



Research Data Management using iRODS in the EUDAT Infrastructure

Johannes Reetz, RZG – Max Planck Society

iRODS User Meeting
Harvard University, Cambridge, MA
June 18-19, 2014



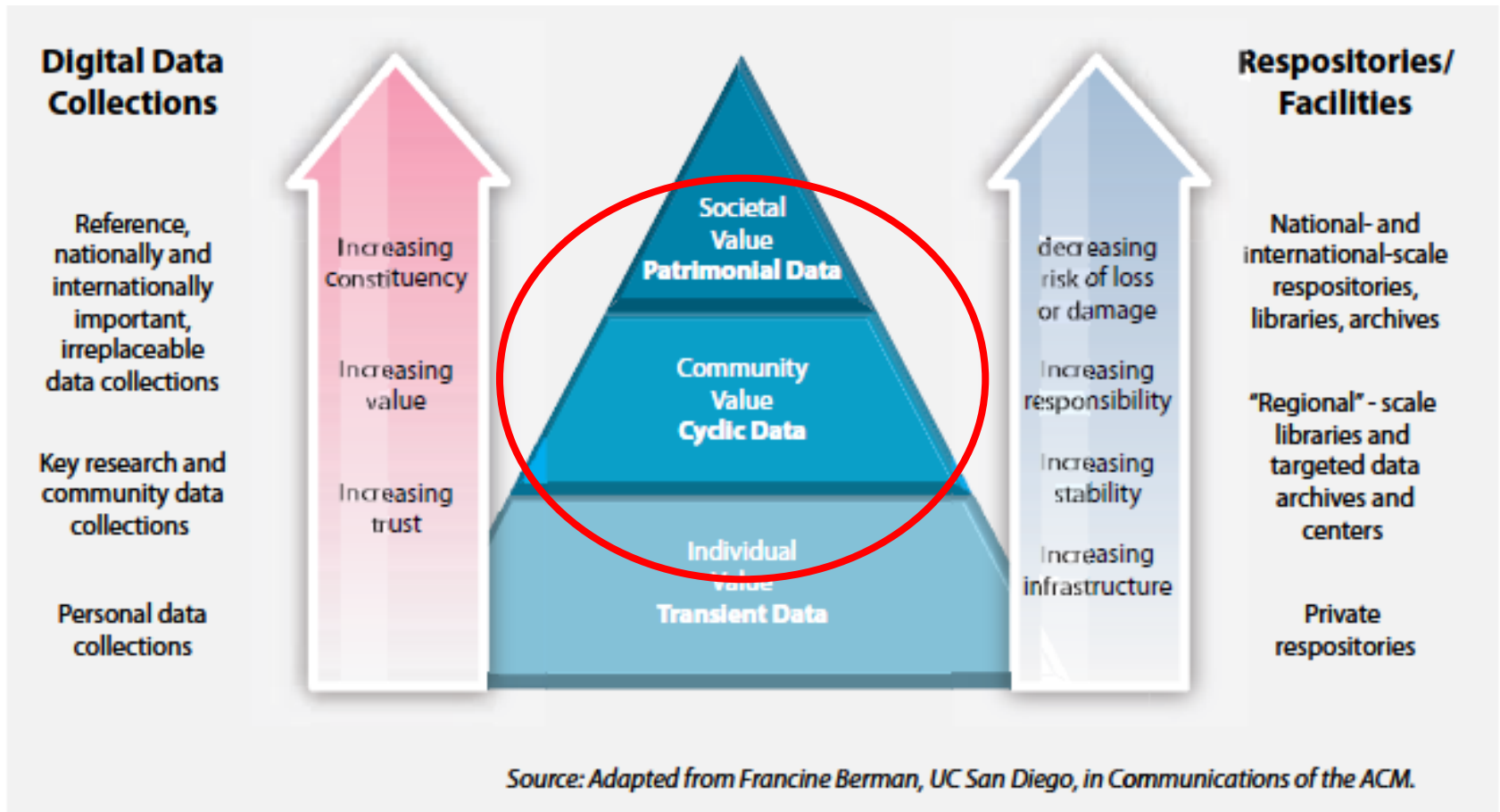
EUDAT Consortium

EC funded project (Oct 2011 - Mar 2015), follow-up project planned (2015+)

25 European partners



EUDAT is mainly focussing on Key Research and Community Data



EUDAT objectives

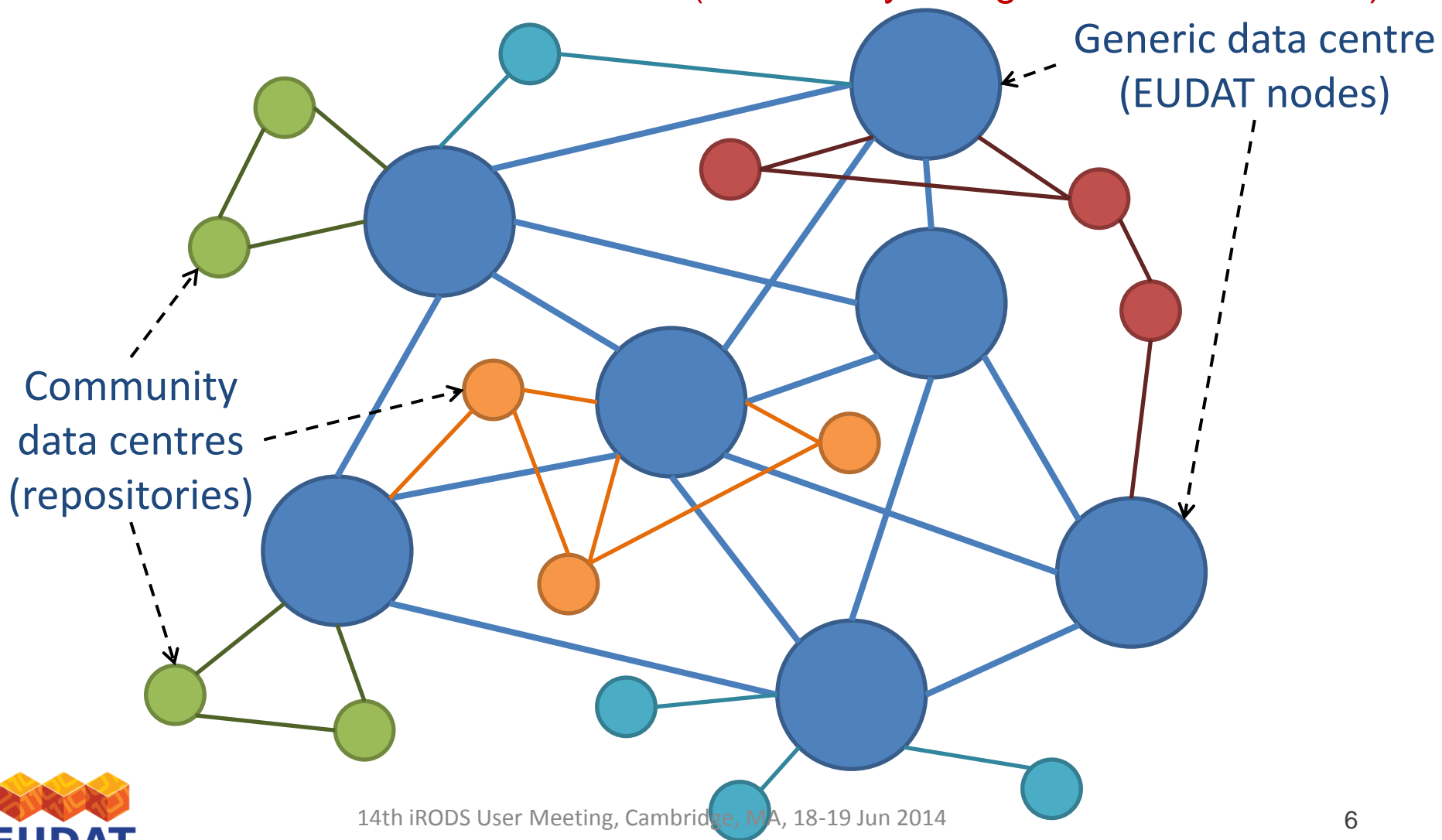
- Establish a European Collaborative Data Infrastructure as a federation of administratively independent cooperating centres (administrative zones)
- Cost-efficient, user-driven, adaptive, resilient, scalable and inclusive, providing an integrated solution for managing data through its full lifecycle including long-term preservation
- Supporting / supported by research infrastructures
- Geographically dispersed researchers and research communities can rely on a single cross-national infrastructure, providing interfaces to national solutions
- Interoperability with other e-Infrastructures

EUDAT Guiding Principles

- Research data deposited with the EUDAT CDI will be preserved for long-term (5, 10, 20yrs, or more)
- Data is typically replicated across different organisational boundaries (administrative zones).
- Data are best curated by their own communities; EUDAT relies on knowledgeable repository managers
- Community Trusted Digital Repositories (TDR) require that the EUDAT CDI is a suitable target for “TDR outsourcing”.
- EUDAT will not assert ownership of any data it holds; this implies a high degree of responsibility regarding policies.
- Infrastructure Security and Service Quality Management rules

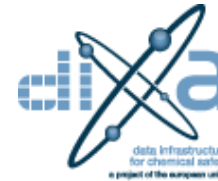
EUDAT Collaborative Data Infrastructure

network of administrative zones (community and generic data centres)



Communities (8+) and Data Centres (12+)

- **EPOS**: European Plate Observatory System
- **CLARIN**: Linguistics
- **ENES**: Climate Modelling
- **LifeWatch**: Biodiversity Data and Observatories
- **VPH**: Biomedicine
- **INCF**: Neuroinformatics
- **DRIHM**: Hydrometeorology
- **DiXA**: Chemical Safety
- more associated research communities

