



EUDAT

Towards a Collaborative Data Infrastructure - Short Training on Key Principles -

Dr.-Ing. Morris Riedel
Federated Systems and Data
Juelich Supercomputing Centre (JSC)



Adjunct Associate Professor
School of Engineering and Natural Sciences
University of Iceland



UNIVERSITY OF ICELAND





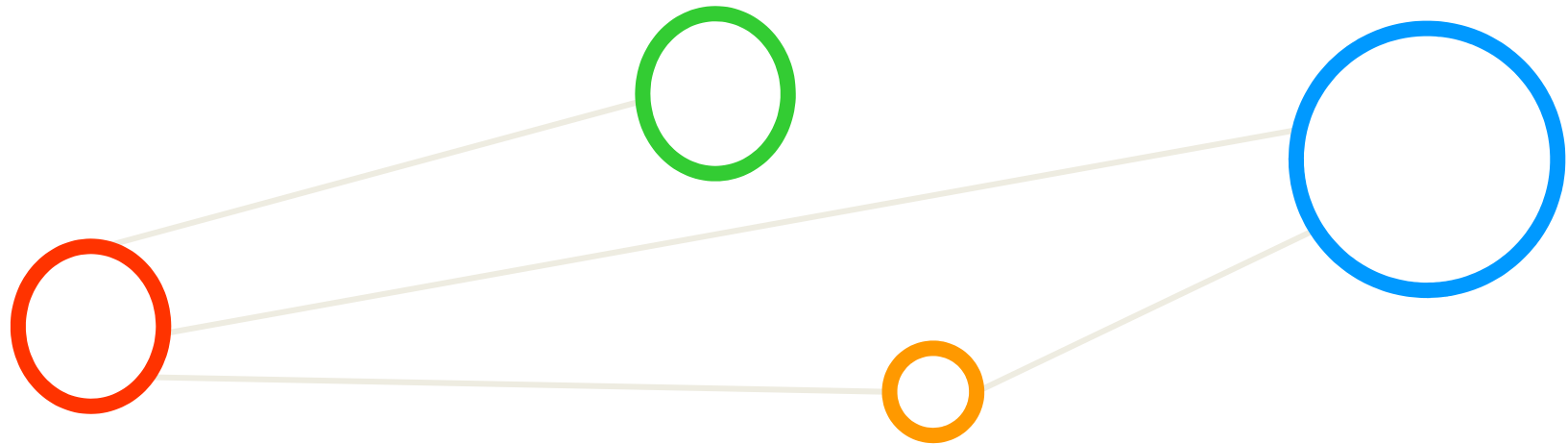
Outline

- Motivation & EUDAT 101
- Short Training on Key Principles
 - How to create a collaborative data infrastructure
 - How to create a registered domain of data
 - How to perform policy-based data replication
- Summary & Actions
- References





Motivation & EUDAT 101



Big Data Waves – Surfboards – Breakwaters

How can we manage the rising tide of scientific data

High Level Expert Group on Scientific Data Report

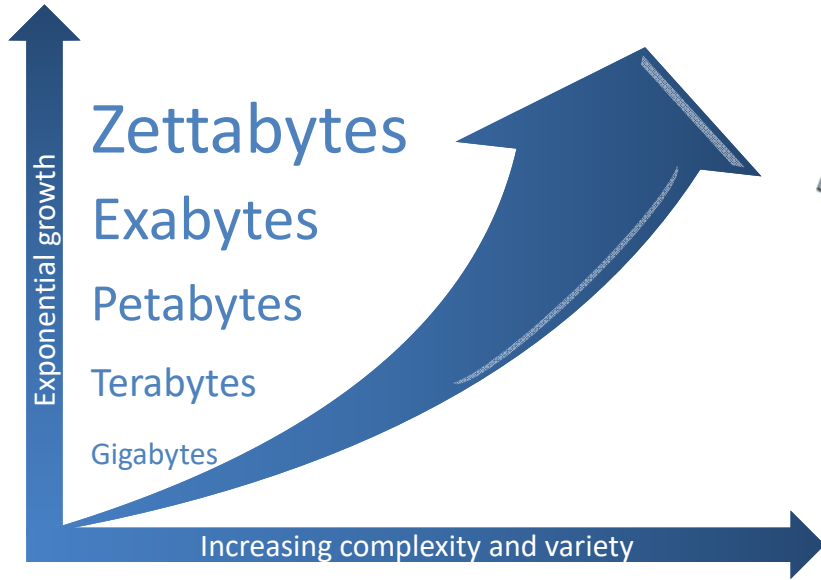
Lists unsolved questions
Outlines challenges
Provides visions

A Surfboard for Riding The Wave Report

Lists 4 key action drivers
Identifies 3 strategic goals
Clarifies Data Scientists

*'Concrete'
Next Steps →*





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



- How to ensure interoperability?

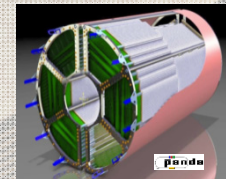
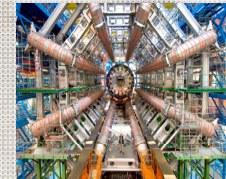
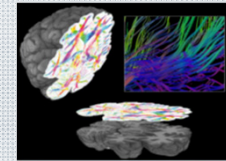




In commercial environments Big Data is all about **Volume – Variety – Velocity**

‘Big Data is data that becomes large enough that it cannot be processed using conventional methods.’

[1] O'Reilly Radar Team, 'Big Data Now: Current Perspectives from O'Reilly Radar'



The next generation radio telescope for science...

The square kilometre array

... 1 PB in 20 seconds



LOFAR

test site Jülich

EUDAT – Collaborate to tackle 'big data'

[2] EUDAT Web page





Training & Working habits with Communities



MINDSET

Do we all think that PIDs for scientific data are important?

Do we agree that safe replication of data is important for unique scientific data sets?



SKILLSET

How do we use PIDs and what type of PID structures are relevant?

What are the techniques to perform data replication and how it relates to PIDs and metadata?



