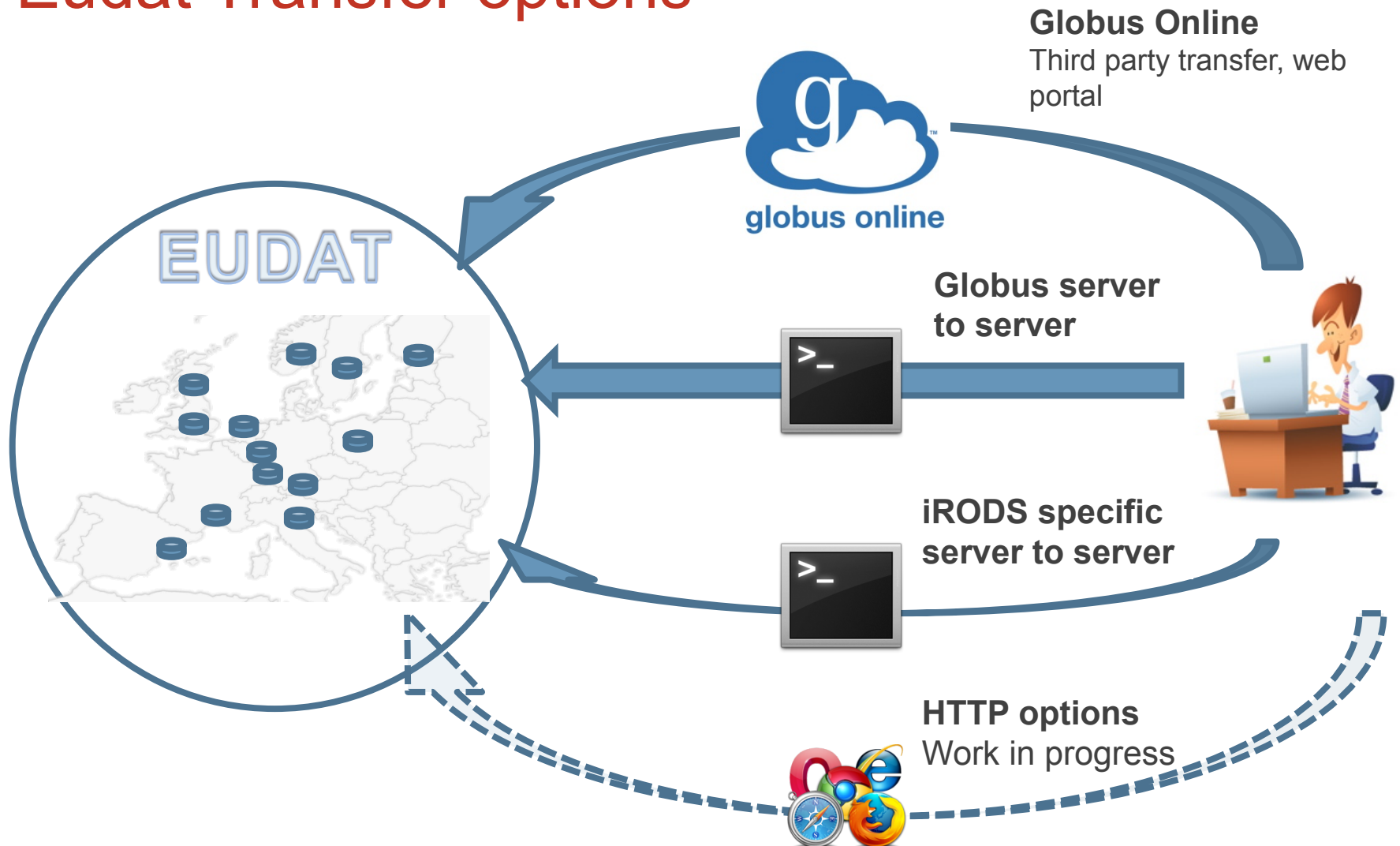If you produce 1 TB of data per day and you want to store it remotely