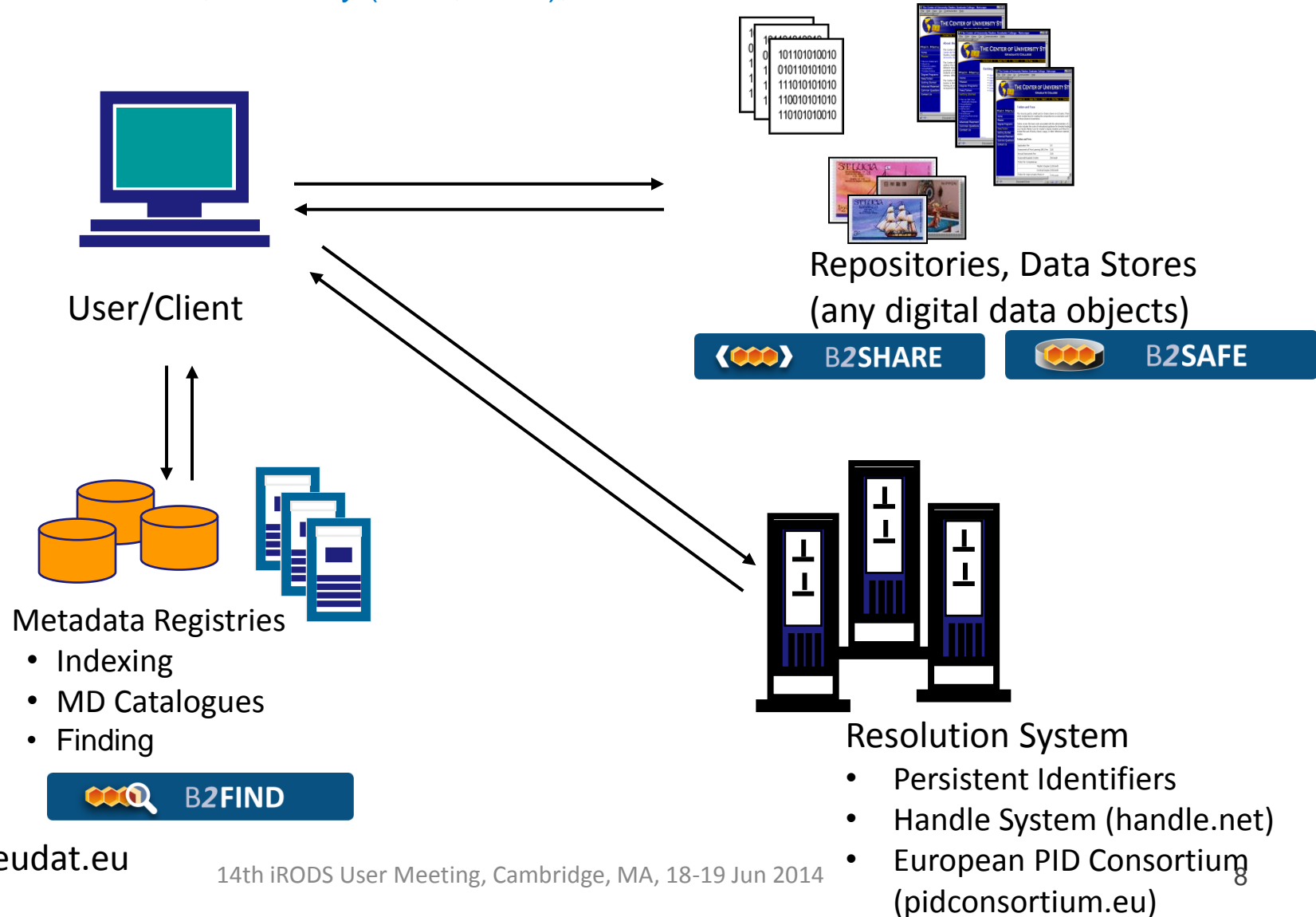


All community infrastructures share common challenges and requirements:

- Data management planning, DM policies
- Metadata management
- Persistent data references
- Long-term preservation, ensuring data integrity, authenticity, security
- Data sharing, distribution/publication, access and interoperability

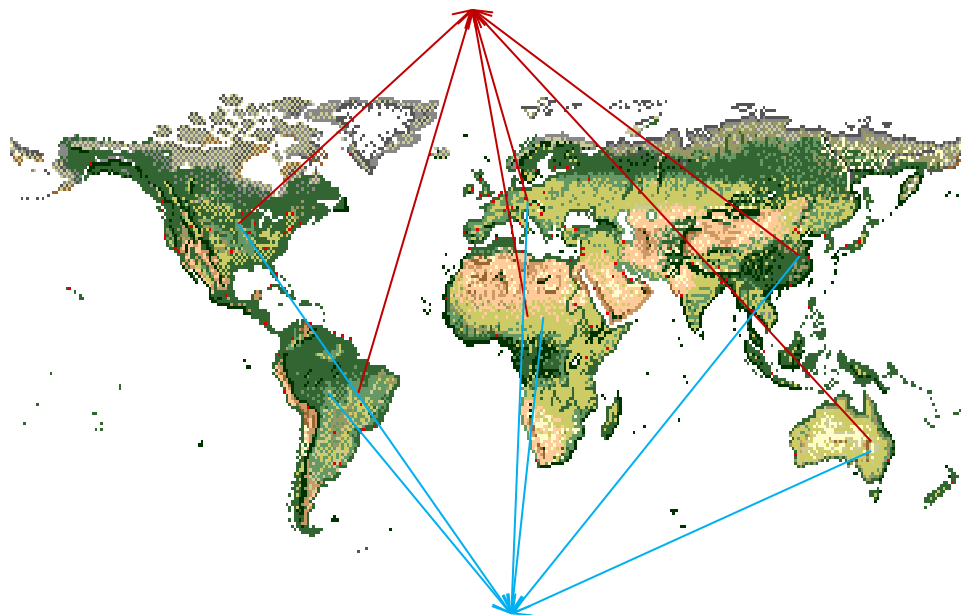
Digital Object Architecture

cf. Kahn, Wilensky (1995, 2006), DOI 10.1007/s00799-005-0128-x



EUDAT CDI relies on a global PID system (like IP)

DONA



PID Name Space Registration Authorities
Datacite, EPIC, ...
with associated PID service providers

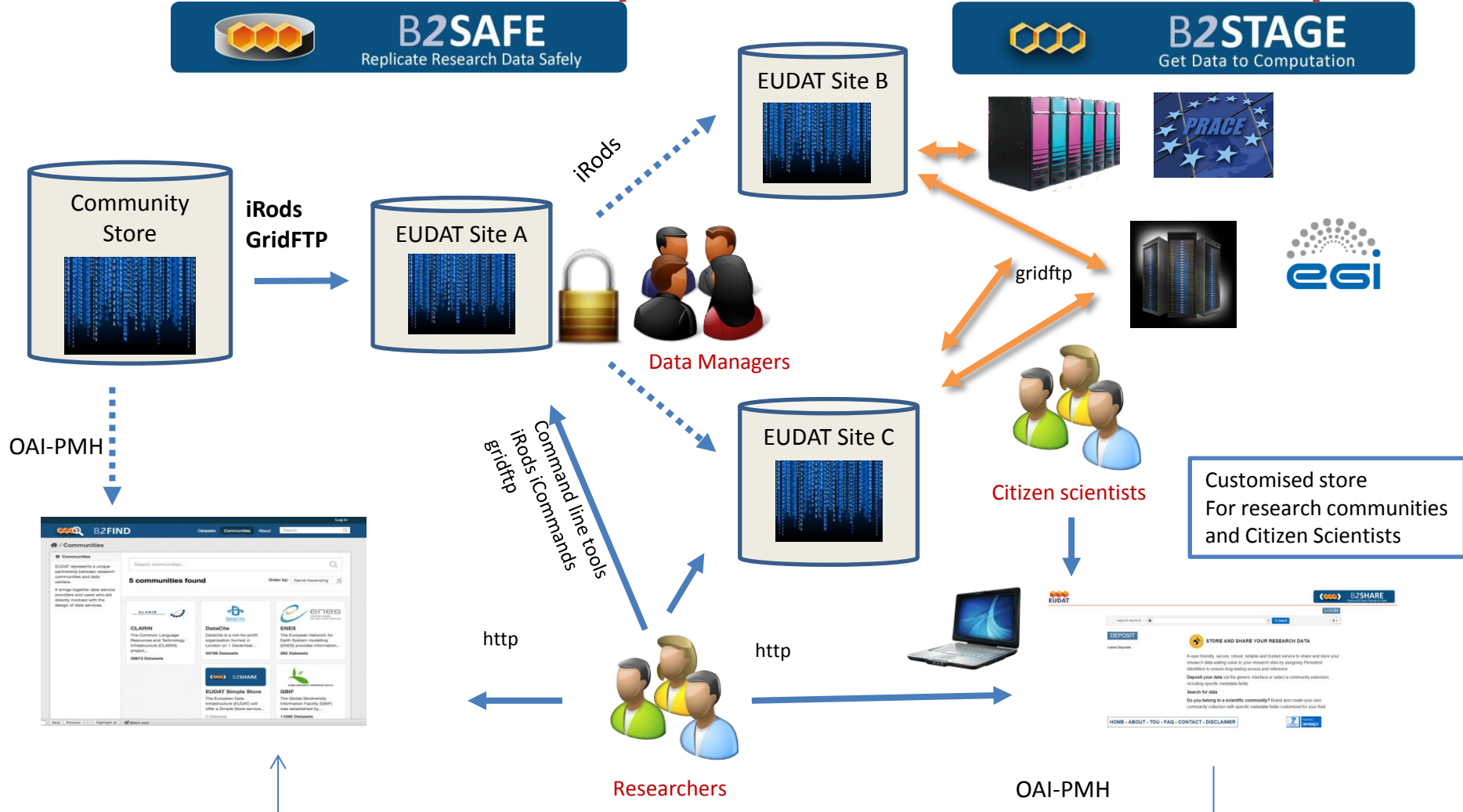
The challenge is to have a worldwide system to register digital objects such as with IP Numbers/nodes and a single protocol stack (like TCP/IP).

Need to be able to identify and proof integrity and authenticity of data, tools, services etc. Handle System offers a powerful solution via an alternative resolution system.

DONA is a foundation under Swiss law that sustains the Handle System independently from CNRI.

Federation of PID prefix registration authorities (RAs) are in the process of being established.

