EUDAT Training

2nd EUDAT Conference, Rome

October 28th

Introduction, Vision and Architecture

SEVENTH FRAMEWORK PROGRAMME

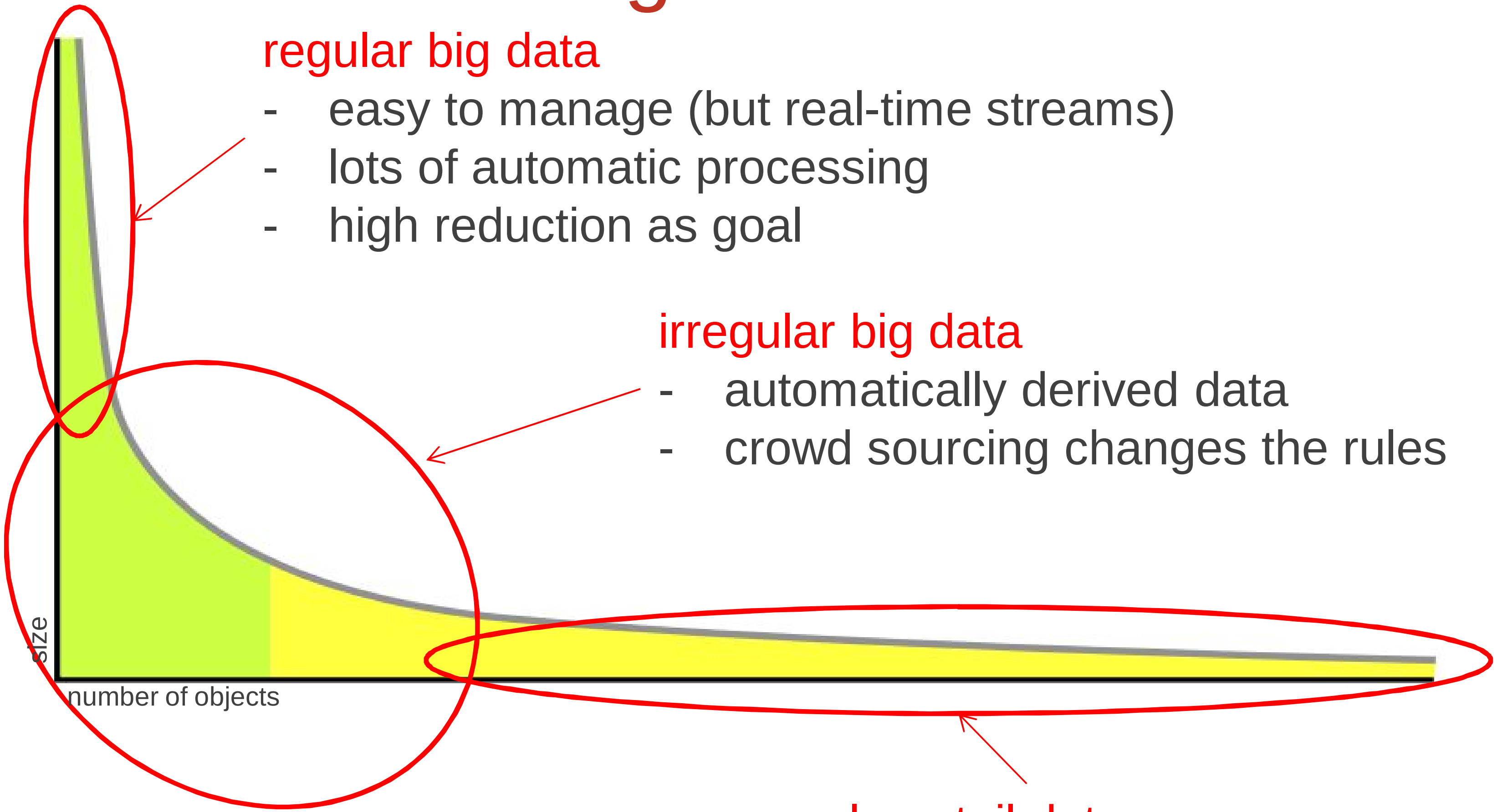
Giuseppe Fiameni – CINECA
Rob Baxter – EPCC
EUDAT members

EUDAT

Agenda

- Background information
- Services
- Common Data Infrastructure

Setting the scene



regular big data

- easy to manage (but real-time streams)
- lots of automatic processing
- high reduction as goal

irregular big data

- automatically derived data
- crowd sourcing changes the rules

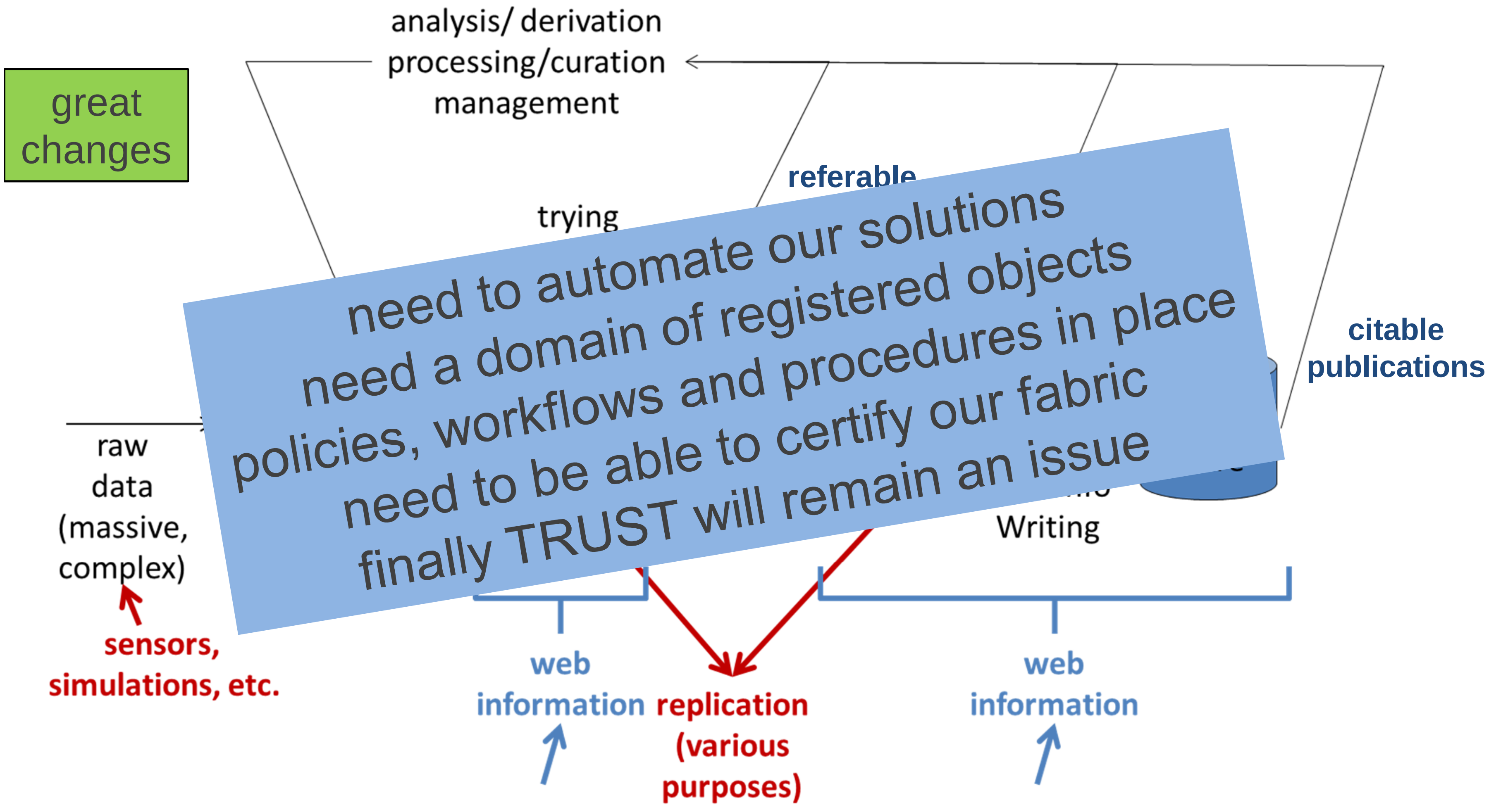
long tail data

- difficult to manage
- lots of relations

all the same for industry, government, public services, citizens, etc.