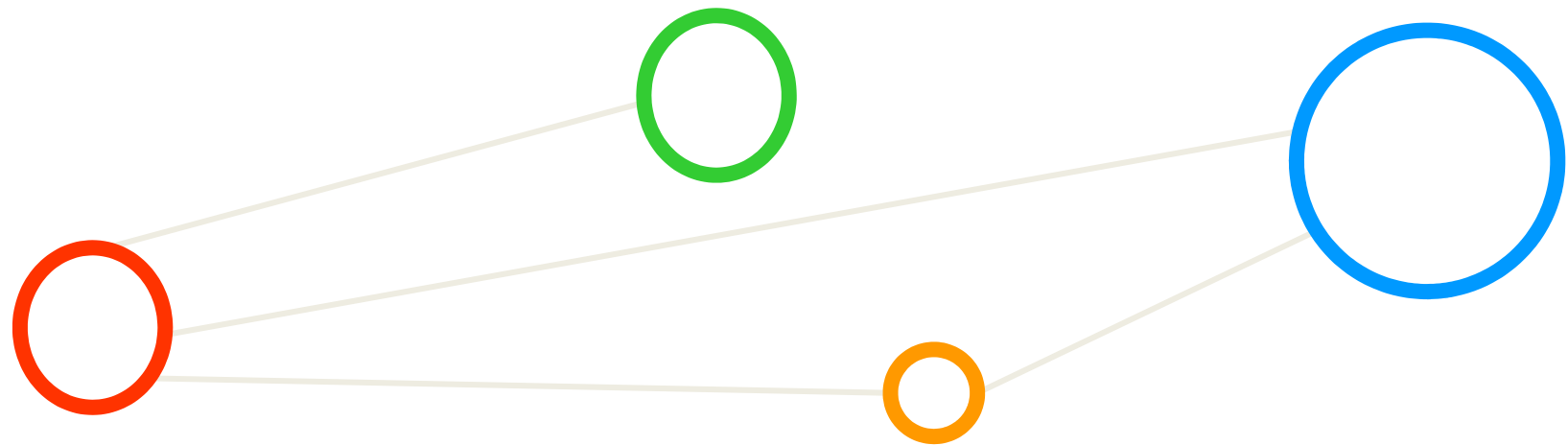
TOOLSET

Do we use (common) services and tools to work with PIDs being part of community practice?

Are there established (common) data replication services and tools available we can use daily?

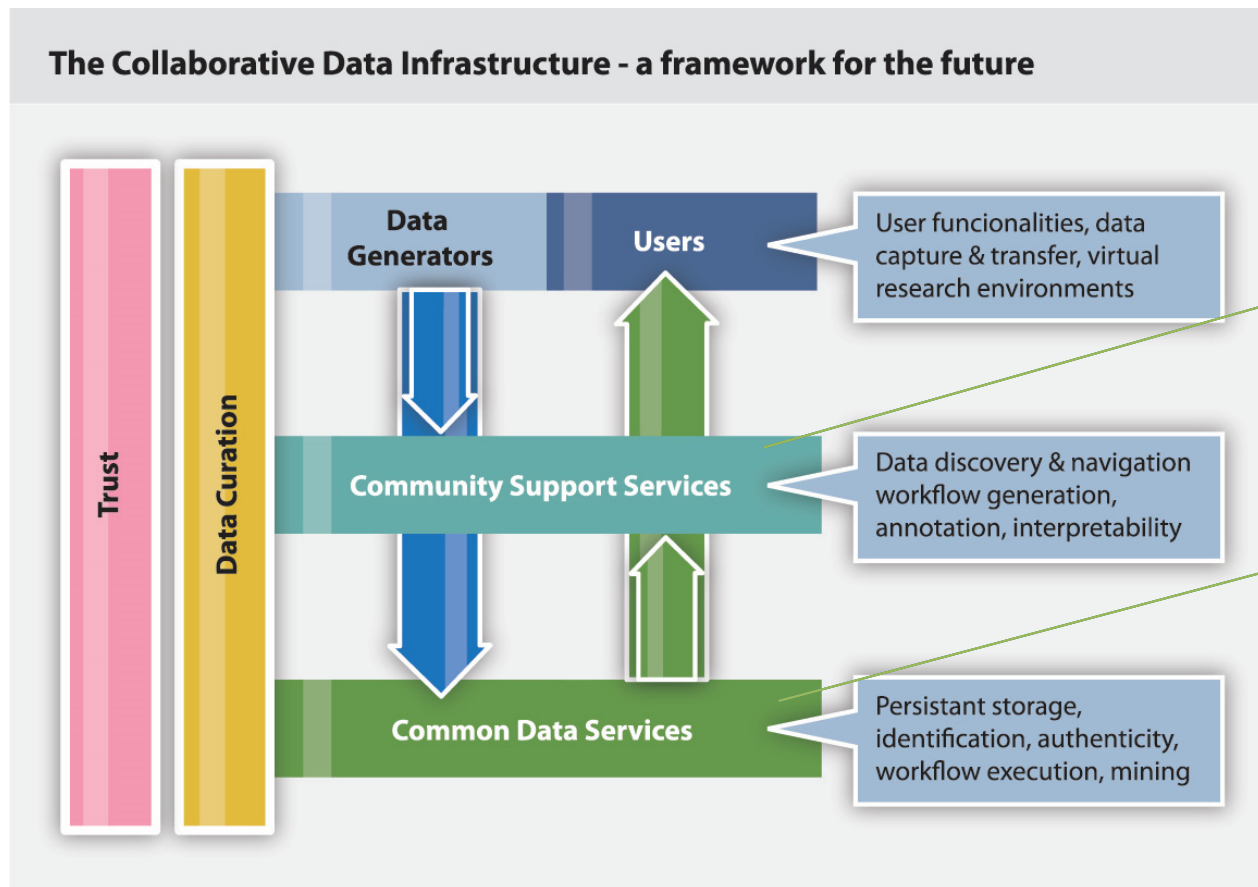


How to create a collaborative data infrastructure



Training on ‘Skillset Level’