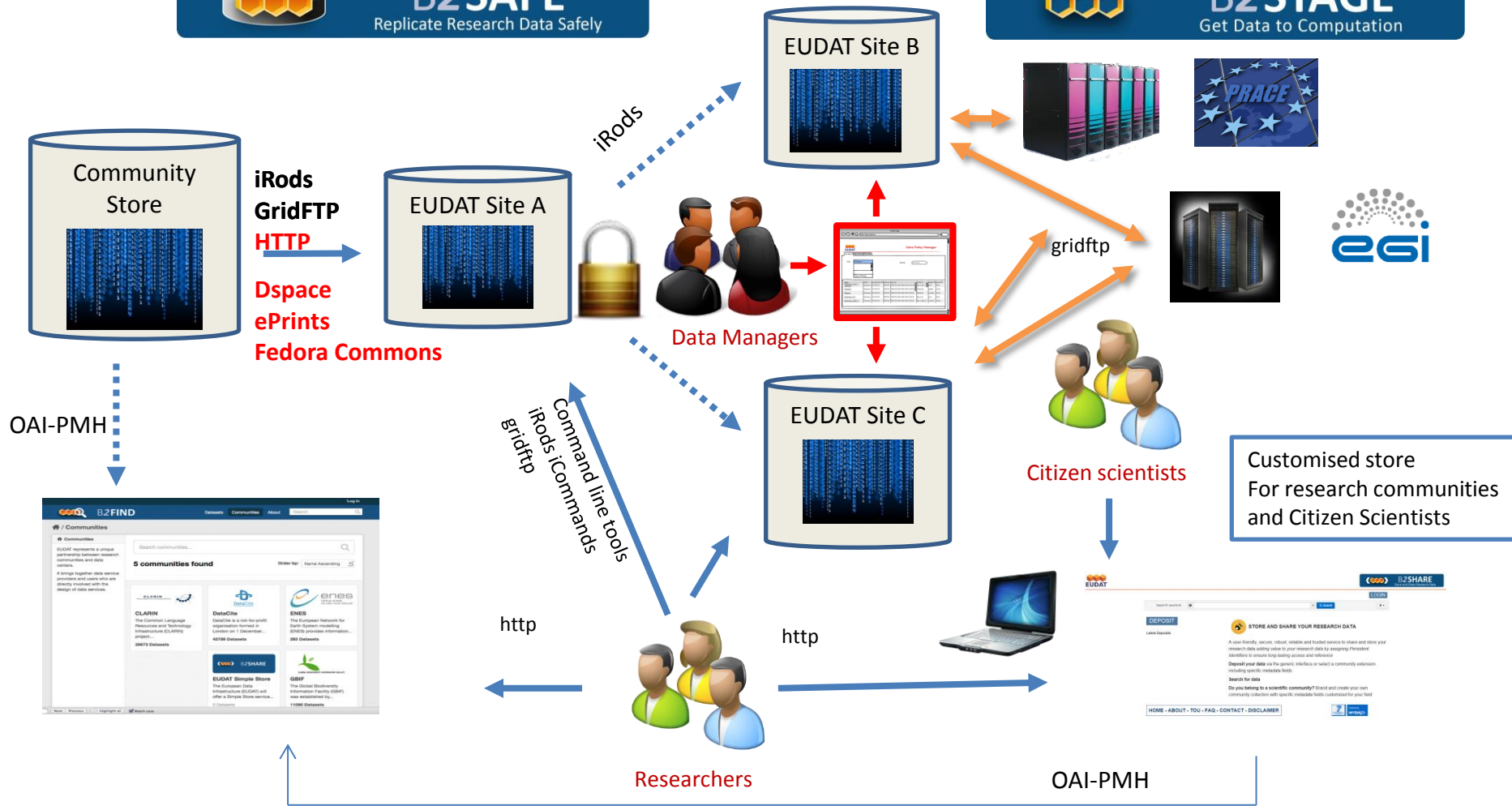
EUDAT service portfolio and landscape



EUDAT B2SAFE evolution

B2SAFE
Replicate Research Data Safely

B2STAGE
Get Data to Computation

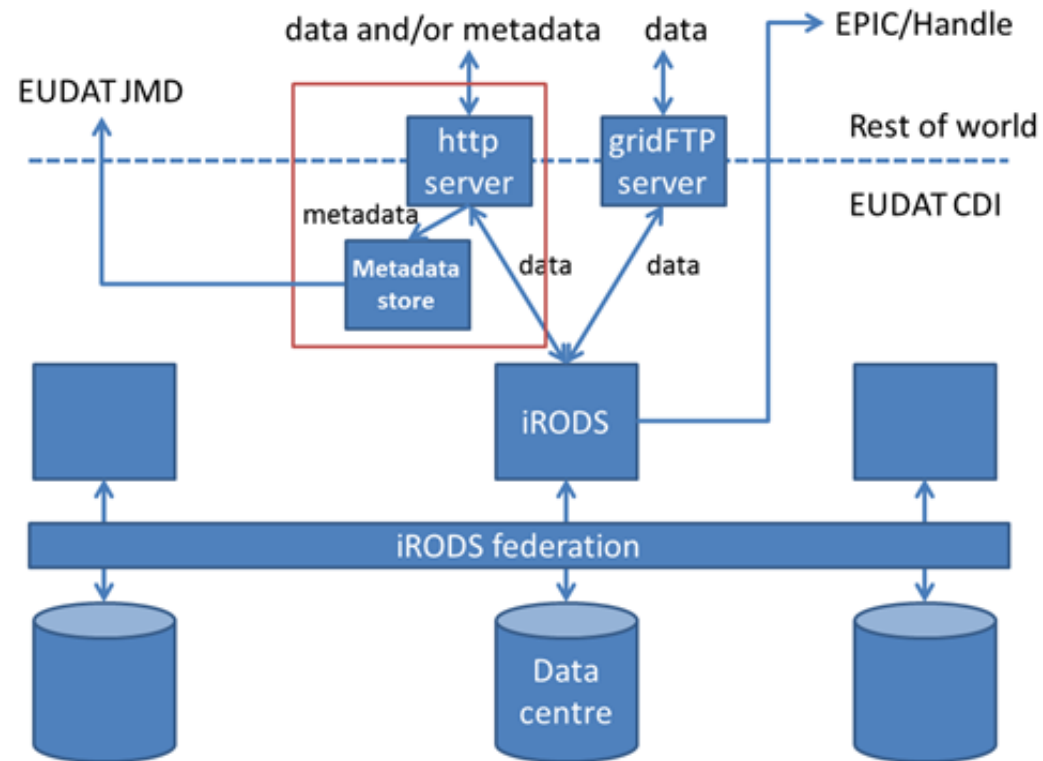


B2FIND
Find Research Data

14th iRODS User Meeting, Cambridge, MA, 18-19 Jun 2014

B2SHARE
Store and Share Research Data

Interface to B2SAFE

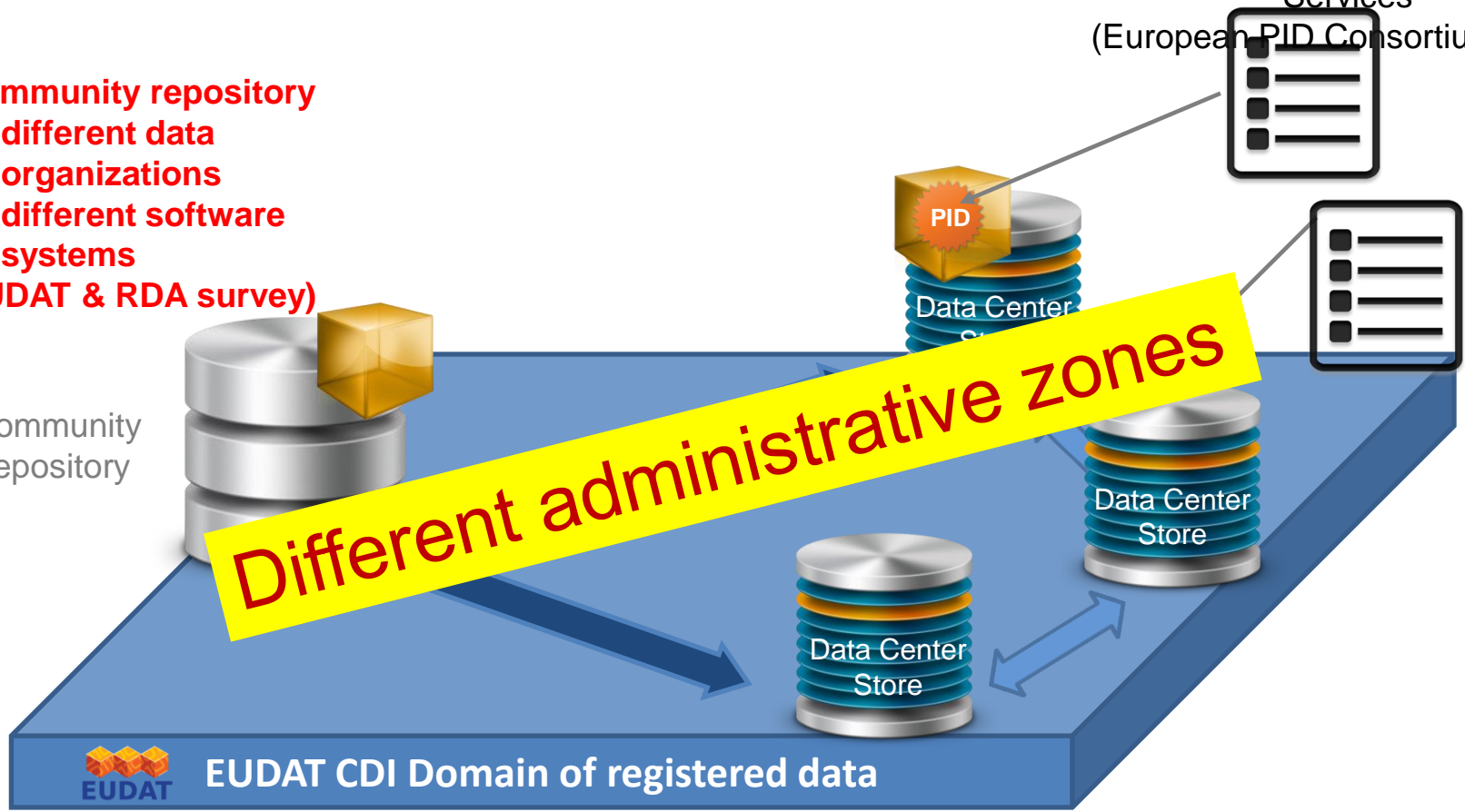


- **iRODS**
- **GridFTP**
- **HTTP/RESTfull API** to the B2Safe service
 - Upload and register a DO and retrieve PID, download via PID
 - Will be based on the open standard SNIA CDMI spec

B2SAFE- safe policy-driven replication

- Community repository**
- different data organizations
 - different software systems
- (EUDAT & RDA survey)

Different Handle Services
(European PID Consortium policy)





A real Use Case: CLARIN (MPI-PL/TLA) (~ 70 TB of data)