big scientific data –> the data fabric

great changes



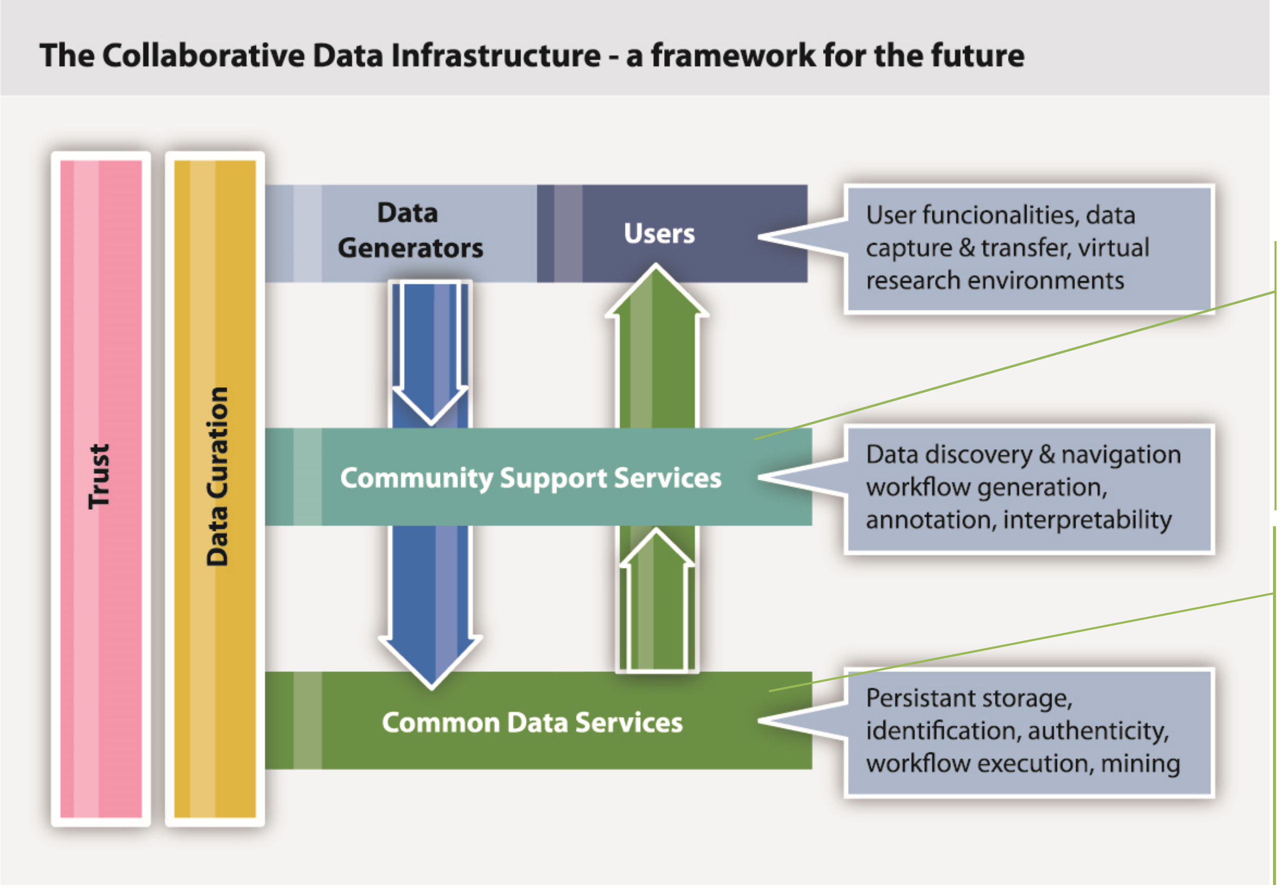
need to automate our solutions
need a domain of registered objects
policies, workflows and procedures in place
need to be able to certify our fabric
finally TRUST will remain an issue

EUDAT: Vision & Architecture

- EUDAT began with the concept of the Collaborative Data Infrastructure
  - See “Riding the Wave” (High Level Expert Group on Scientific Data, Final Report, 2010)
- This identified a handful of core Service Cases
  - See http://www.eudat.eu/services-and-technologies
- And the implementation of the Service Cases led to our current distributed Architecture
  - See later ☺

EUDAT

EUDAT’s mission: common services in CDI



The Collaborative Data Infrastructure - a framework for the future

Trust

Data Curation

Data Generators

Users

User funcionalities, data capture & transfer, virtual research environments

Community Support Services

Data discovery & navigation workflow generation, annotation, interpretability

Common Data Services

Persistant storage, identification, authenticity, workflow execution, mining

CLARIN, LifeWatch, ENES, EPOS, VPH, INFC etc.
6 Core Infrastructures
about 20 infrastructures

⇒ 12 EUDAT data centers

⇒ **and/or cross-disciplinary initiatives**

diagram taken from EC’s HLEG report “Riding the Wave”

What is the EUDAT CDI?

- The EUDAT Collaborative Data Infrastructure is
  - a ***pan-European***, ***cross-disciplinary*** domain of research data for both ***big community*** researchers and ***“long tail”*** scientists
  - where data are ***registered, preserved***, ***accessible*** and made ***re-usable***

What does this mean?

- ***Pan-European***
  - Fundamentally, a wide-area distributed architecture

- ***Cross-disciplinary***
  - Five core stakeholder communities, many other interested; many sources of conflicting requirements!
  - Including simplified services to encourage the “long tail” to participate
  - All implies a significant systems integration challenge!