And it takes one week to move the data

Then any policy is useless
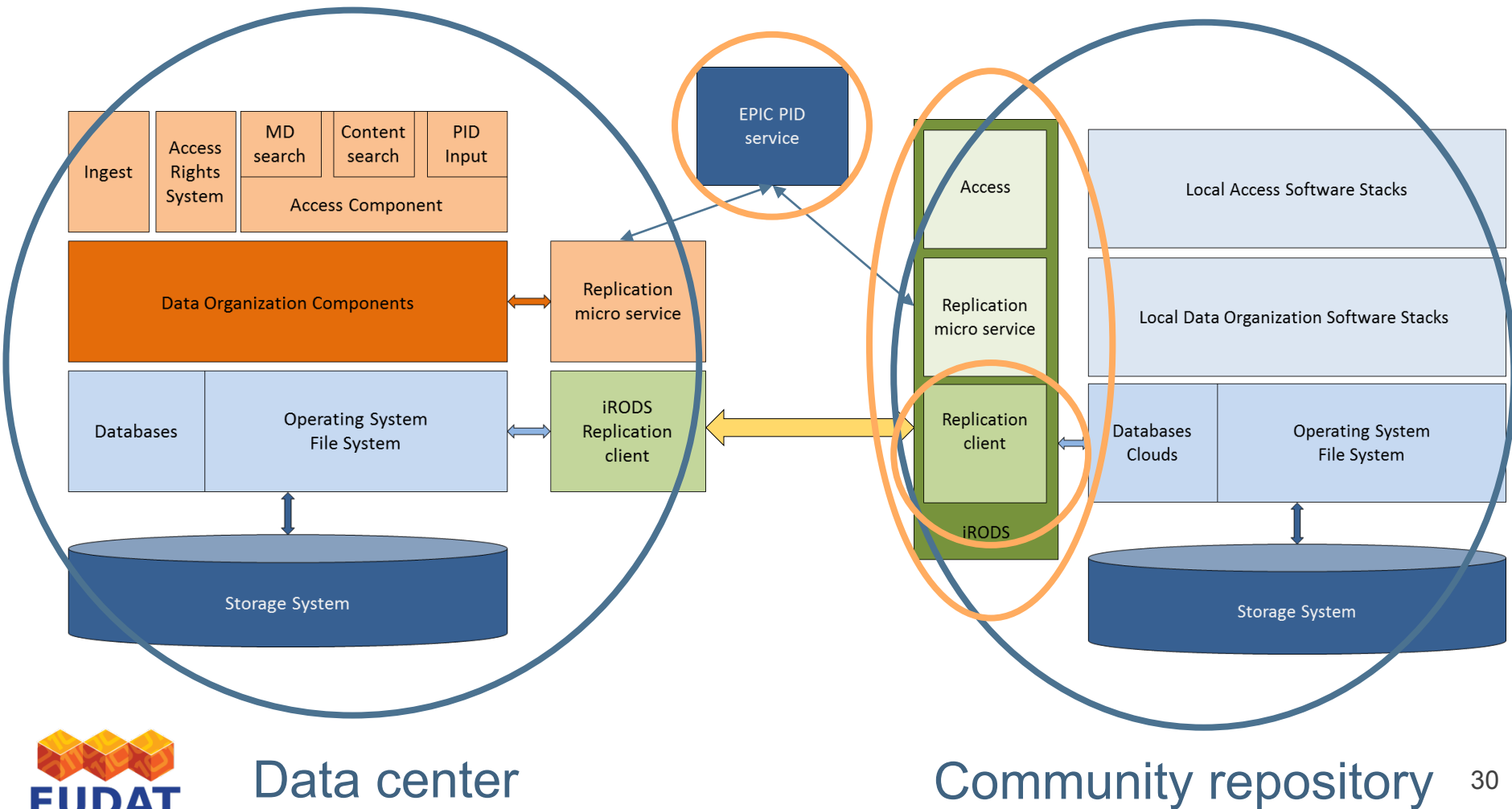
EUDAT

High Performance Transfers

Eudat Transfer options