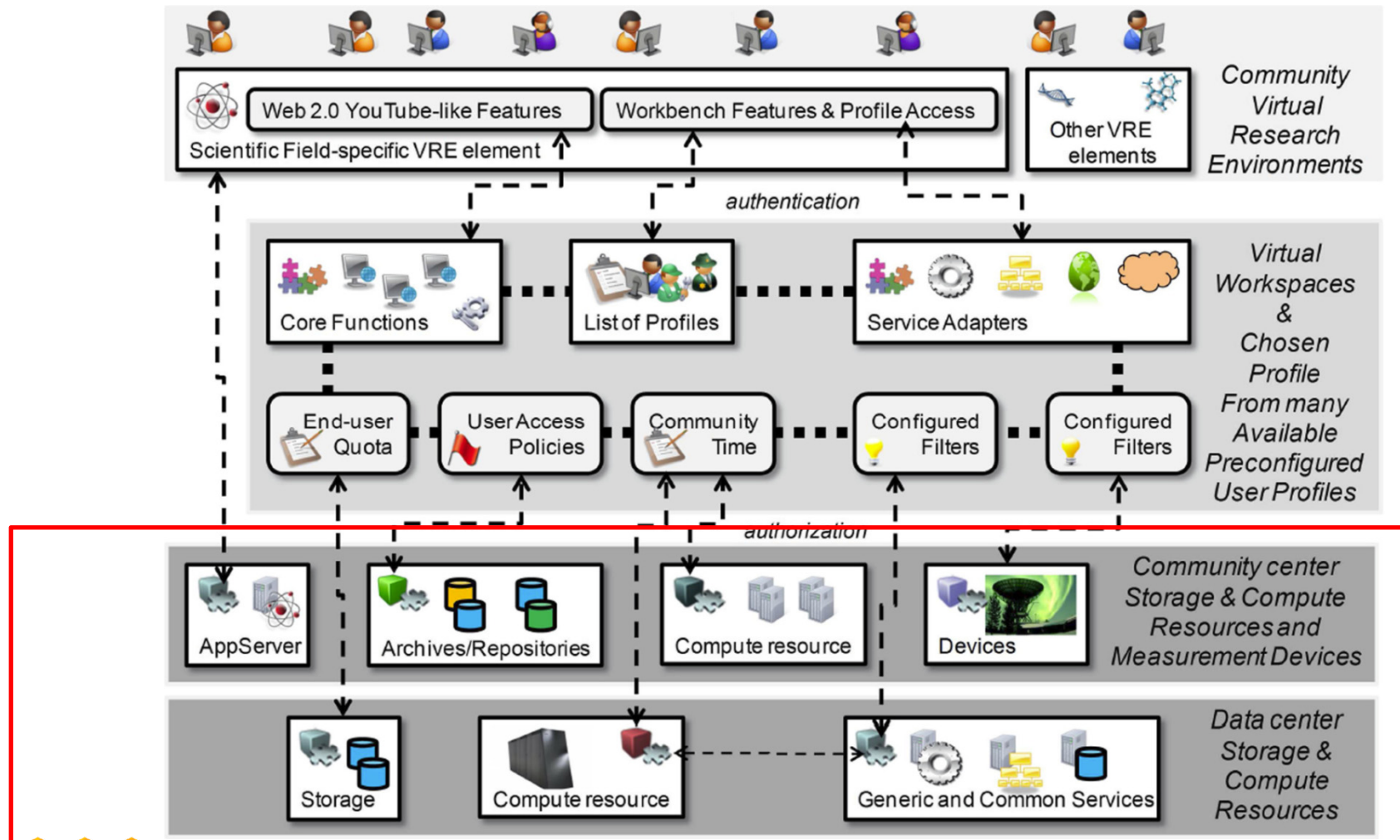
Blueprint of a Collaborative Data Infrastructure



CLARIN, LifeWatch, ENES, EPOS, VPH, etc.
5 Core Infrastructures
more second round infrastructures

=> 12 EUDAT data centers

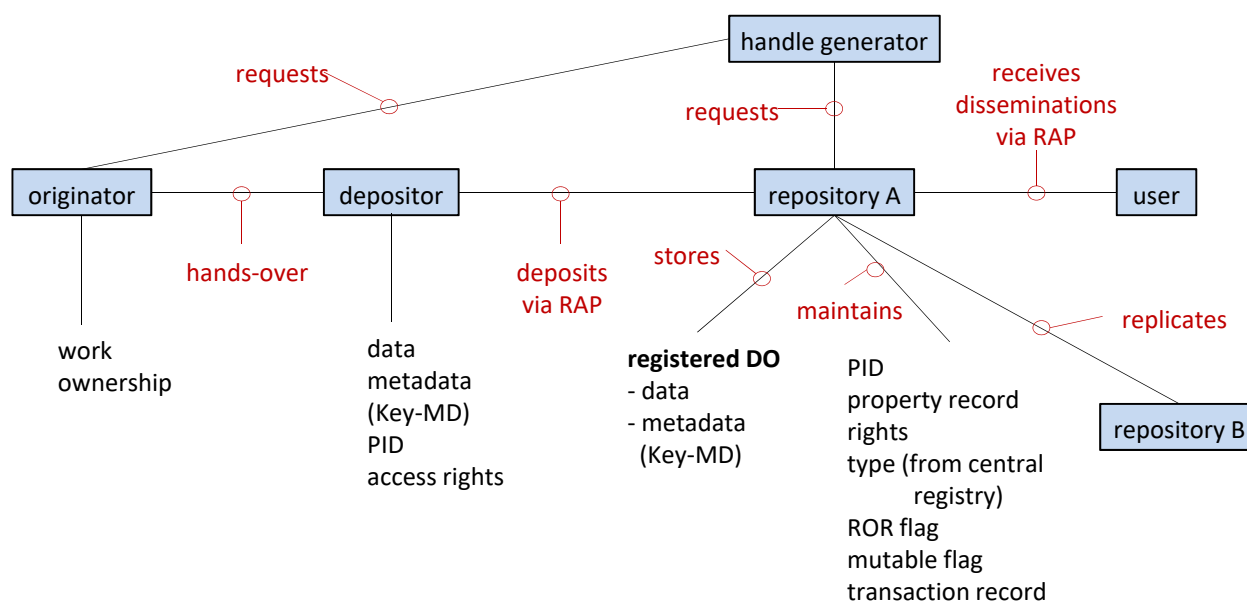
Conceptual View of a CDI





Analysis of existing community data infrastructure

- ❑ community interactions based on abstract model (Kahn & Wilensky, 2006)
 - 'triple': Data + Metadata + Handle (PID) – use it as 'orientation point'!
- ❑ used in many meetings and interactions - accepted quickly as reference model
- ❑ helped even in improving community organization plans





Definitions/Entities

originator = creates digital works and is owner;
depositor = forms work into DO (incl. metadata),
digital object (DO) = instance of an abstract data type;
registered DOs are such DOs with a Handle;
repository (Rep) = network accessible storage to store DOs;
RAP (Rep access protocol) = simple access protocol
Dissemination = is the data stream a user receives
ROR (repository of record) = the repository where data was stored first;
Meta-Objects (MO) = are objects with properties
mutable DOs = some DOs can be modified
property record = contains various info about DO
type = data of DOs have a type
transaction record = all disseminations of a DO