What does this mean? (2)

- ***Registered*** means EUDAT data are
  - – Globally identified and discoverable (the PID Service)

- ***Preserved*** means EUDAT data are
  - – Stored at big European HPC and data centres
  - – Replicated for safety (the Safe Replication Service)
  - – Governed by policy rules (the Policy Management Service)

What does this mean? (3)

- ***Accessible*** means EUDAT data are
  - – Identifiable and findable (the PID Service)
  - – Retrievable efficiently (the Data Staging Service)
  - – Governed by suitable access control (the AAI Service)

- ***Re-usable*** means EUDAT data are
  - – Findable (the PID Service)
  - – Comprehensible (the Joint Metadata Service)
  - – Composable and combinable (future workflow and computational services)

EUDAT

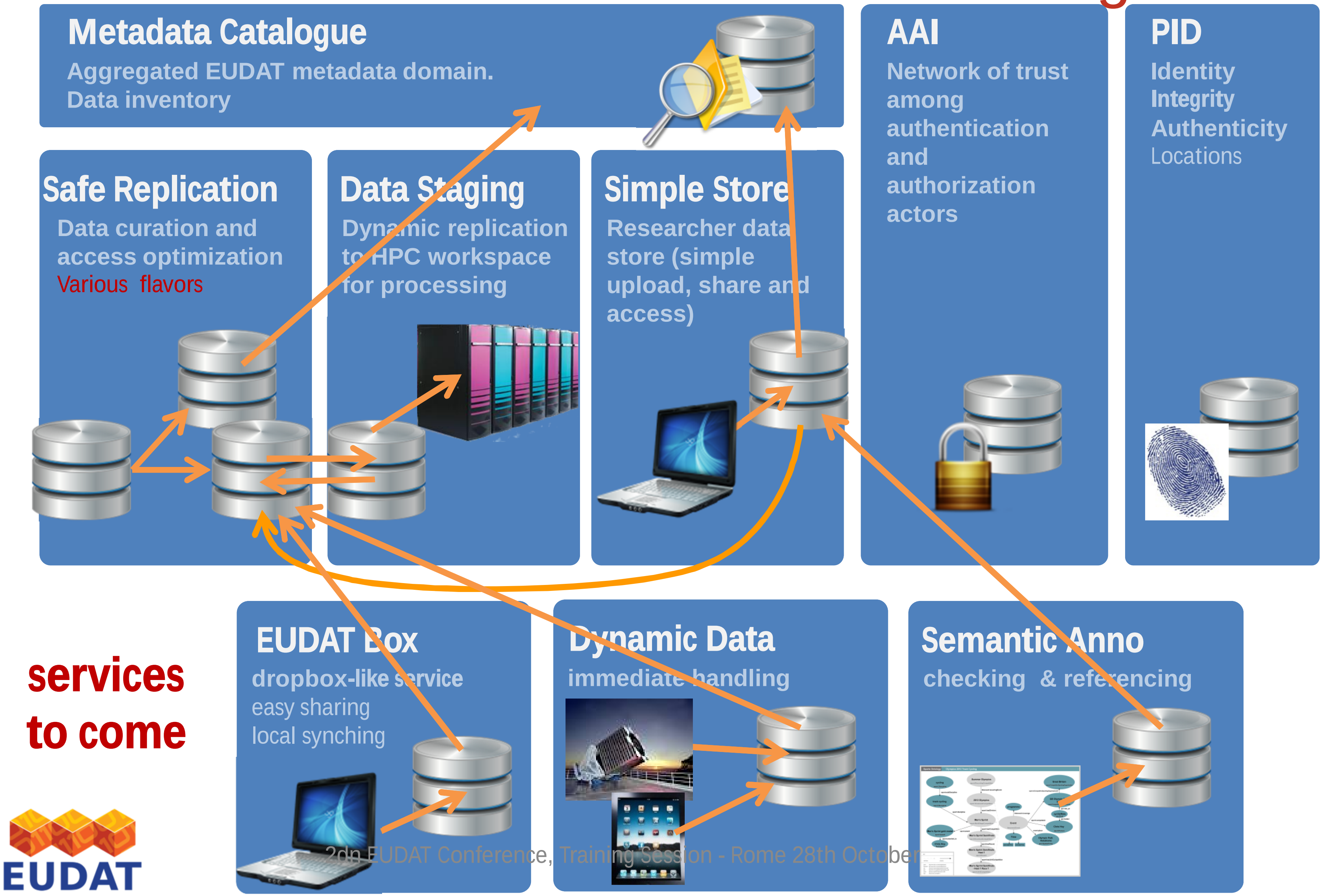
What does this mean? (4)

- For both ***big communities*** and ***“long tail”*** means
  - Stable, web-service APIs for existing tool-stacks to use (the Common Service Layer Interface)
  - Low barriers to use (the Simple Store Service)

- Hence the core EUDAT service cases

- Identifying solutions for these cases *that work with our stakeholder communities’ existing solutions* led us to the current CDI architecture

common services EUDAT is working on



Metadata Catalogue

Aggregated EUDAT metadata domain.
Data inventory

AAI

Network of trust among authentication and authorization actors

PID

Identity
Integrity
Authenticity
Locations

Safe Replication

Data curation and access optimization
Various flavors

Data Staging

Dynamic replication to HPC workspace for processing

Simple Store

Researcher data store (simple upload, share and access)

services to come

EUDAT Box

dropbox-like service
easy sharing
local synching

Dynamic Data

immediate handling

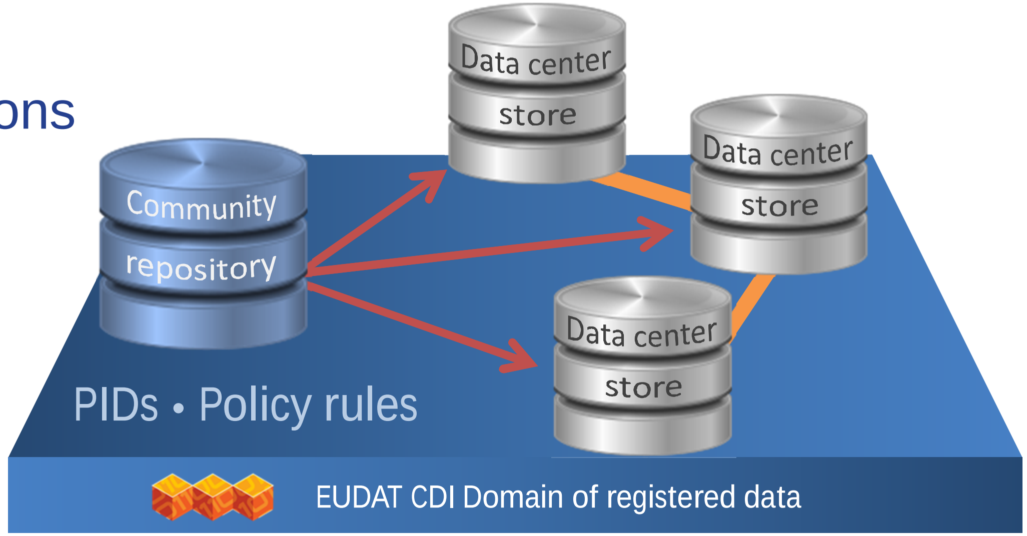
Semantic Anno

checking & referencing

EUDAT

Safe Replication Service

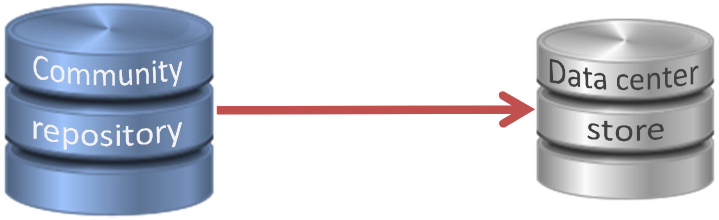
- Robust, safe and highly available data replication service for small- and medium- sized repositories
  - To guard against data loss in long-term archiving and preservation
  - To optimize access for user from different regions
  - To bring data closer to powerful computers for compute-intensive analysis



http://eudat.eu/safe-replication | eudat-safereplication@postit.csc.fi

Safe Replication Service (B2SAFE)

PIDs • Policy rules • iRODS



communities/departments

- do not have IT people
- can't install iRODS
- can't adapt their repository solution
- are partly using ready-made solutions (Fedora, D-SPACE, WikiMedia, etc.)
- don't know how to register PIDs
- etc.