Globus Online
Third party transfer, web portal

Globus server to server

iRODS specific server to server
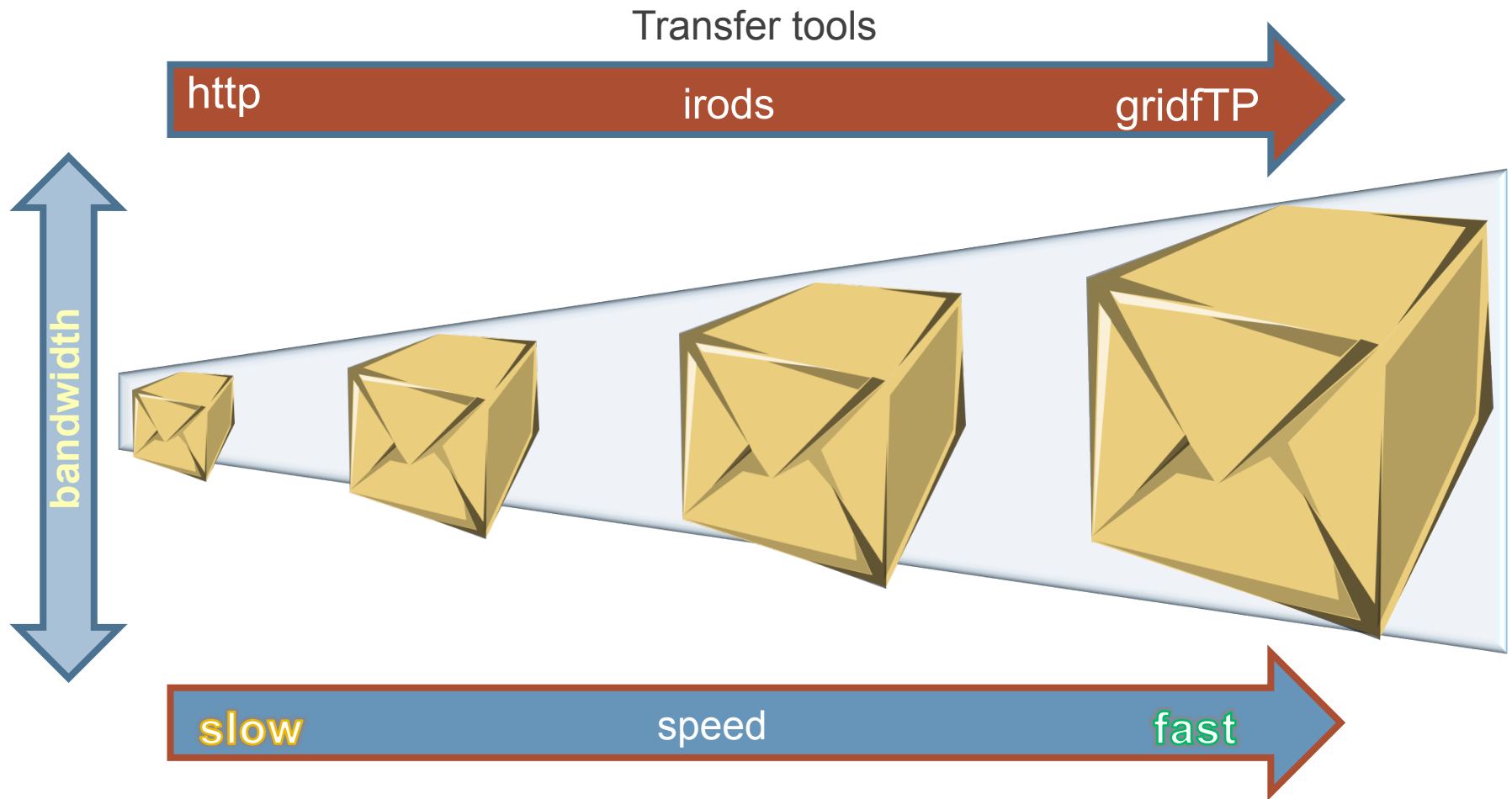
HTTP options
Work in progress