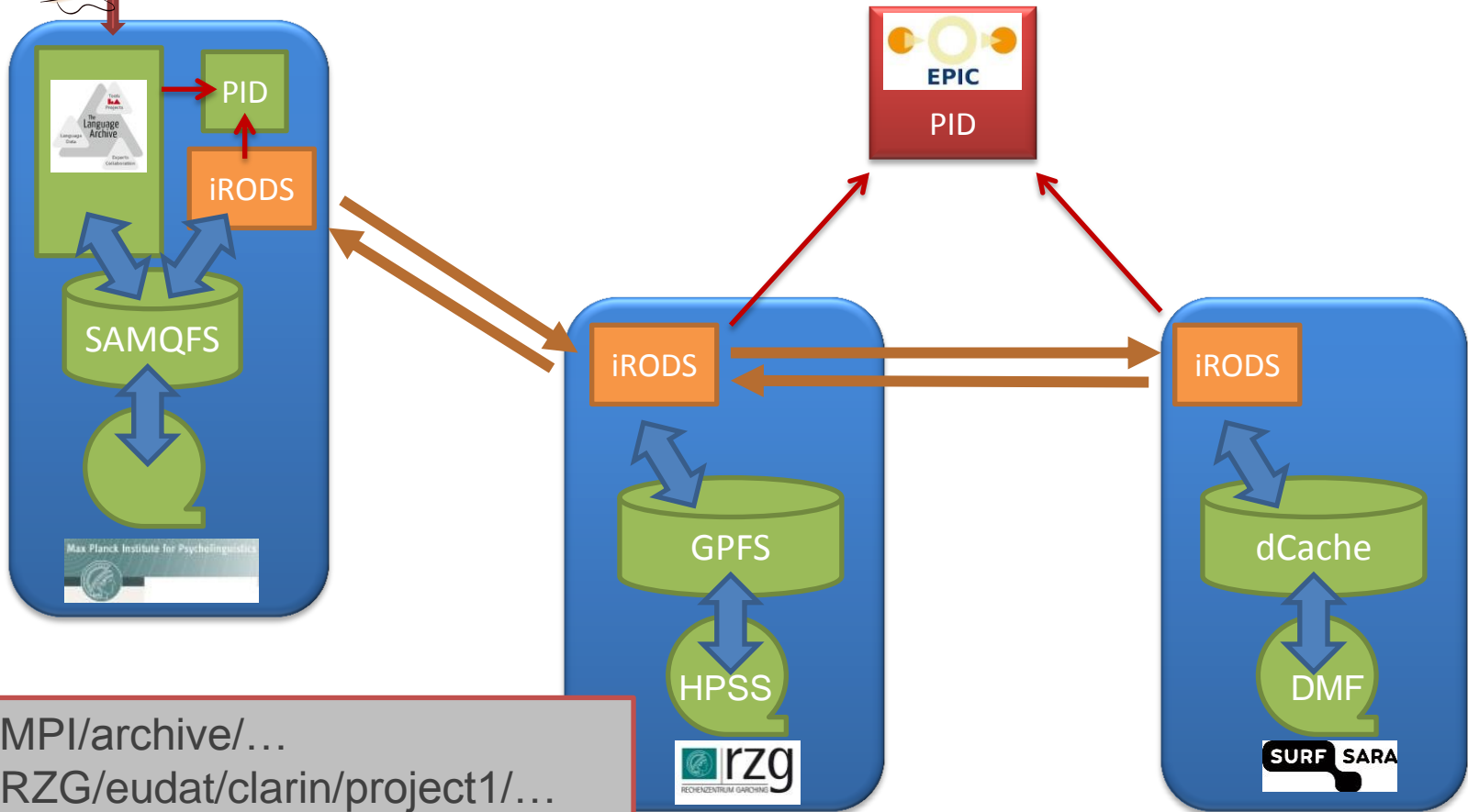


iRODS is suitable for EUDAT

- Data management is policy driven, user and system level rules
 - EUDAT replication policies
- Extensible via specific user defined rules and micro services
 - EUDAT Replication and EPIC PID micro service
- Scalable architecture: from single server to large scale (clustered) storage systems
- Integration with existing research data repositories, MPI-PL “The Language Archive”, EPOS, ENES, VPH, INCF, and mass storage systems (HPSS, dCache/DMF, TSM, S3, ...)
- iRODS is open source: BSD license
 - DICE, CC-IN2P3, EU SHAMAN, Australian ARCS, UK e-Science, King’s College and others



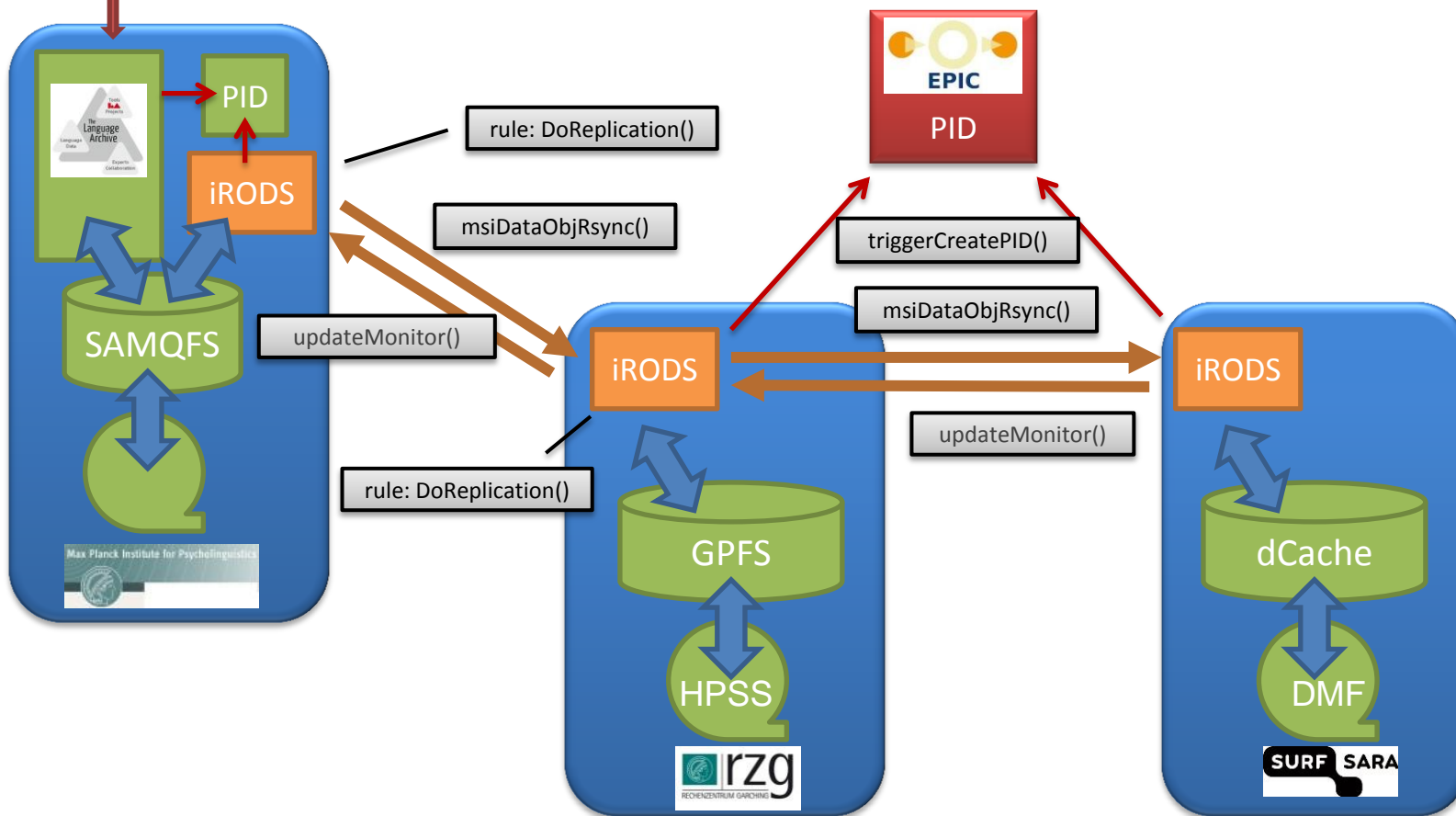
Use Case: CLARIN B2SAFE



`/vzMPI/archive/...`
`/vzRZG/eudat/clarin/project1/...`
`/vzSARA/eudat/clarin/project1/...`

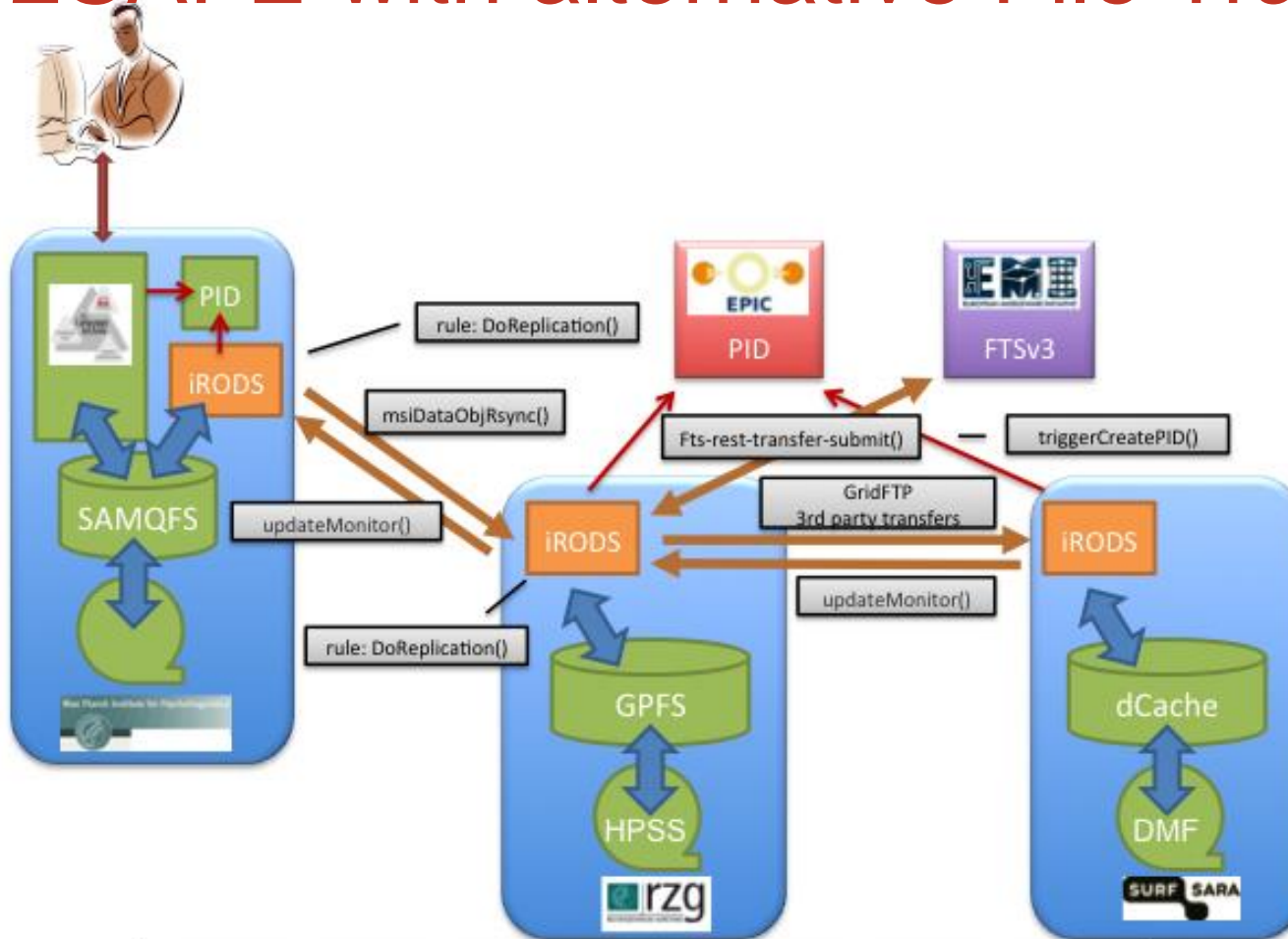


Use Case: CLARIN B2SAFE



```
doReplication(*pid,*source,*destination,*status) {
  msiDataObjRsync(*source, "IRODS_TO_IRODS", "null", *destination, *rsyncStatus);
  triggerCreatePID(*collectionPath*child.pid.create", *pid, *destination);
  updateMonitor(*collectionPath*filepaths\*.pid.update");
}
```