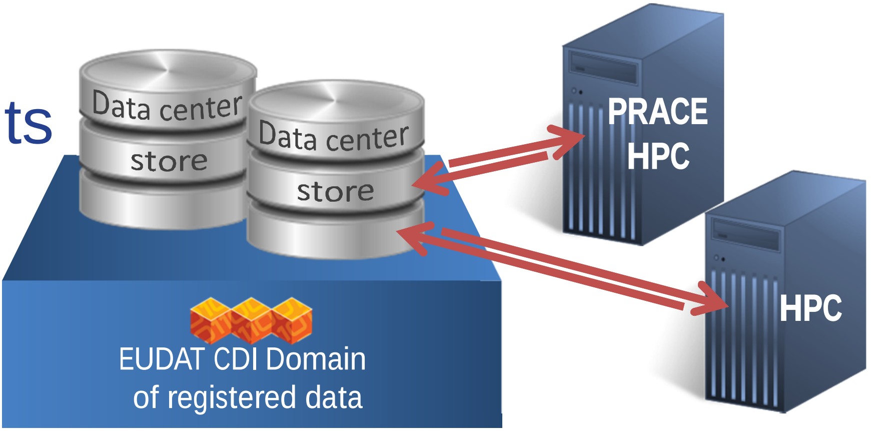
- not as easy as thought
- so what to do?
- created/are creating various flavors

– FULL SR via iRODS policies	ready
– Light SR via GridFTP client	ready
– Lighter SR via HTTP client	in p
– Packages for	
• Fedora	ready
• D-SPACE	in p
• WikiMedia	tbd
• ?	

http://eudat.eu/safe-replication | eudat-safereplication@postit.csc.fi

Data Staging Service

- Support researchers in transferring large data collections from EUDAT storage to HPC facilities
- Reliable, efficient, and easy-to-use tools to manage data transfers
- Provide the means to re-ingest computational results back into the EUDAT infrastructure

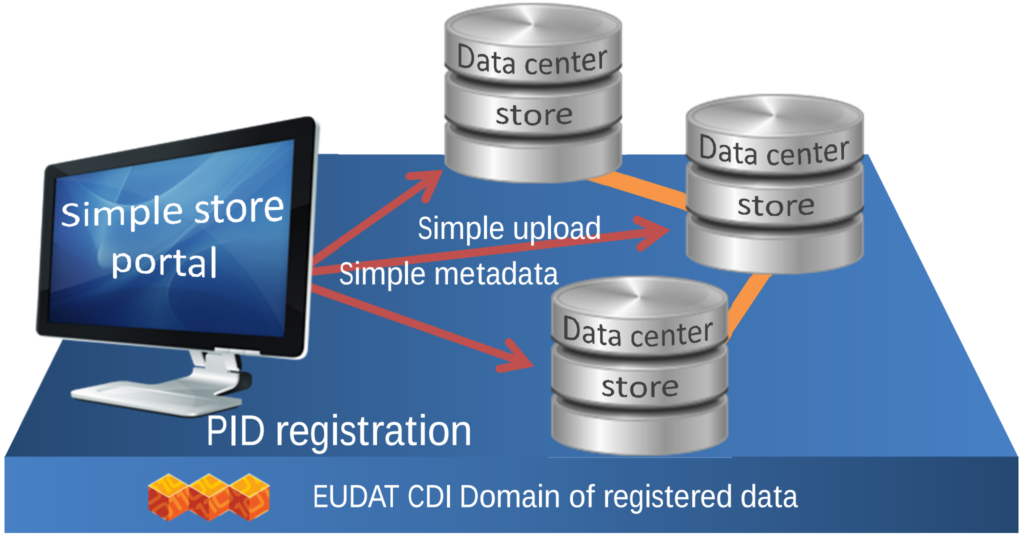


- not a simple service!
- politics involved (access to HPC)

http://eudat.eu/datastaging | eudat-datastaging@postit.csc.fi

Simple Store Service

- Allow registered users to upload ”long tail” data into the EUDAT store
- Enable sharing objects and collections with other researchers
- Utilise other EUDAT services to provide reliability

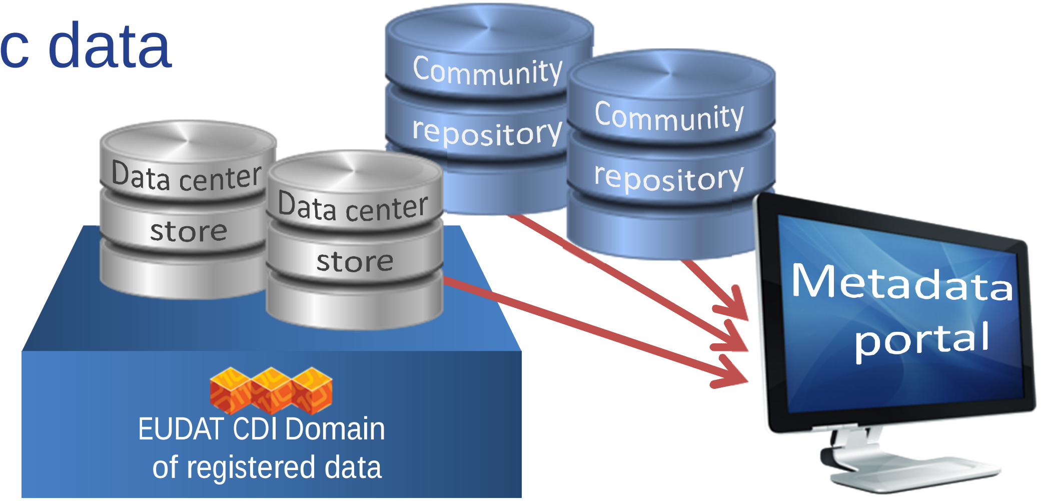


- much competition
- see it as complementary – finally it is about trust

http://eudat.eu/simplestore | eudat-simplestore@postit.csc.fi

Metadata Service

- Easily find collections of scientific data – generated either by various communities or via EUDAT services
- Access those data collections through the given references in the metadata to the relevant data stores
- Europeana of scientific data

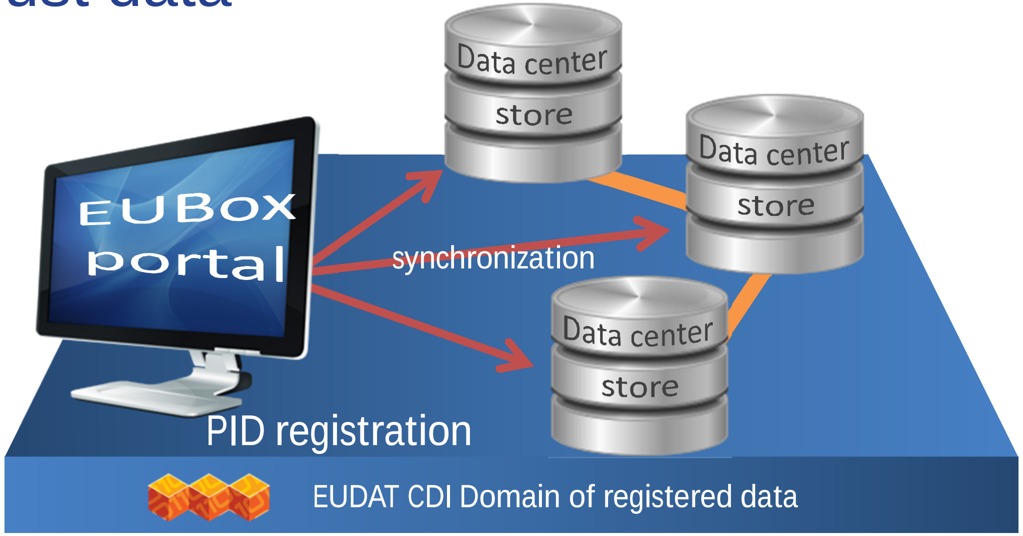


- how to offer metadata in a cross-disciplinary space?
- scalability issue?