To join or not to join

Join: automated data movement server to server

Data center   Community repository
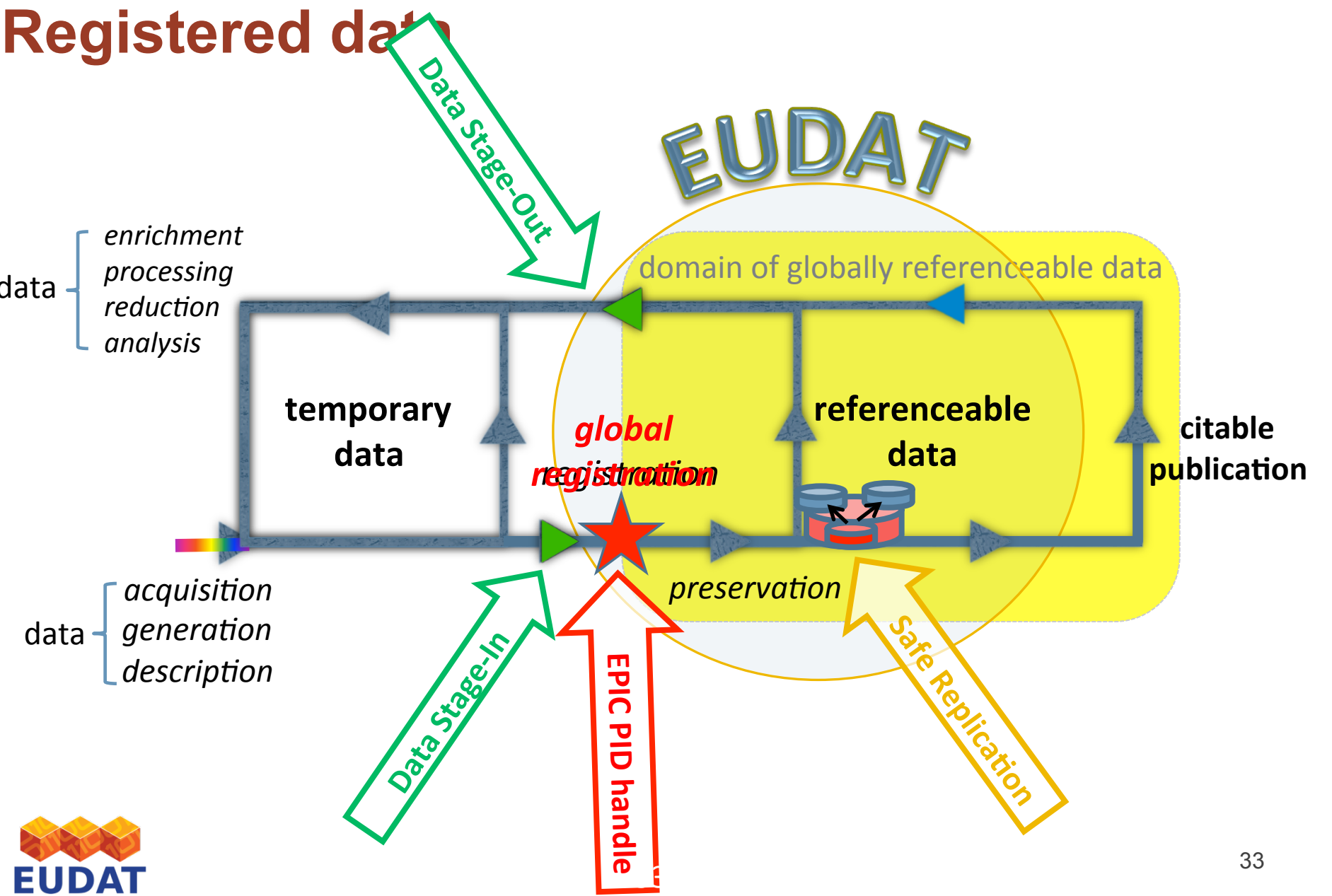
Persistent Identifiers (PID) – definition …

- See the previous presentation "Persistent Identifiers" by Maurizio Lunghi for the general concept.

- EUDAT relies on the **EPIC service** to associate persistent identifier to digital objects.