B2SAFE with alternative File Transfer



```
doReplication("pid,"source,"destination,"status) {
  msiExecCmd("fts-rest-transfer-submit", ""source "destination", "null", "null", "null", *out);
  triggerCreatePID("collectionPath*filepathslash.pid.create", "pid,"destination);
  updateMonitor("collectionPath*filepathslash.pid.update");
}
```

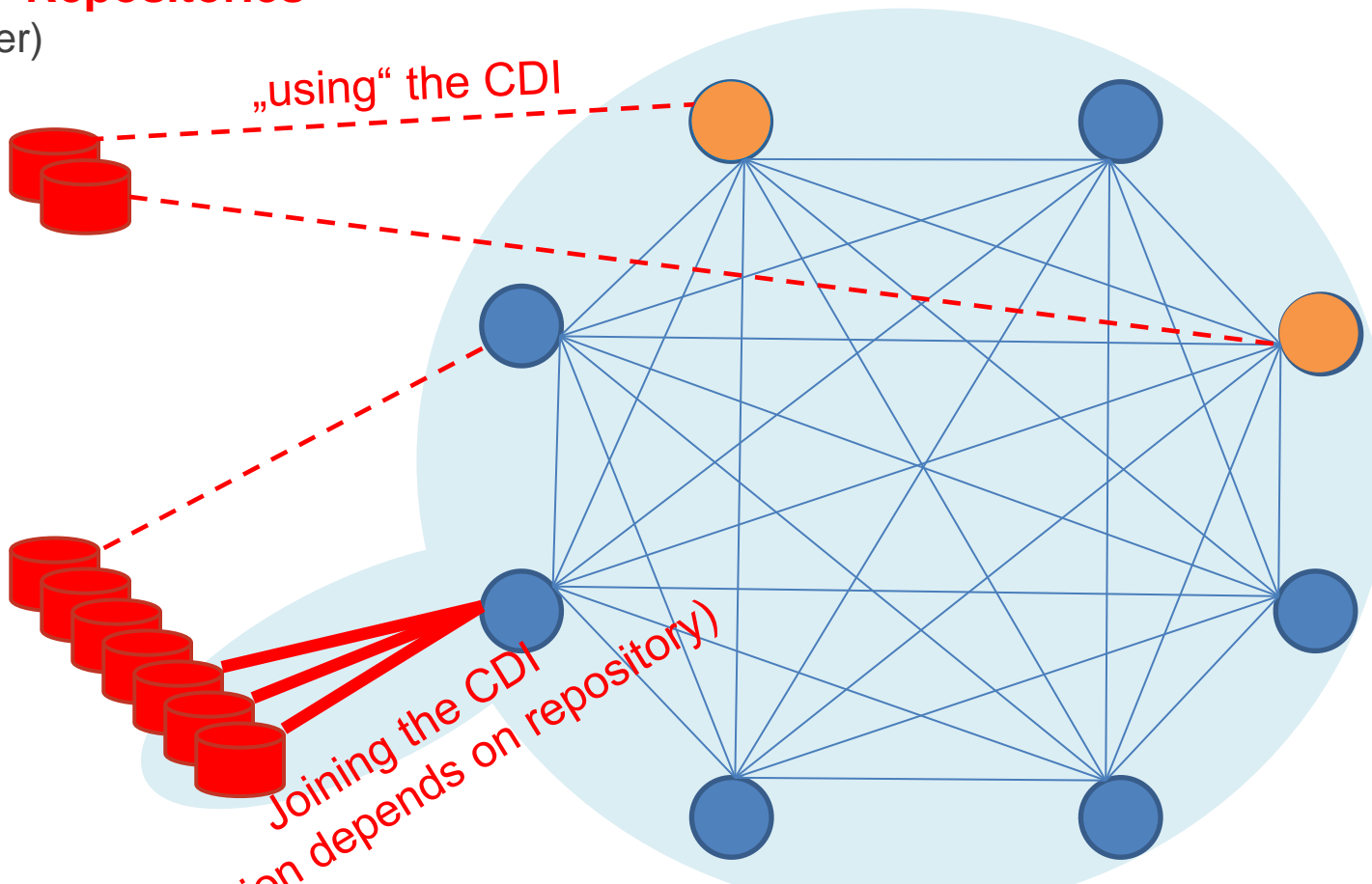


Different community-specific approaches: Using or Joining the CDI

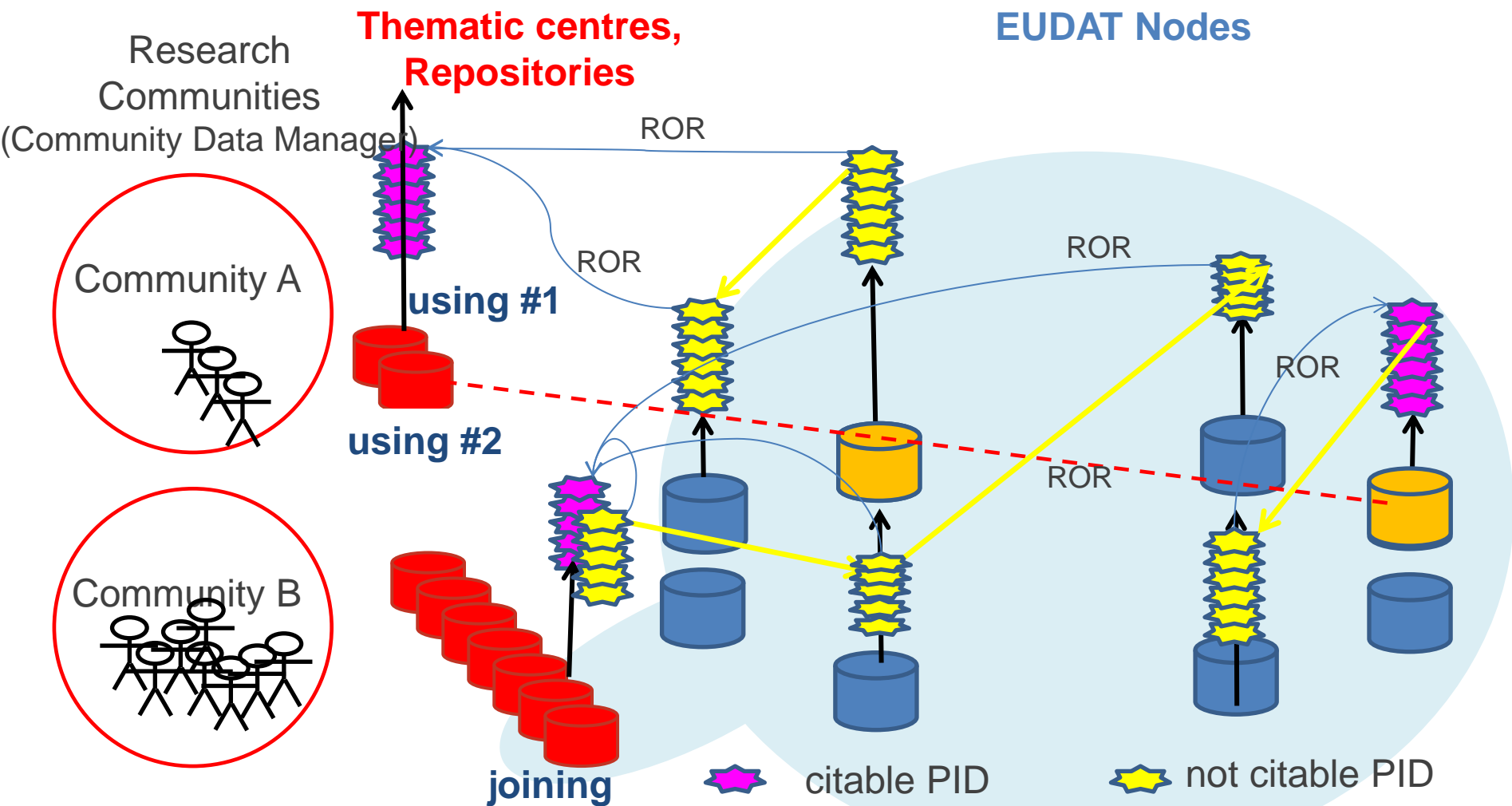
Research Communities
(Community Data Manager)

**Thematic Nodes
Repositories**





EUDAT Nodes



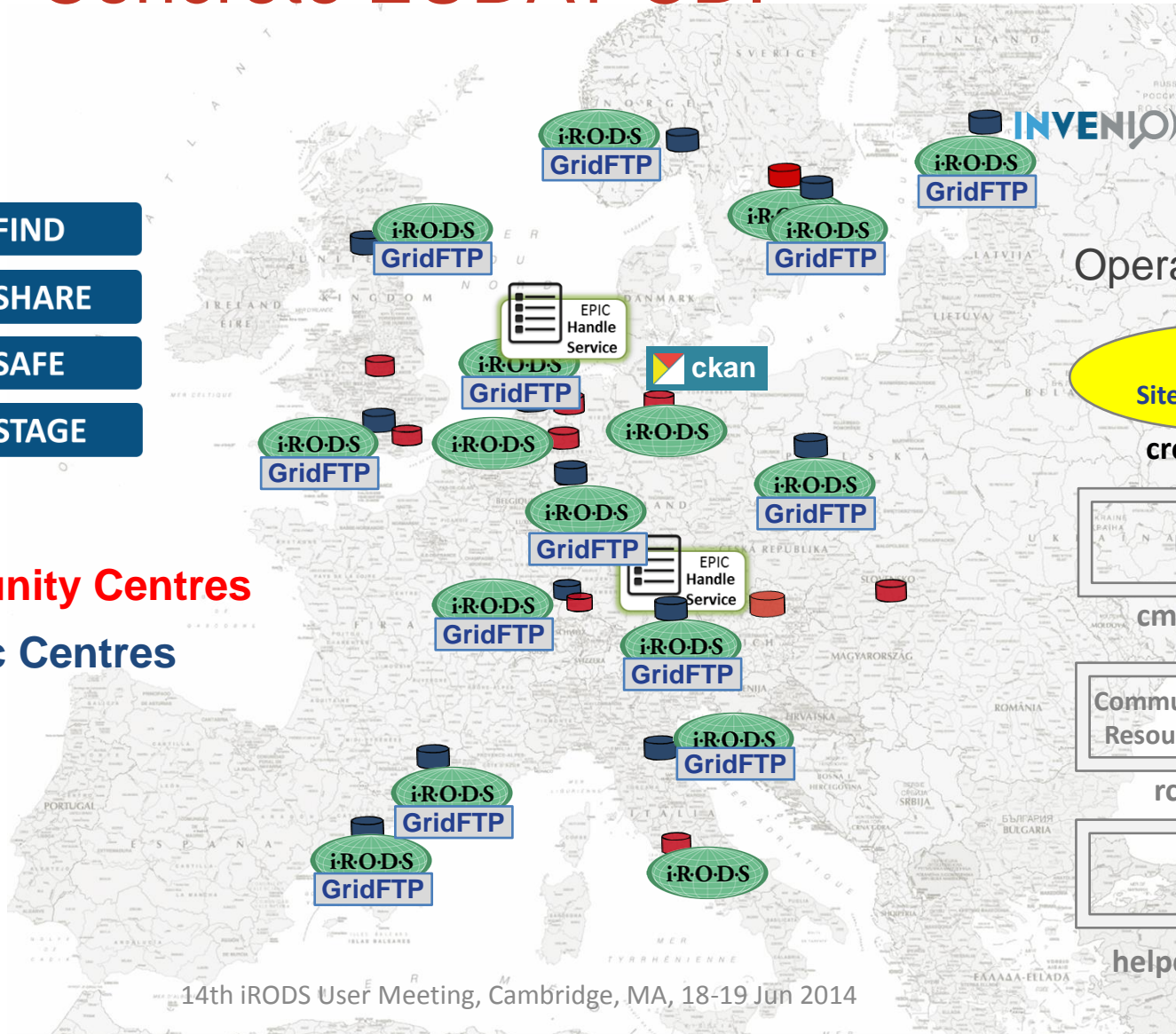
Different options of PID provisioning and linking