EUDAT Box Service

- some similarity to SimpleStore of course
- just similar to Dropbox incl. load balancing and replication
- there is no metadata – just data

- how to integrate into registered domain of data?

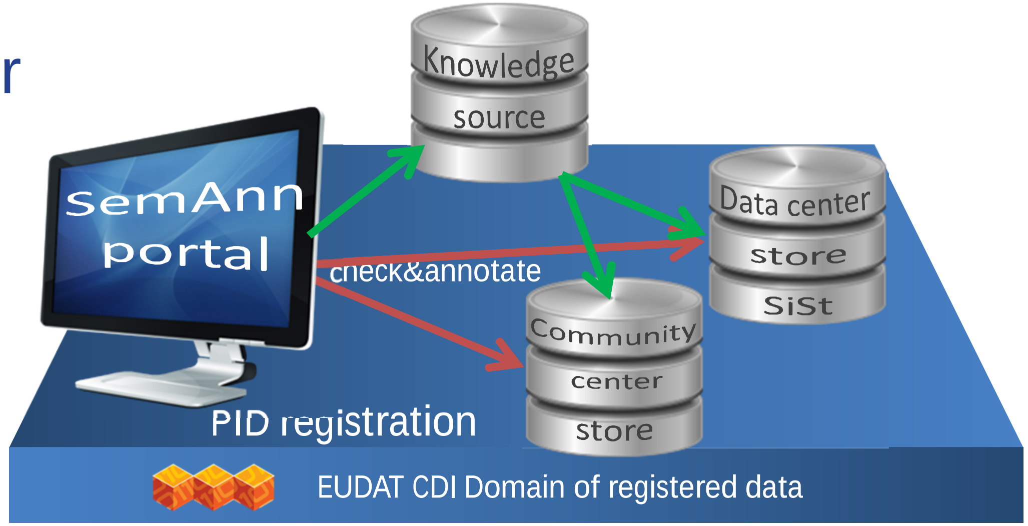


- much competition
- see it as complementary – finally it is about trust

Semantic Annotation Service

- acts as a plugin component to be executed before uploading a resource with tags (crowd sourcing etc.)
- check tags against Knowledge Source & correct/refer/etc.

- could be used as trigger in Simple Store
- plugin available to everyone



- not center dependent

Service Targeting

- **Replication**: targeted at data managers/archivists/ projects/departments without facilities
- **Data Staging**: same plus “easy” access to HPC
- **SimpleStore**: place for individuals/projects/groups to store & exchange data
- **EUBox**: share data via synchronization
- **Metadata**: EUDAT data & everyone interested
- **SemAnn**: individual/projects working with massive amounts of human created data

data stored in domain of registered data is not EUDAT’s data!

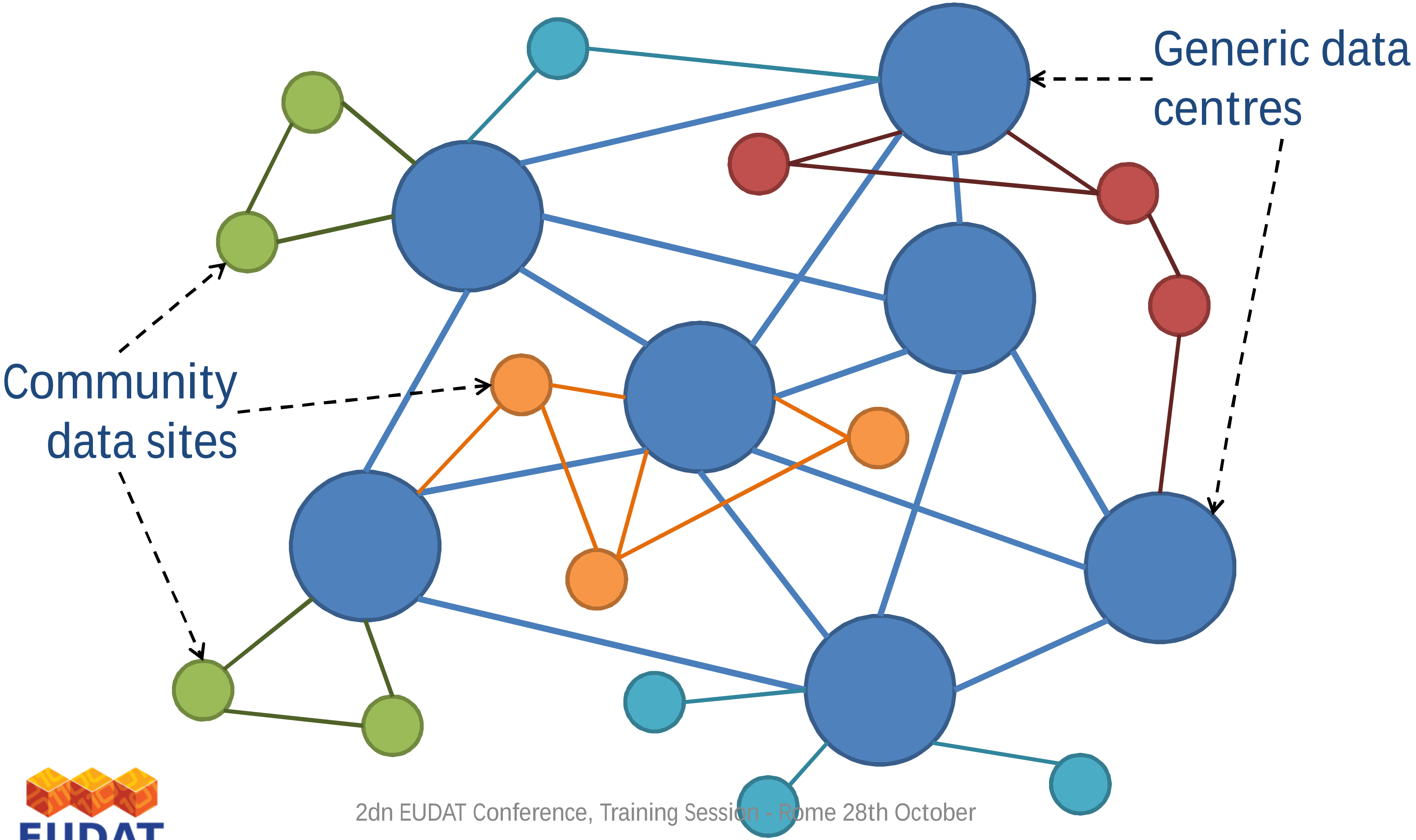
COMMON DATA INFRASTRUCTURE

The CDI network architecture

- The CDI is a connected network of European research institutions and data centres (collectively *Nodes*) each offering one or more common EUDAT data services to both participating research communities and independent researchers

- Data centre Nodes have lots of connections
- Research community Nodes need only one
- Connections have both technical & policy agreement aspects