[2] EUDAT Web page

Example: EUDAT Centres

-  community centre
-  EUDAT centre





Clear Task: Identify Common Services

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



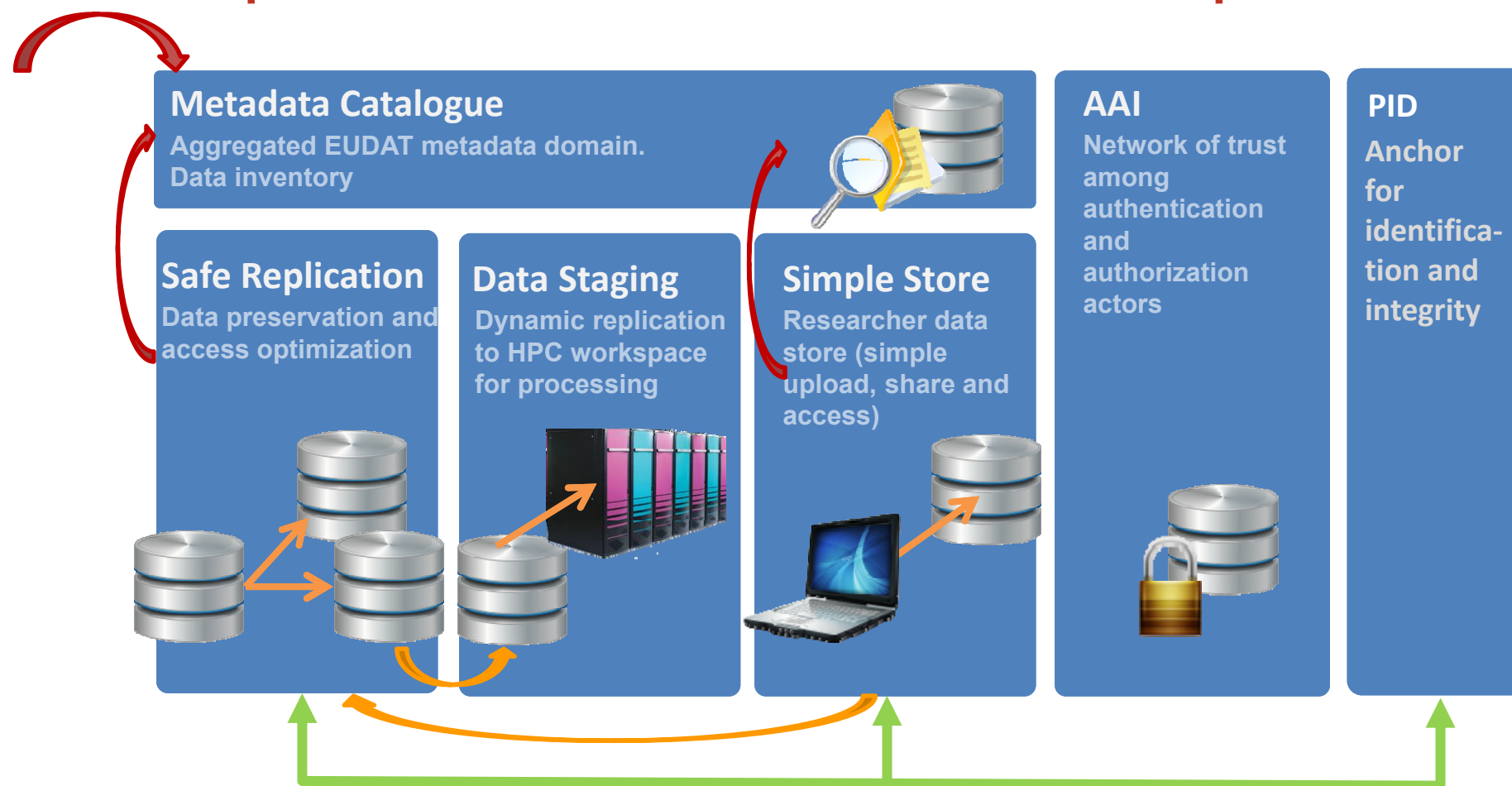


Example: Current EUDAT Services Focus

Common Services	CLA RIN	LW	VPH	EN ES	EP OS	IN CF	EC RIN	Bio Vel	Dixa	CESS DA	DAR IAH	Pan Data	BB MRI	EM SO
Safe Replication	X	o	X	X	X	X			x		x			
Data Staging	o	o	X	X	X									
SimpleStore	X	X	X	X	X	x	x	x	x	x	x		x	
Metadata	X	X	o	X	x	x	x	x	x	x	x	x	x	x
Web-service platform	X	o		X	o									

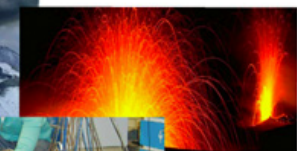
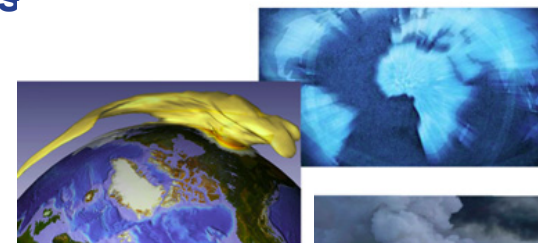
X = needed now, x = interested, o = interest, not direct priority