Concrete EUDAT CDI

-  B2FIND
-  B2SHARE
-  B2SAFE
-  B2STAGE

-  **Community Centres**
-  **Generic Centres**



Operational Tools

REGISTRY
Sites and Services
creg.eudat.eu

Monitoring
cmon.eudat.eu

Community Data Project
Resource Coordination
rct.eudat.eu

Helpdesk
helpdesk.eudat.eu

List of iRODS services



EUDAT
powered by
GOCDDB5+

Browse

- My Sites
- Admin Domains
- Projects
- Sites
- Service Groups
- Services

Add

- Add Site
- Add Service Group
- Add Service
- Add Downtime

Downtimes

- Active & Imminent

About GOCDDB5

- Doc, Help & Support

Search

Submit

User Status

Registered as:
Johannes Reetz

View Details
Manage Roles



Services

All Services in GOCDDB

Filter (clear)

Service Type: Project: Search:

Production: Monitored: Scope: Certification: Include Closed Sites:

Extension Name:

20 Services (Showing 1 - 30)

Hostname	Service Type	Production	Monitored	Scope(s)	Host Site
» ariane.cines.fr	iRODS	✓	✓	EUDAT	CINES
» bscirops.bsc.es	iRODS	✗	✓	EUDAT	BSC
» clarin.uni-tuebingen.de	iRODS	✓	✓	EUDAT	EKUT
» data.repo.cineca.it	iRODS	✓	✓	EUDAT	CINECA
» dtn01.hector.ac.uk	iRODS	✓	✓	EUDAT	EPCC
» ed-res-01.csc.fi	iRODS	✓	✓	EUDAT	CSC
» ed-res-02.csc.fi	iRODS	✓	✓	EUDAT	CSC
» eudat-icat.pdc.kth.se	iRODS	✓	✓	EUDAT	PDC
» eudat-node1.esc.rl.ac.uk	iRODS	✗	✓	EUDAT	STFC
» eudat-s1.fz-juelich.de	iRODS	✓	✓	EUDAT	JUELICH
» eudat1.dkrz.de	iRODS	✓	✓	FIIDAT	DKRZ



Operational Tools

REGISTRY
Sites and Services

creg.eudat.eu

Monitoring

cmon.eudat.eu

Community Data Project
Resource Coordination

rct.eudat.eu

Helpdesk

helpdesk.eudat.eu



List of Service Groups for Scoping Domains of Services

EUDAT
powered by
GOCD85+

Service Groups
All Service Groups in GOCD85.
What is a service group?

Filter (clear)
Scope: EUDAT Extension Name: (none)

23 Service Groups

Name	Description	Scope(s)
» EUDAT_COLLABORATIVE_SERVICES	EUDAT Collaborative Services	EUDAT
» EUDAT_CORE_AAI_SERVICES	Service endpoints which belong or are related to the EUDAT AAI services	EUDAT
» EUDAT_CORE_B2FIND	EUDAT Joint Metadata Service	EUDAT
» EUDAT_CORE_B2SAFE	All service endpoints which belong or are related to the EUDAT B2SAFE service	EUDAT
» EUDAT_CORE_B2SHARE	EUDAT Simple Store Services	EUDAT
» EUDAT_CORE_B2STAGE	Service endpoints which belong or are related to the EUDAT core service B2STAGE	EUDAT
» EUDAT_CORE_EPICPID	The EUDAT coordinated core PID service using the EPIC service	EUDAT
» EUDAT_CORE_MYPROXY	EUDAT coordinated core MyProxy services	EUDAT
» EUDAT_CORE_SR4CLARIN	Service endpoints which belong to the EUDAT B2SHARE for all CLARIN repositories	EUDAT
» EUDAT_CORE_SR4CLARIN_CUNI	Service endpoints which belong to the EUDAT core service "Safe Replication" for CLARIN project: CUNI (Repository: LINDAT)	EUDAT
» EUDAT_CORE_SR4CLARIN_EKUT	Service endpoints which belong to the EUDAT core service "Safe Replication" for CLARIN project: EKUT (Repository: EKUT)	EUDAT
» EUDAT_CORE_SR4CLARIN_REPLIX	Service endpoints which belong to the EUDAT core service "Safe Replication" for CLARIN project: REPLIX (Repository: MPI-PL)	EUDAT
» EUDAT_CORE_SR4ENES	Service endpoints which belong to the EUDAT core service	EUDAT
» EUDAT_CORE_SR4EPOS	Service endpoints which belong to the EUDAT core service "Safe Replication" for EPOS	EUDAT
» EUDAT_CORE_SR4INCF	Service endpoints which belong to the EUDAT core service	EUDAT