The CDI network architecture


Generic data centres

Community data sites

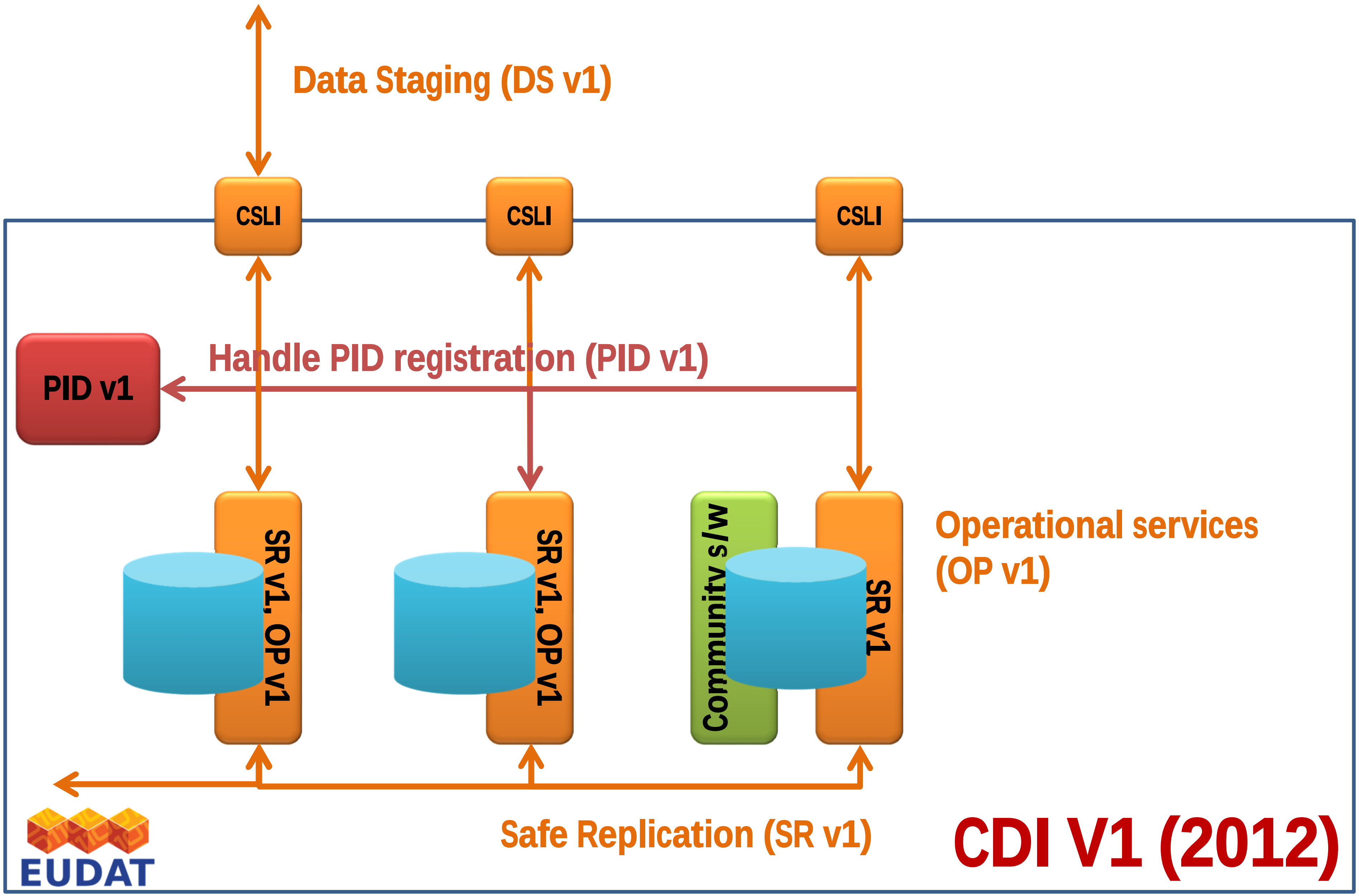
CDI Node architecture

- Nodes run parts of the CDI Node software suite, depending on which services they want to offer
- All Nodes should offer Safe Replication and PID
  - This is really what being in the CDI is all about
- Others are optional
  - Depends on what a Node’s expected user base requires
- (Some data centre Nodes also need to run the Operational Services suite)

CDI service development timeline

Year	EUDAT Services	CDI
2012	Safe Replication v1, Data Staging v1, PID v1, Operational Services v1; AAI (design)	V 1
2013	AAI v1, Joint Metadata v1, Simple Store Service v1, Safe Replication v2, Data Staging v2, Operational Services v2; Common Service Layer Interface (design), workflow and computation services (design)	V 2
2014	AAI v2, Joint Metadata v2, Simple Store Service v2, Data Staging v3, Operational Services v3, CSLI v1, computational and workflow services v1	V 3
2015 +	PID v2, Safe Replication v3, computational and workflow services v2	V 4

CDI V1 (2012)



Data Staging (DS v1)

CSLI

CSLI

CSLI

PID v1

Handle PID registration (PID v1)