Example: EUDAT Services in Preparation

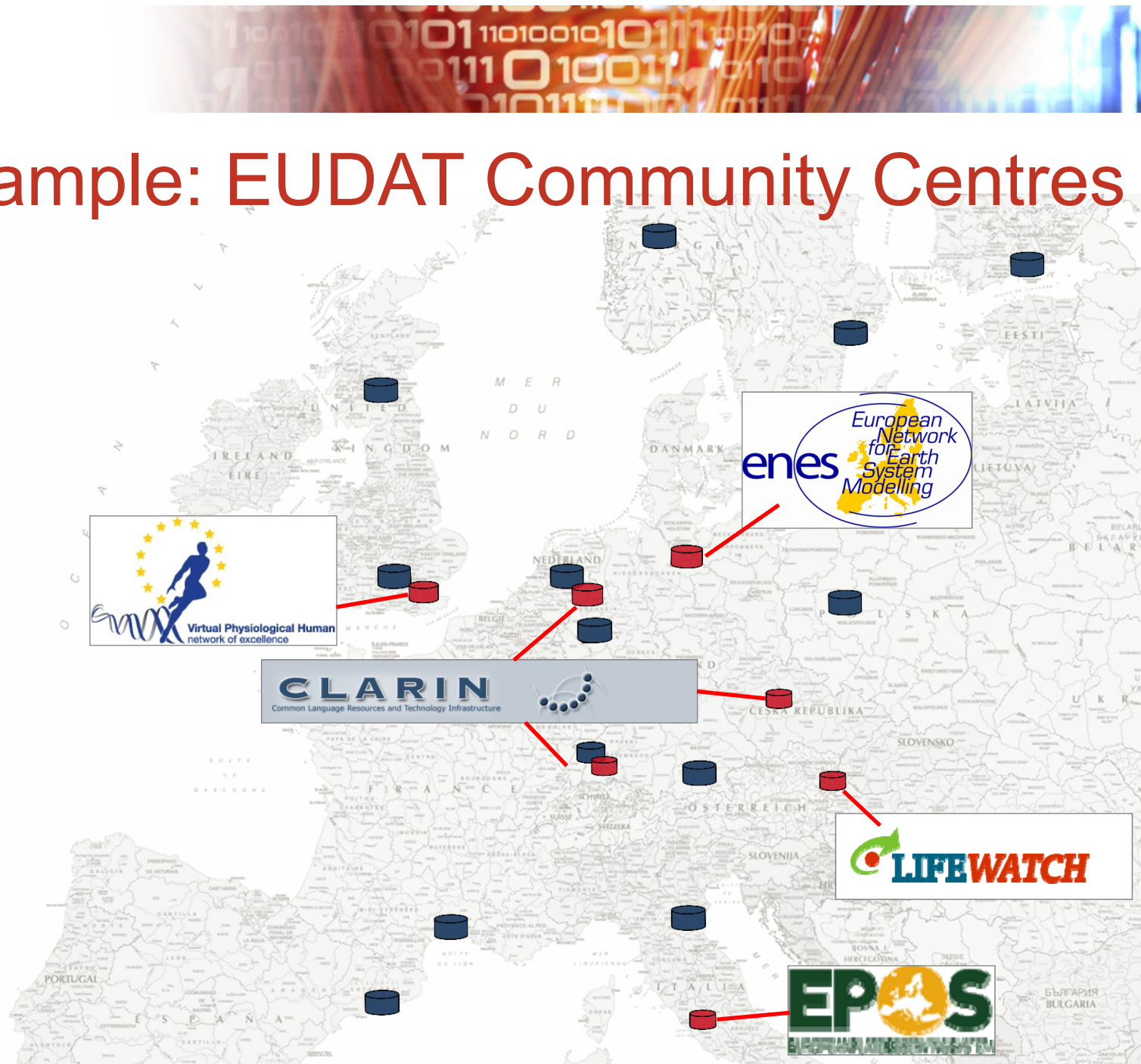


Example: EUDAT user communities

- **EPOS**: European Plate Observatory System
- **CLARIN**: Common Language Resources and Technology Infrastructure
- **ENES**: Service for Climate Modelling in Europe
- **LifeWatch**: Biodiversity Data and Observatories
- **VPH**: The Virtual Physiological Human
- **All share common challenges:**
 - Reference models and architectures
 - Persistent data identifiers
 - Metadata management
 - Distributed data sources
 - Data interoperability



Example: EUDAT Community Centres



'ScienceTube': User perspective of CDI

The screenshot displays the ScienceTube web interface. At the top, there's a search bar with 'Data wave' entered and buttons for 'Suchen', 'Kategorien', and 'Video hochladen'. Below the search bar, a video player shows a 3D pie chart titled 'Scientific Evaluation of an accurate measurements last year'. The pie chart has six segments with values: 8 (purple), 5 (red), 10 (green), 4 (blue), 10 (pink), and 6 (grey). To the right of the video player is a 'Vorschläge' (Suggestions) sidebar with video thumbnails.

Below the video player, there are interaction buttons: 'Mag ich', a comment icon, and a view count of '7.557.864 Aufrufe'. Below this is a comment section with two comments from 'Prof. Dr. Knowm' and 'Citizen from Berlin'.

On the right side of the interface, there are several panels:

- Warning:** A yellow warning icon and text: '76% of your data in this storage is unused since 90 days – remove data?'
- Available Resources (Filter #443):** A list of resources with checkboxes: STORAGE: (RZG), STORAGE: (JUELICH), and STORAGE: (RZG).
- Available Services (Filter #43):** A list of services with checkboxes: Data-mining service and Workflow service.
- 10:46 - CEST Meetings Today:** A clock icon and text: '9:00 Steering Group, see Meeting Minutes'.
- sensor ship #4455** and **sensor airplane #4711** with corresponding colored squares.
- HPC: (SARA)**, **HPC: (JUELICH)**, **HTC: (EGI)**, **STORAGE: (JUELICH)**, and **STORAGE: (RZG)** with corresponding colored squares.
- Most Viewed Today:** Three small charts: 'MPI toolset usage' (a 3D pie chart), 'Paper about aging' (a bar chart), and 'Device #46 data' (a bar chart).