Operational Tools

REGISTRY
Sites and Services

creg.eudat.eu

Monitoring

cmon.eudat.eu

Community Data Project
Resource Coordination

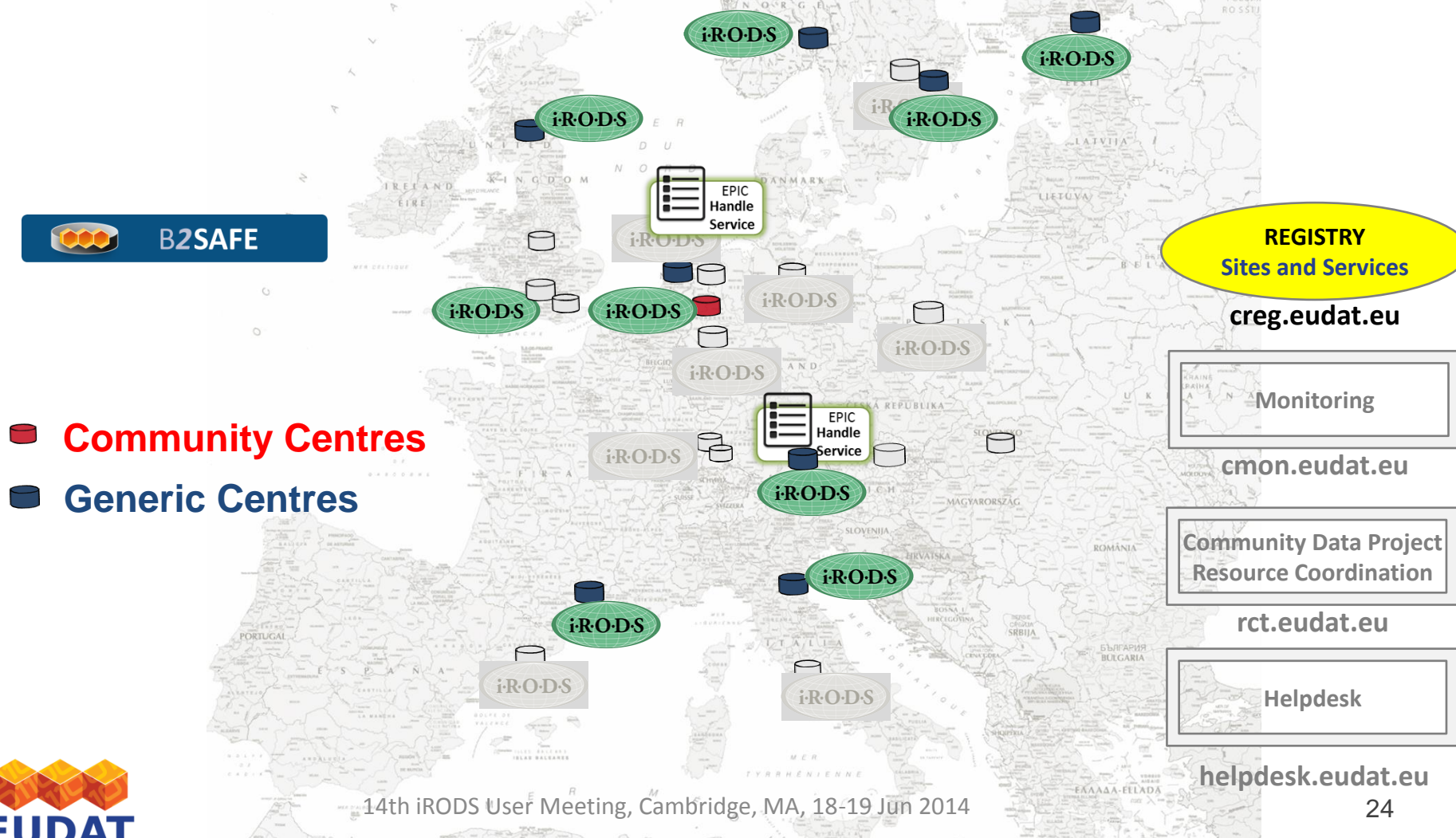
rct.eudat.eu

Helpdesk

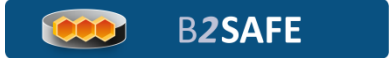
helpdesk.eudat.eu

Islands of Trust 1

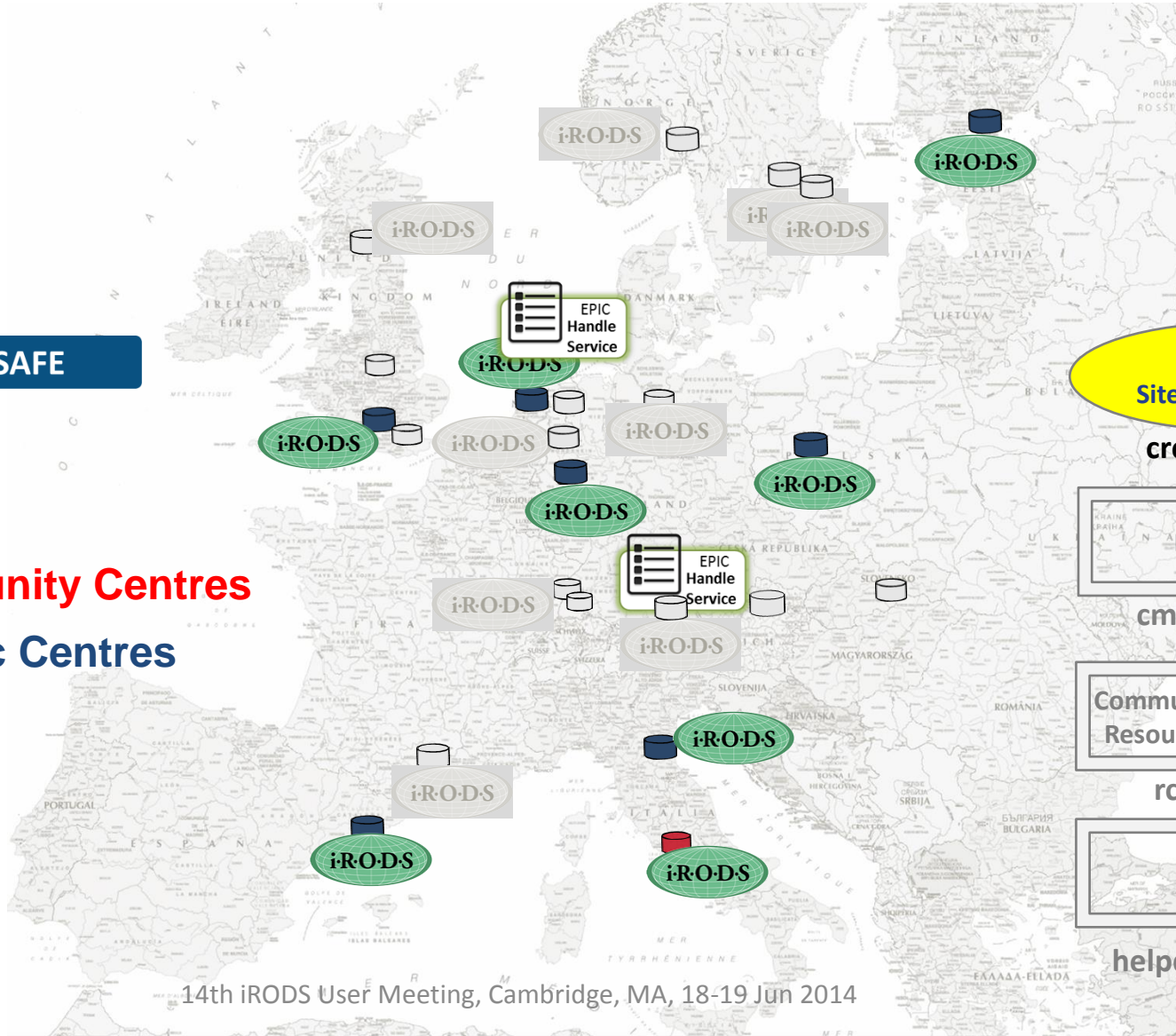
Repositories/Depositors need control over where their data is stored and managed.



Islands of Trust 2



-  **Community Centres**
-  **Generic Centres**



REGISTRY
Sites and Services

creg.eudat.eu

Monitoring

cmon.eudat.eu

Community Data Project
Resource Coordination

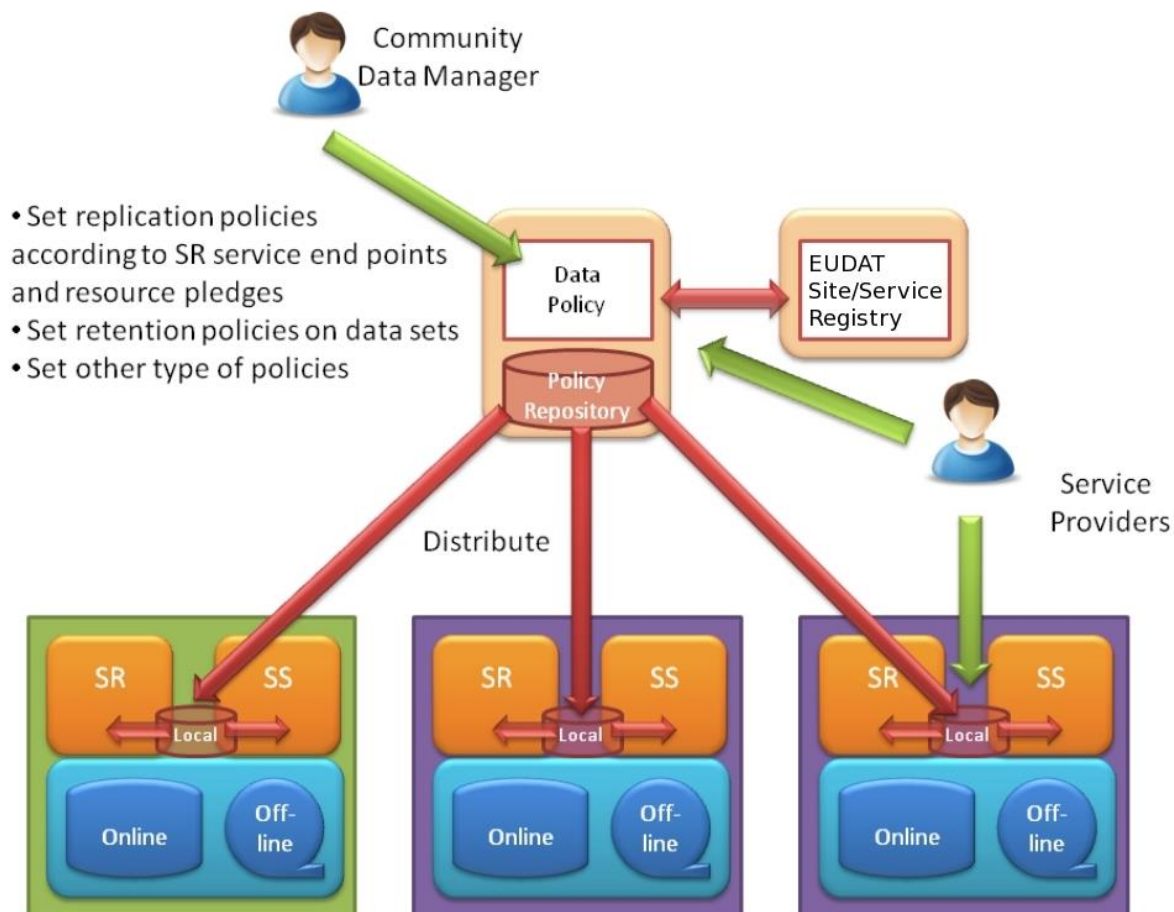
rct.eudat.eu

Helpdesk

helpdesk.eudat.eu



Data Policy Manager



Objectives of the DPM

- allow Community Managers (CM) to specify data management policies
- allow Community Managers (CM) to manage (define, assign, monitor) data management policies on distributed administrative zones via a web-portal

Replication

EUDAT



1839/abc
29db...279b4a
1.10.13 00:00

CM wants to specify the type of replication, the replica sites and how often data shall be replicated

Replication from A to B and from B to C
Replication from A to B and from A to C

456/abc
29db...279b4a
1.10.13 02:00

789/abc
29db...279b4a
1.10.13 02:00

Integrity

EUDAT



1839/abc
29db...279b4a
1.10.13 00:00

CM wants to specify the periodicity of integrity checking

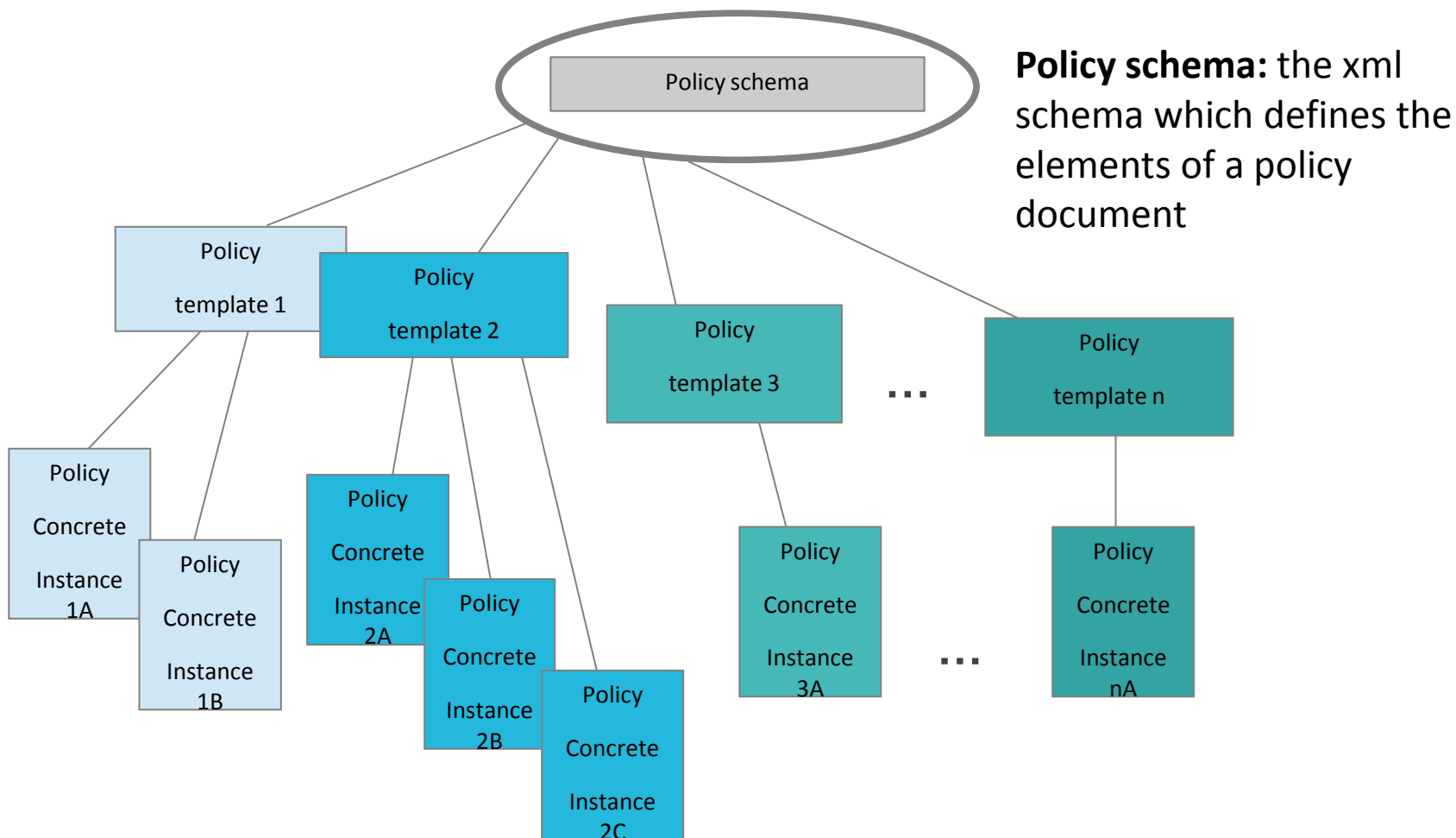
456/abc
29db...279b4a
1.10.13 02:00

789/abc
29db...279b4a
5.10.13 05:00

Checksum recalculation on the physical file



Policy hierarchy



Abstract Policy (template)