SR v1, OP v1

SR v1, OP v1

Community s/w

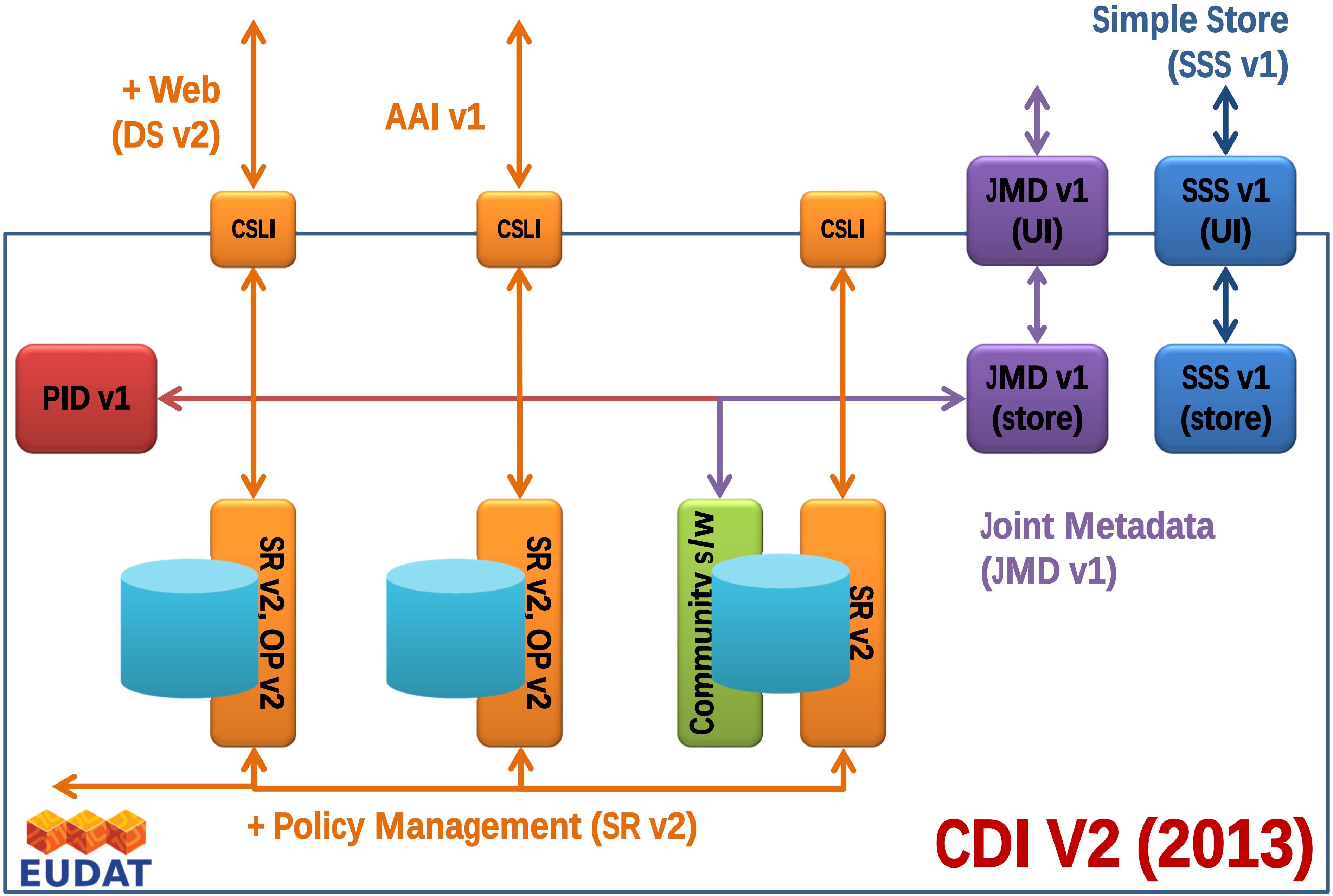
SR v1

Operational services (OP v1)

Safe Replication (SR v1)

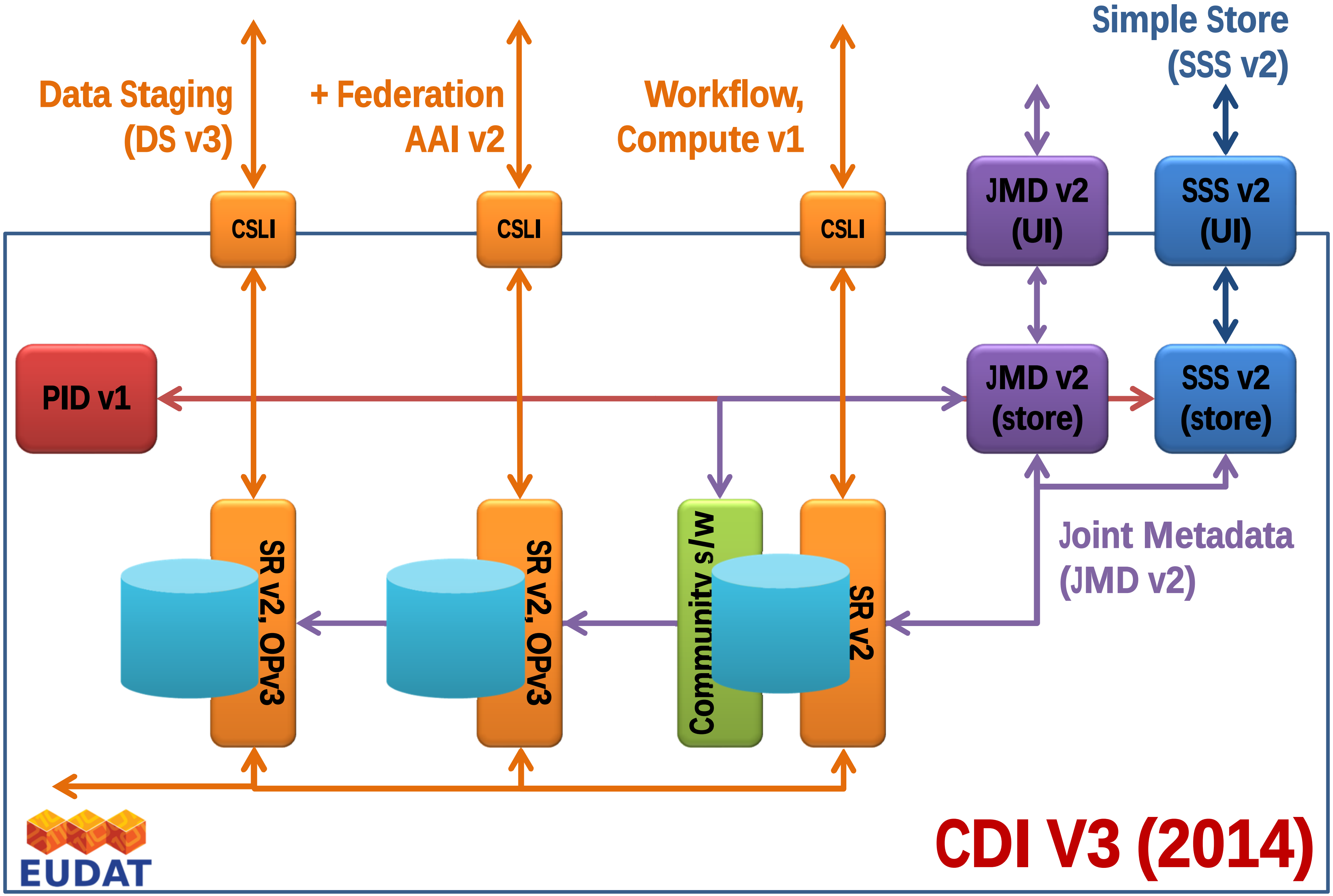
EUDAT

CDI V2 (2013)



+ Web
(DS v2)
AAI v1
Simple Store
(SSS v1)
CSLI
CSLI
CSLI
JMD v1
(UI)
SSS v1
(UI)
PID v1
JMD v1
(store)
SSS v1
(store)
SR v2, OP v2
SR v2, OP v2
Community s/w
SR v2
Joint Metadata
(JMD v1)
+ Policy Management (SR v2)
EUDAT

CDI V3 (2014)



Data Staging (DS v3)

+ Federation AAI v2

Workflow, Compute v1

Simple Store (SSS v2)

CSLI

CSLI

CSLI

JMD v2 (UI)

SSS v2 (UI)

PID v1

JMD v2 (store)

SSS v2 (store)