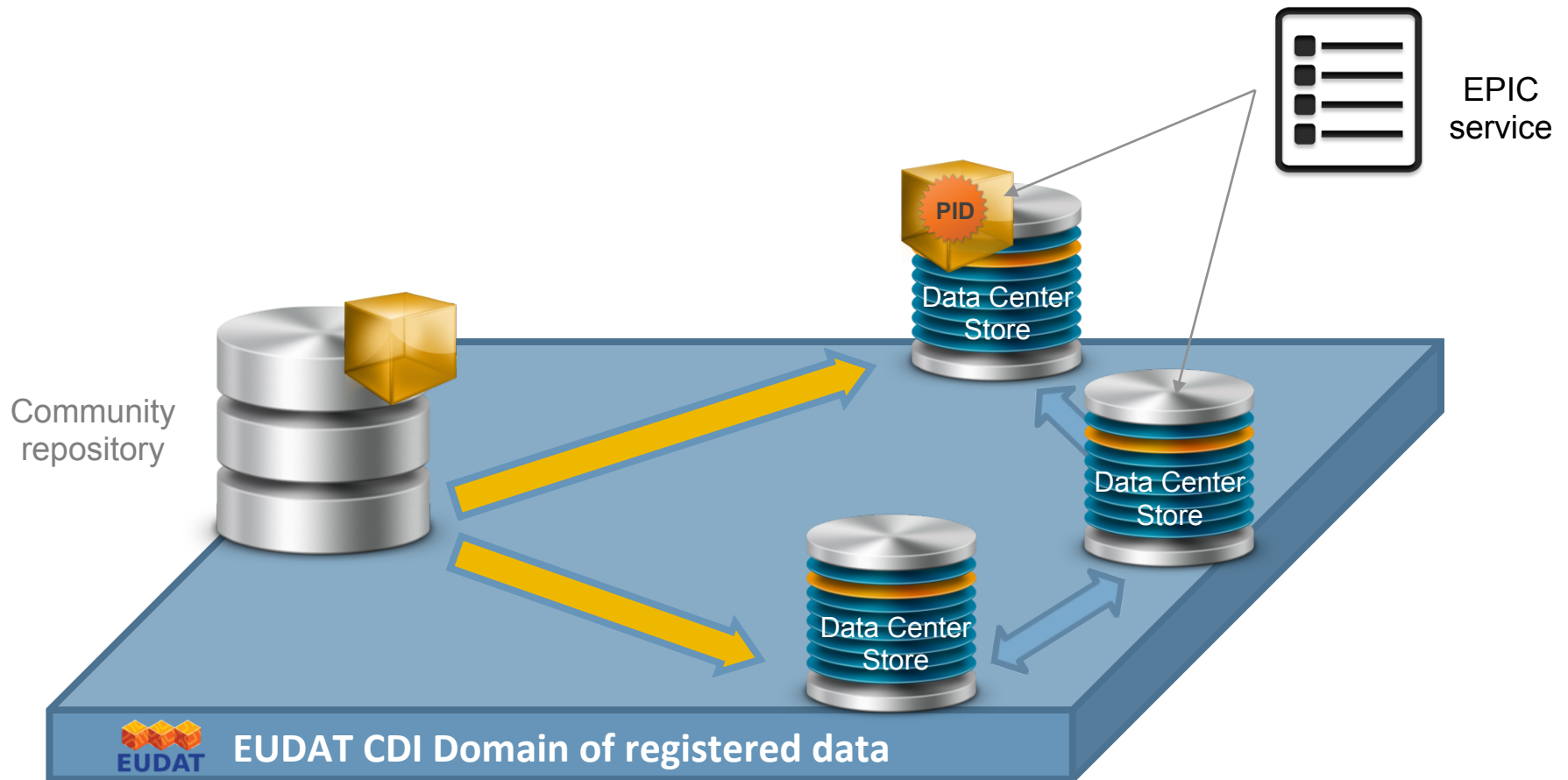- EPIC is an identifier system using the **Handle infrastructure.**