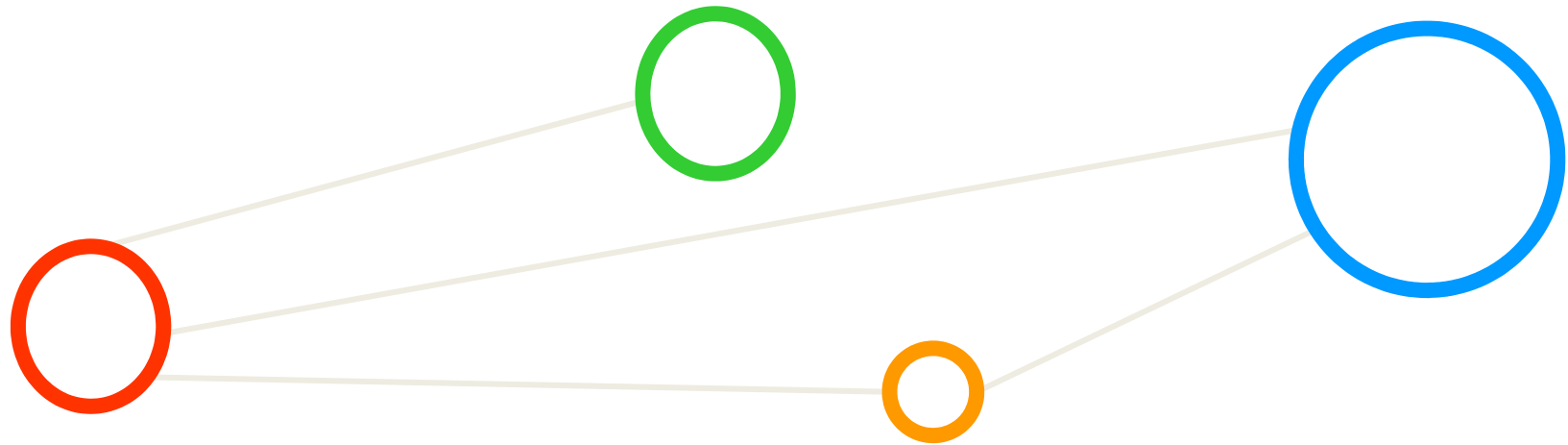


Lessons Learned in this Training Section

- ✓ **Accept that many communities have already a data infrastructure, so we need to connect it**
- ✓ **Knowing triple to organize/understand data plans**
- ✓ **Understand the major blueprint of a Collaborative Data Infrastructure (CDI)**
- ✓ **Capable of identifying common data services**
- ✓ **Knowing the difference between mono-thematic community center and multi-disciplinary centers**