SR v2, OPv3

SR v2, OPv3

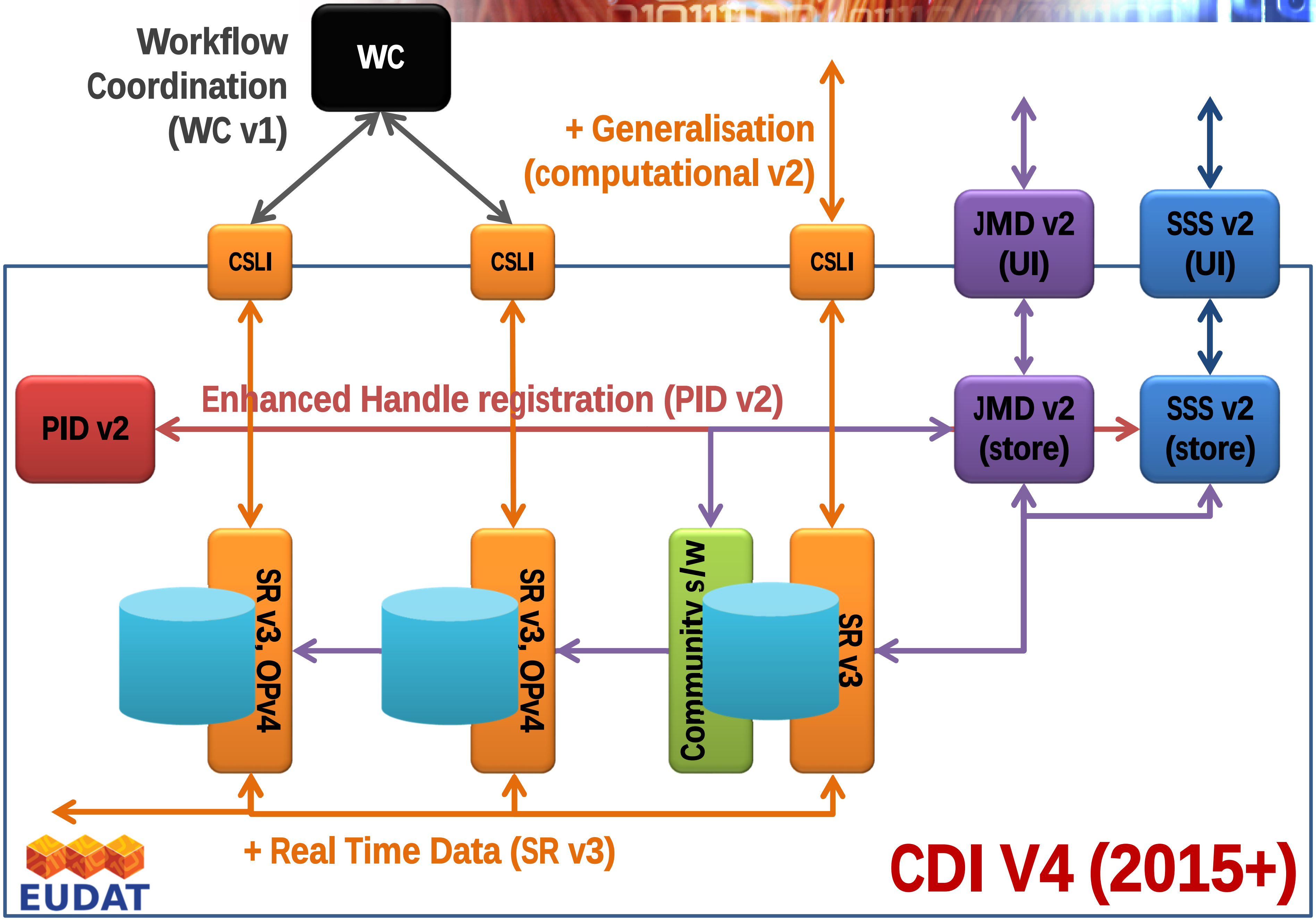
Community s/w

SR v2

Joint Metadata (JMD v2)

EUDAT

CDI V4 (2015+)



Workflow Coordination (WC v1)

WC

+ Generalisation (computational v2)

CSLI

CSLI

CSLI

JMD v2 (UI)

SSS v2 (UI)

PID v2

Enhanced Handle registration (PID v2)

JMD v2 (store)

SSS v2 (store)

SR v3, OPv4

SR v3, OPv4

Community s/w

SR v3

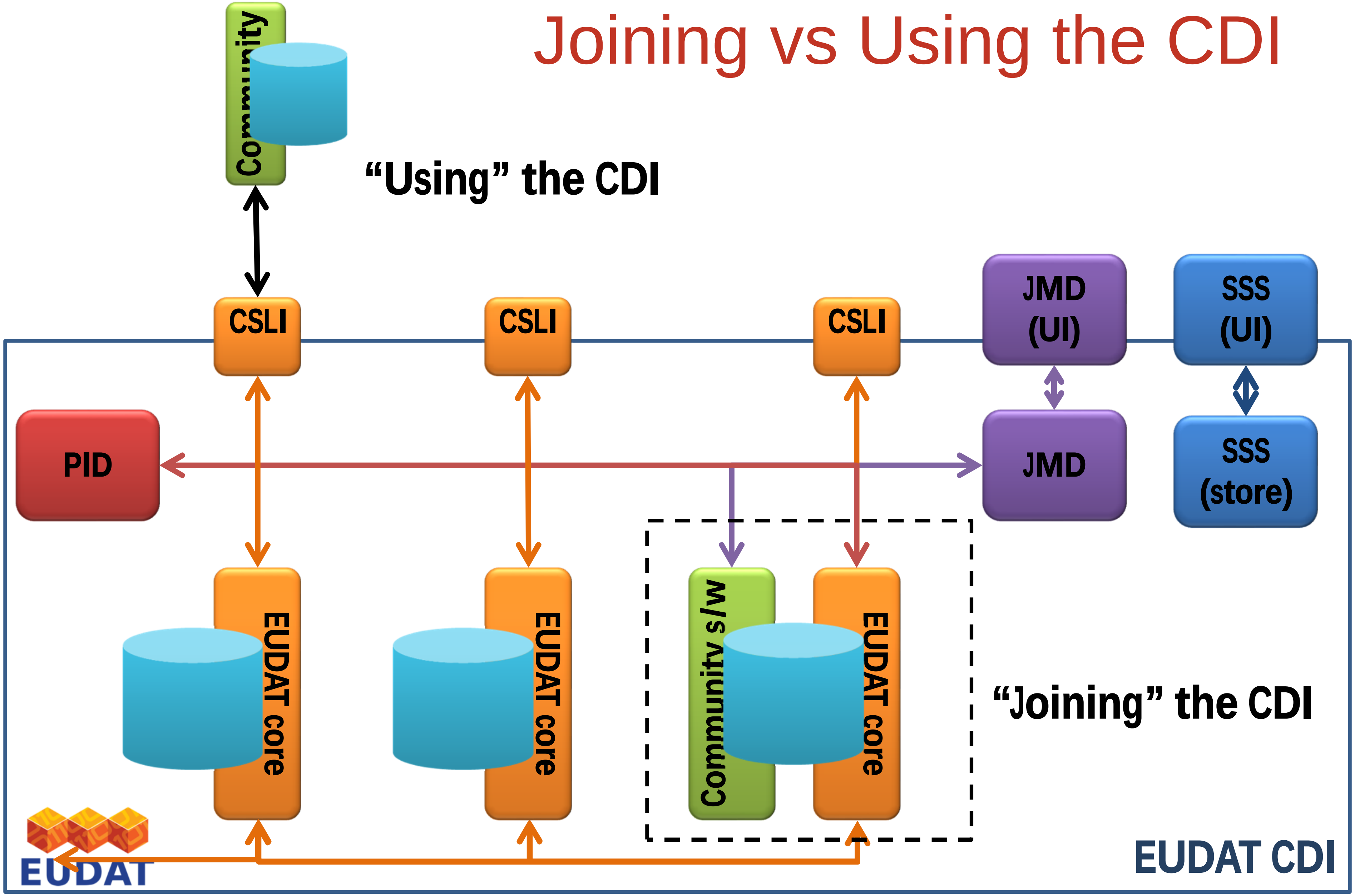
+ Real Time Data (SR v3)

EUDAT

Current Node software suite

- **Safe Replication**
  - iRODS v3.x + EUDAT replication microservices + GSI (X.509) security
- **PID**
  - EPIC/Handle system (external service) + EUDAT EPIC client microservices for iRODS
- **Data Staging**
  - GridFTP iRODS DSI + EUDAT Data Staging script
- **Joint Metadata**
  - CKAN + Apache Solr + OAI-PMH
- **Simple Store**
  - Invenio + EUDAT faceted user interface layer

Joining vs Using the CDI



Many thanks for your attention!

QUESTIONS?