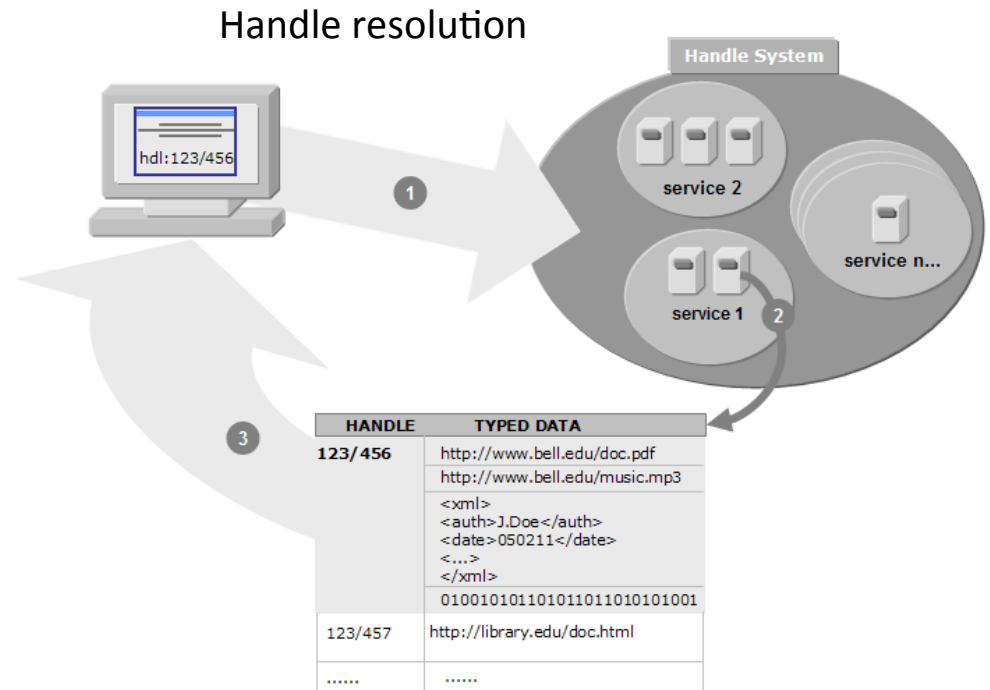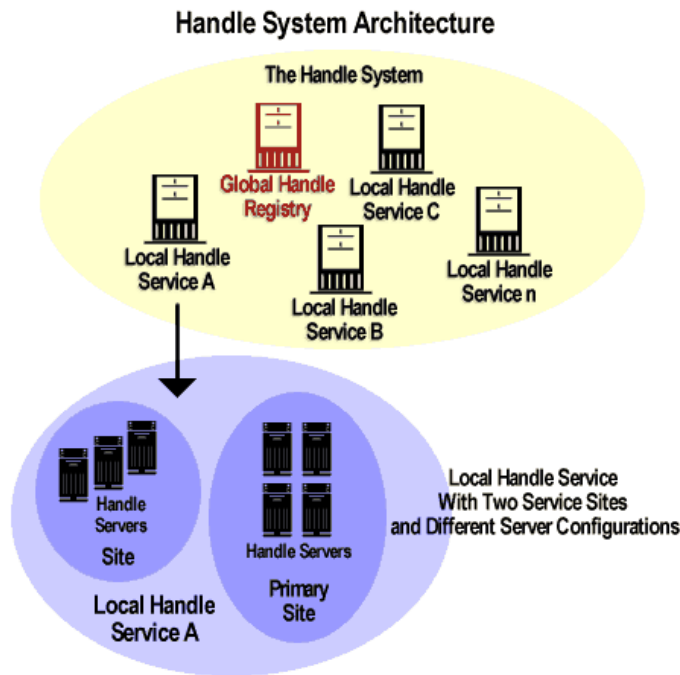
… and focus