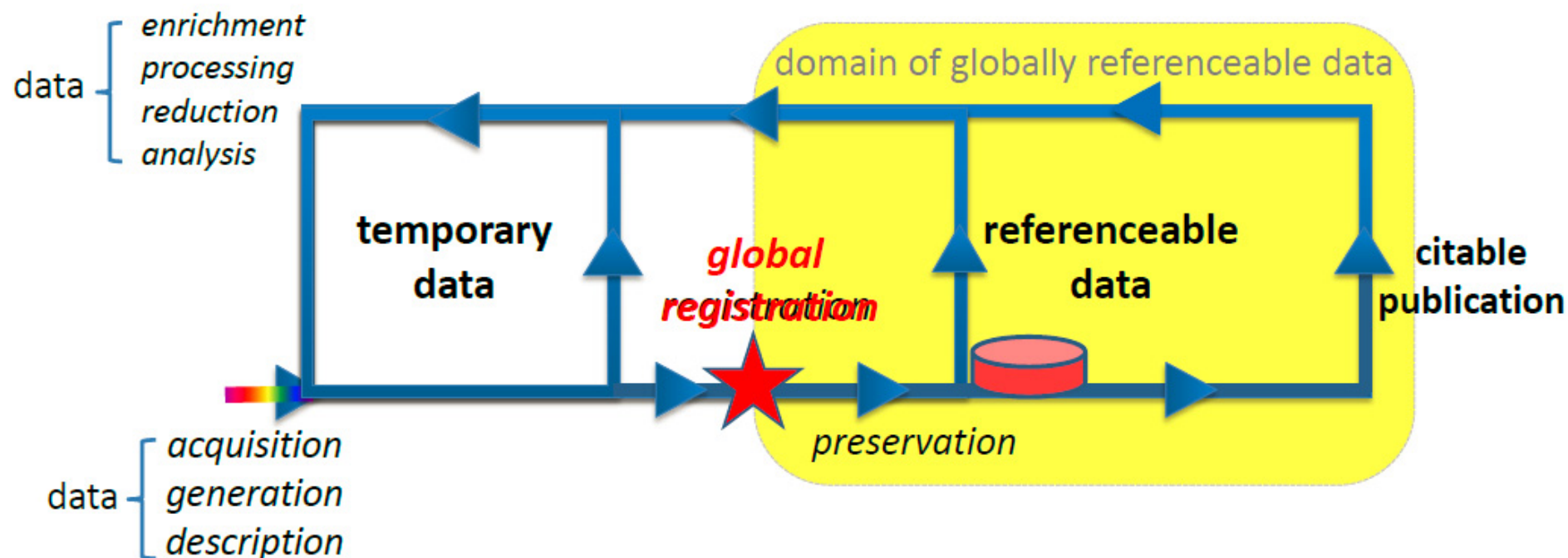


How to create a registered domain of data

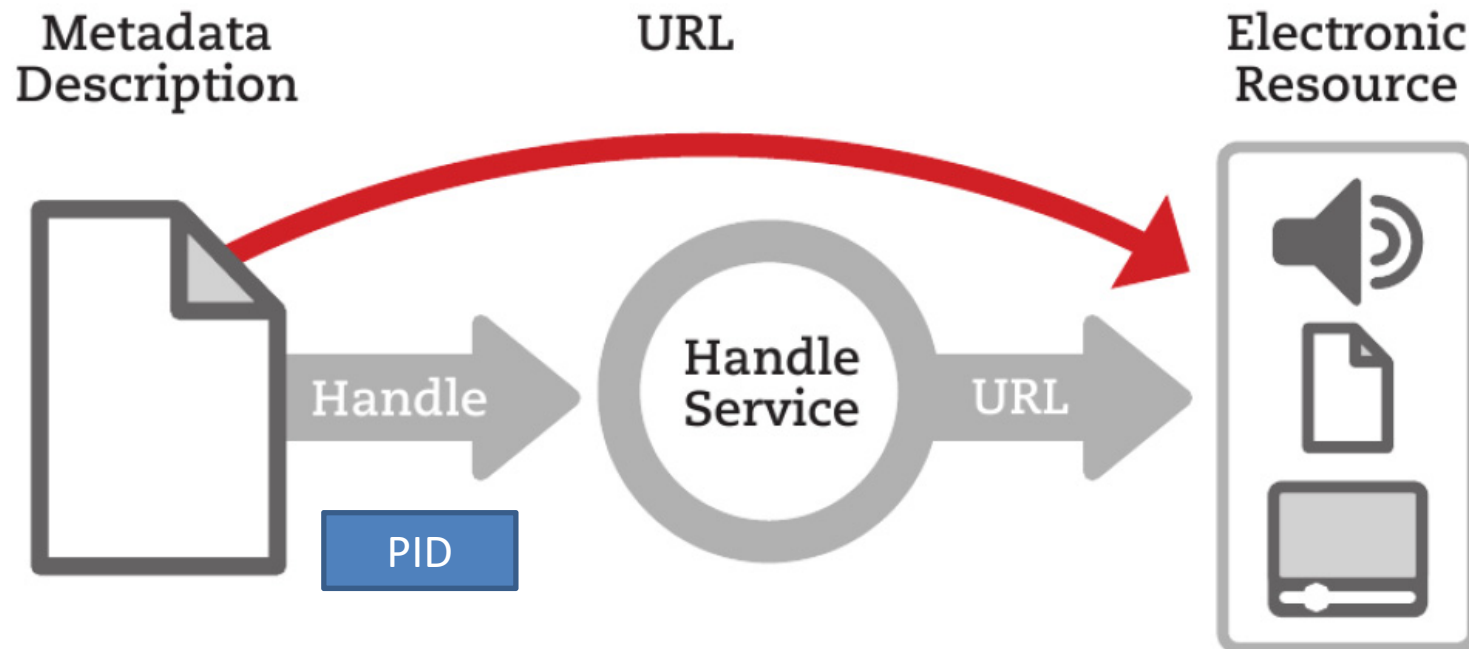


Training on ‘Skillset Level’