- **Policy template:** the policy document which defines a policy process, but without specific parameters.
- Therefore the tasks are defined, but without input/output parameters. For example:

```

• <dataset>
•   <collection id="0">
•     <persistentIdentifier type="PID"></persistentIdentifier>
•   </collection>
• </dataset>
• <actions>
•   <action name="replication onchange">
•     <type>replicate</type>
•     <trigger>
•       <action>modify object</action>
•     </trigger>
•     <targets>
•       <target id="0">
•         <location xsi:type="irodsns:coordinates">
•           <irodsns:site type="EUDAT"></irodsns:site>
•           <irodsns:path></irodsns:path>
•           <irodsns:resource></irodsns:resource>
•         </location>
•       </target>
•     </targets>
•   </action>
• </actions>

```

Define data sets

Policy type

Define action

Target descriptons

Concrete Policy (Instance)

- **Policy instance:** the policy document which defines a policy process, but with specific parameters.

```

<dataset>
  <collection id="0">
    <persistentIdentifier type="PID">
      11100/6c8ac19e-c982-11e2-b3cb-e41f13eb41b2
    </persistentIdentifier>
  </collection>
</dataset>
<actions>
  <action name="replication onchange">
    <type>replicate</type>
    <trigger>
      <action>modify object</action>
    </trigger>
    <targets>
      <target id="0">
        <location xsi:type="irodsns:coordinates">
          <irodsns:site type="EUDAT">CINECA</irodsns:site>
          <irodsns:path>/path/to/destination</irodsns:path>
          <irodsns:resource>defaultResc</irodsns:resource>
        </location>
      </target>
    </targets>
  </action>
</actions>

```


Web-based portal for managing Data Policies

- tool to be used by community/repository data managers
- independent from any protocols
- integrated with EUDAT services which apply data management policies
- authentication and authorization (security)
- policy instances according to research domain requirements

B2SAFE core

Replication supporting

- Synchronization based on checksum, timestamp, size
- Multiple geographically distributed replicas with possibility to choose the locations among the nodes of the EUDAT CDI
- Registration of data through persistent identifiers (PIDs)
- Auditable policy rules
- Both legacy and open standard protocols for data transfers
- Multiple back-ends for data storage

B2SAFE core work

- Integration with Data Policy Manager (DPM)
- Implementation of ACL mechanism for policy enforcement
- Integrity check mechanisms adoption
- Integration with AAI framework
- Consolidation of an EUDAT policy set to harmonize data management and PID registration.

EUDAT Rules

- Rules for Replication and PID handling

getEpicApiParameters

getSharedCollection

writeFile

logInfo

logDebug

logError

logWithLevel

readFile

updateCommandName

updateMonitor

retrieveChecksum

triggerReplication

triggerCreatePID

triggerUpdateParentPID

processReplicationCommandFile

processPIDCommandFile

doReplication

createPID

createPIDgriffin

addPIDWithChecksum

searchPID

searchPIDchecksum

CheckReplicas

updatePIDWithNewChild

getRorPid



Remarks concerning EUDAT AAI requirements and the EUDAT AAI pilot

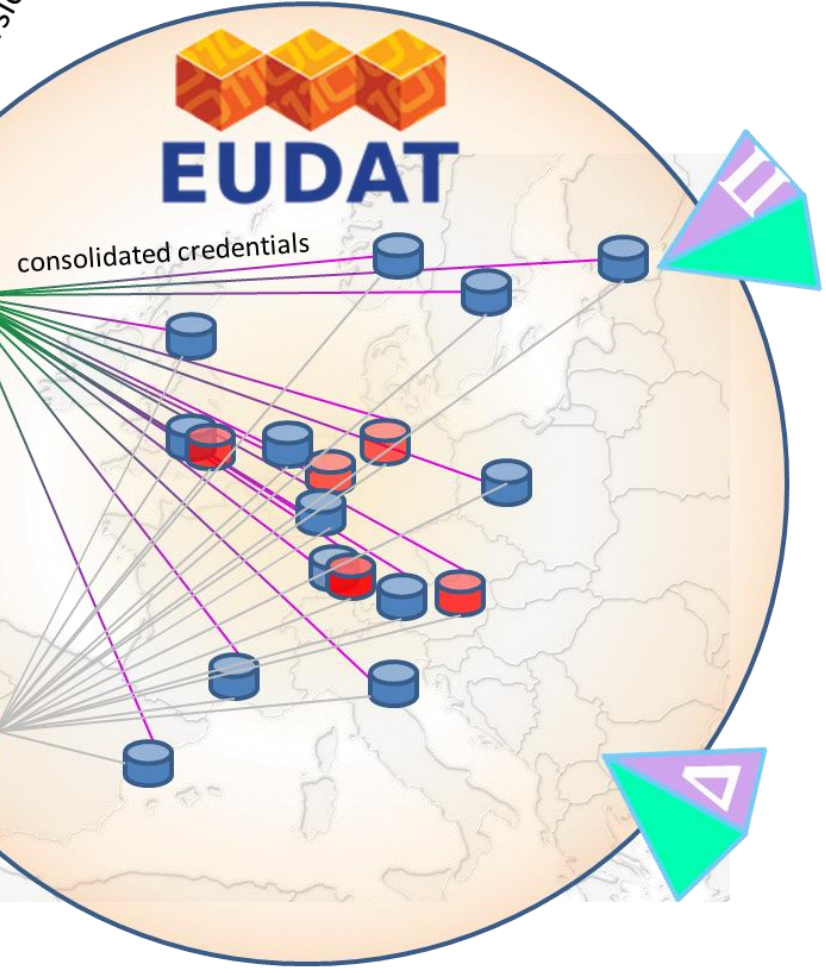
Different types of Identity Providers
AuthN



- zoned credential conversion service
- unique user Ids, project-wise mapped to
- attribute based access control information

Identity credential conversion

COMMUNITIES

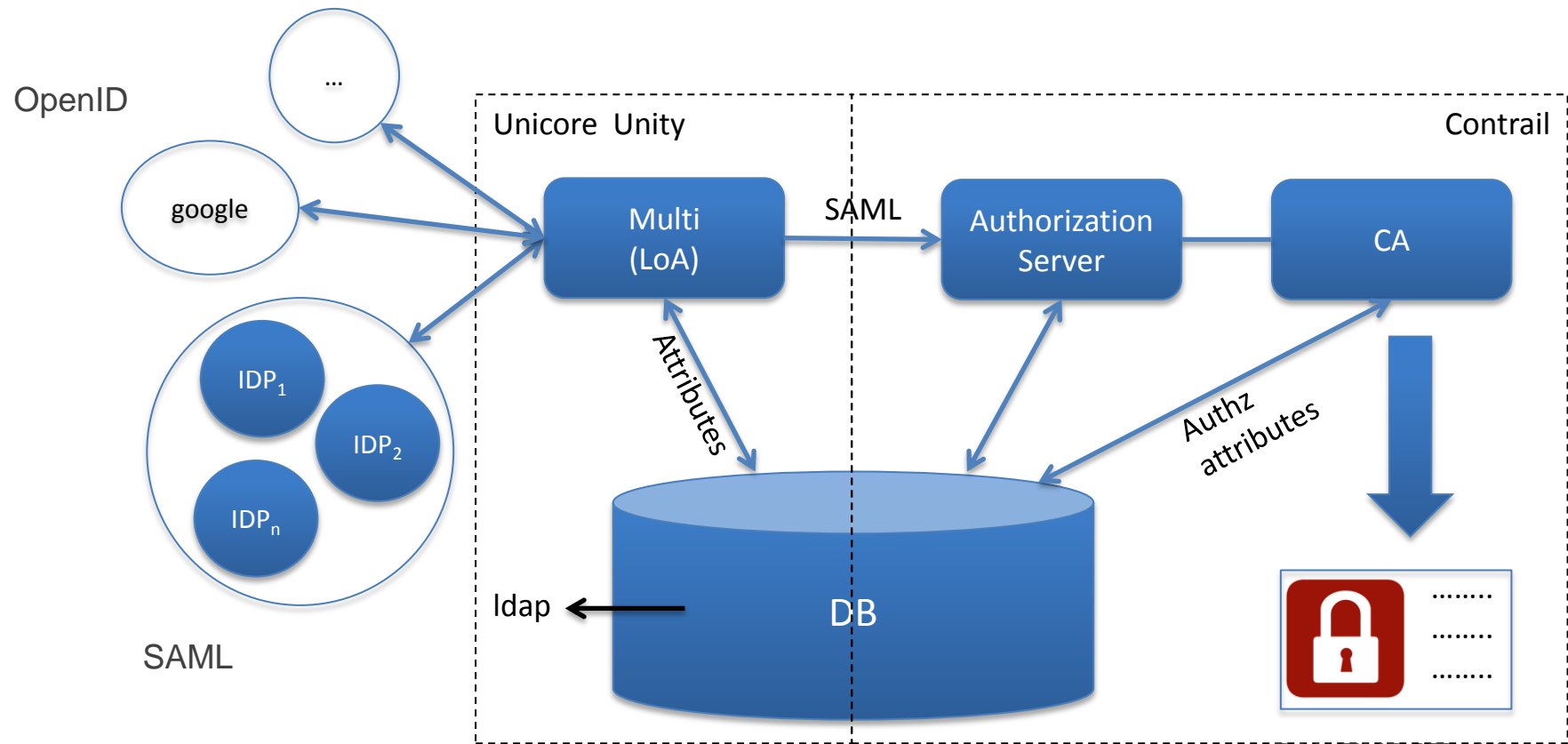


Attribute Provider **AuthZ**

either community-managed or (*) attributes provided by user's home IdP are reused



Currently piloted EUDAT AAI



DN: EUDAT uid
 Attributes:
 • Community uid
 • ...

References

Claudio Cacciari, CINECA

Stephane Coutin, CINES

Francesca Iozzi, Adil Hasan, UiO

Willem Elbers, MPI for Psycholinguistics

Johannes Reetz, RZG/MPS

Robert Verkerk, SURFsara; Dejan Vitlacil, KTH/PDC

Giuseppe Fiameni, Giacomo Mariani, CINECA

Shaun de Witt, STFC; Martin Helmich, CERN

Jens Jensen, STFC; Shiraz Memon

John Kennedy, RZG/MPS

Mark van de Sanden, SURFsara, NL

...

B2SAFE core (MS, Rules,..)

B2SAFE/Repository packages

Data Policy Manager

Data Management Policies

B2SAFE product

PID Services in EUDAT

B2STAGE product

HTTP/B2SAFE interface

EUDAT AAI

EUDAT Site and Service Reg.

Technologies, Service Building