- Its focus is the registration of data in an early state of the scientific process, where lots of data is generated and has to become referable to collaborate with other scientific groups or communities, but it is still unclear, which small part of the data should be available for a long time period.

Registered data

Safe replication explained

EPIC service

PID

Data Center Store

Data Center Store

Data Center Store

Community repository

EUDAT CDI Domain of registered data

Finally, all together

The set up of the automated data transfer between EPOS community and EUDAT touched all the aspects we mentioned so far:

EPOS *joined* the EUDAT CDI
We defined a *specific policy* with them
We *tuned the transfer tools* to achieve the best performance, but the HP tool (GridFTP) was useless since the bottleneck was the bandwidth
So we chose a *more flexible* tool like iRODS irsync protocol
In fact in order to achieve a hourly synchronization we exploited checksum sync and file age limit options.

Why we love data replication

– *data bit-stream preservation*

– *more optimal data curation*

– *better accessibility of data*

– *identification of data through Persistent Identifiers (PIDs)*
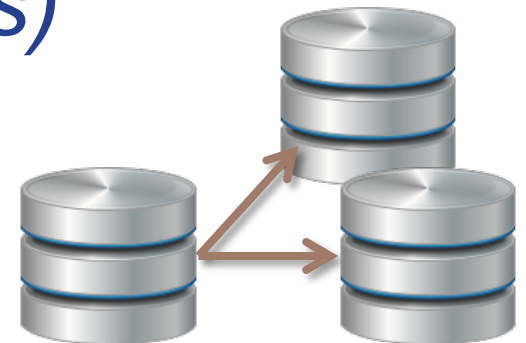
EUDAT is the solution, or maybe not
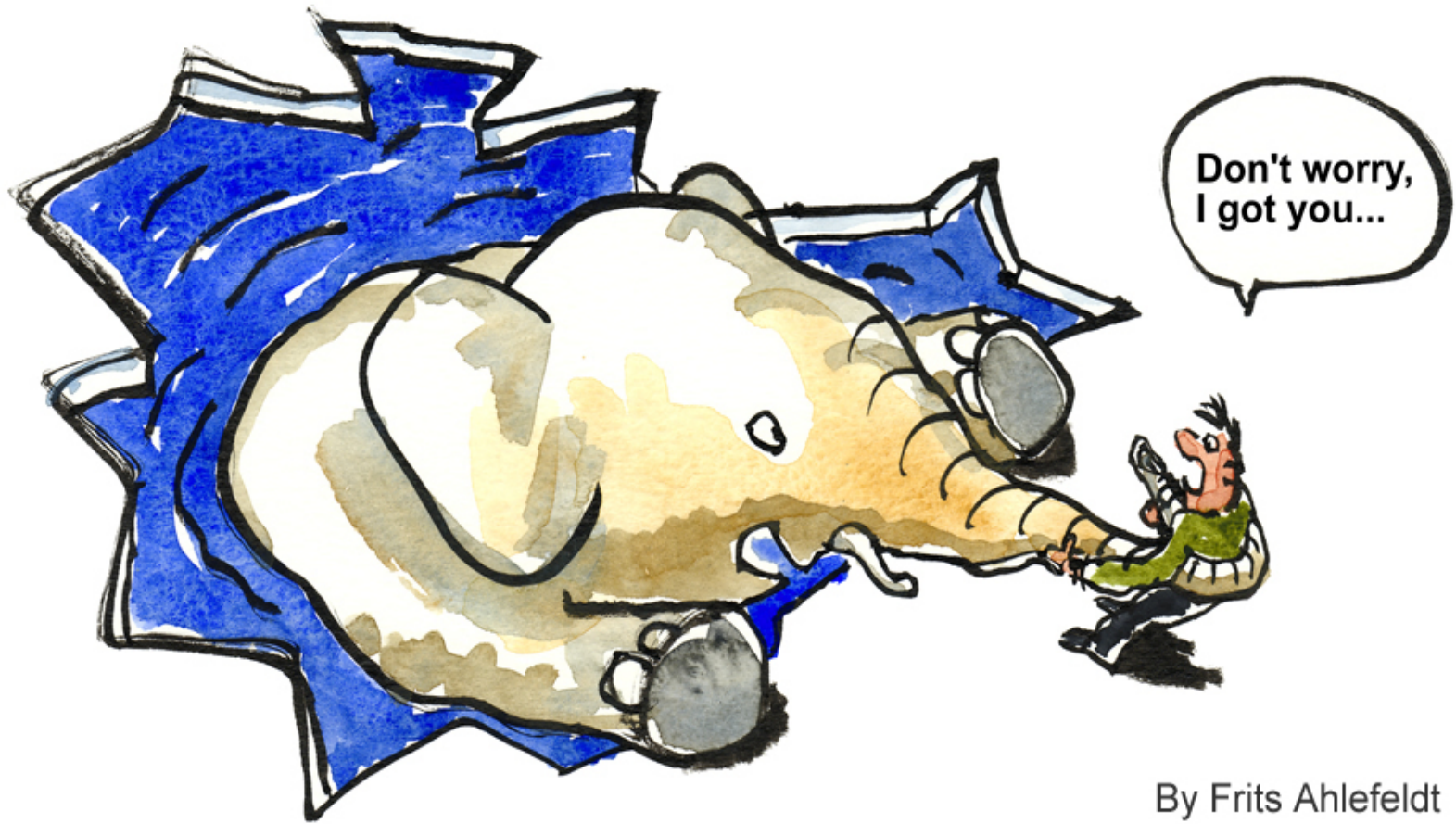
By Frits Ahlefeldt

Contact Us

http://www.eudat.eu

eudat-info@postit.csc.fi

Project Coordinator: Kimmo Koski
kimmo.koski@csc.fi

Scientific Coordinator: Peter Wittenburg
peter.wittenburg@mpi.nl

Project Manager: Damien Lecarpentier
damien.lecarpentier@csc.fi

Dissemination Manager: Nagham Salman
nagham.salman@bsc.es

Industry Task Force: David Manset
dmanset@maatg.fr

EUDAT

Thank you!