Blueprint for a registered domain of data



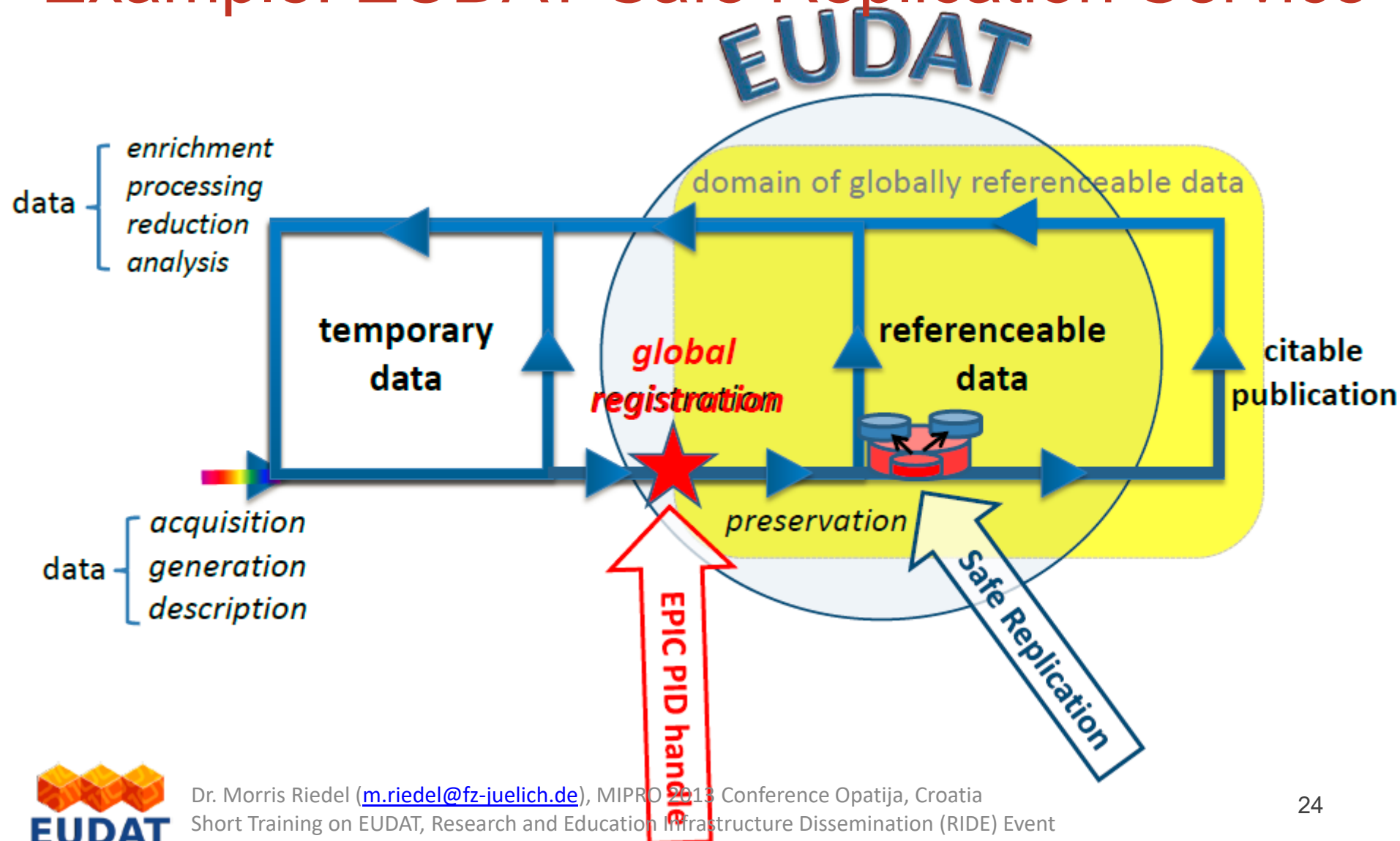
Key Principle of Handle System



- URN, ARK, Handle, DOI, PURL (by HTTP-redirect)
- Critical: Resolution



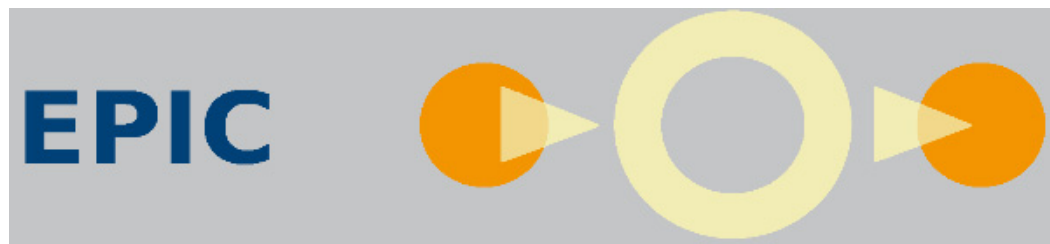
Example: EUDAT Safe Replication Service



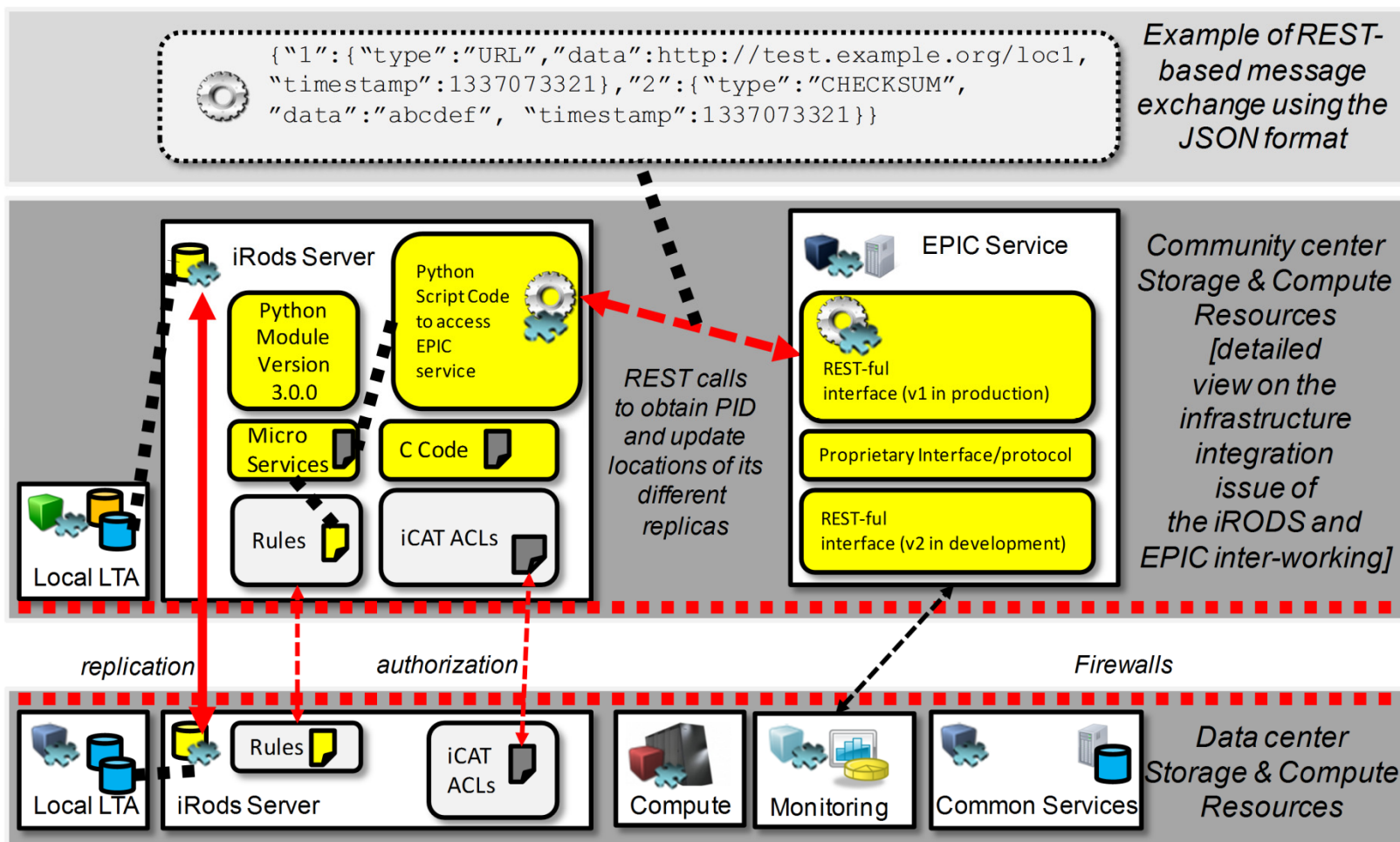


Use Persistent Identifiers to Identify Data

- Use Persistent Identifiers (PIDs)
 - Based on the Handle System
 - Used to reference data, including different locations
- Requires a PID Service
 - One example is the EPIC PID service
 - E.g. register a PID specifying a URI
 - EPIC = European Persistent Identifier Consortium



Example: EUDAT Use of EPIC Service



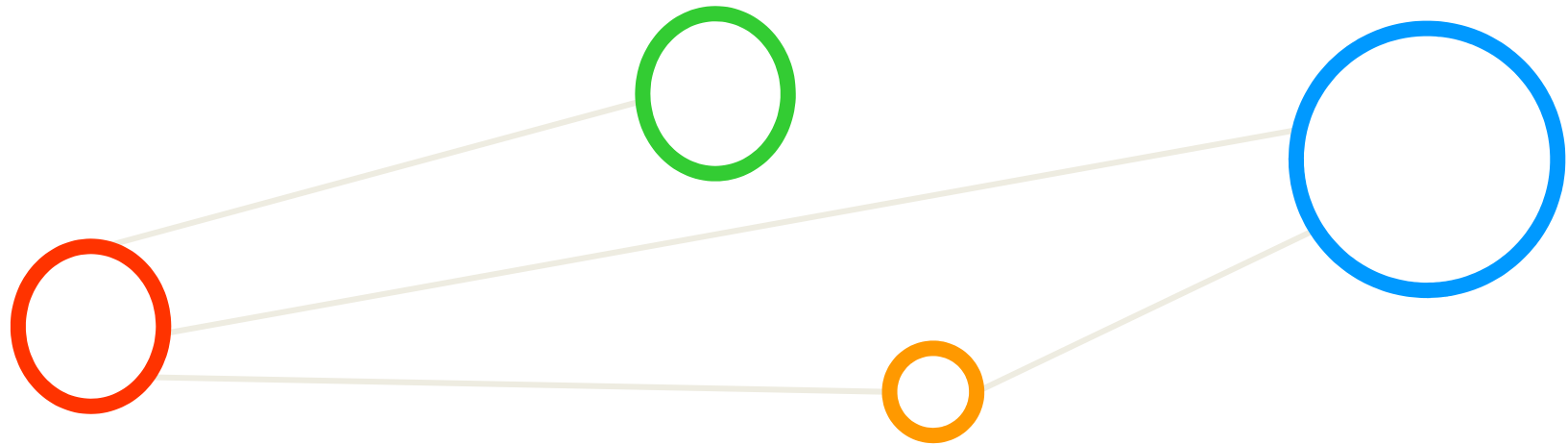


Lessons Learned in this Training Section

- ✓ Understand the structure of one possible registered domain of (scientific) data
- ✓ Accept that the handle system is a pragmatic way to identify data not bound to location
- ✓ Knowing that you need Persistent Identifier (PIDs) as reference to digital objects (data)
- ✓ Capable of creating a theoretical use case that is using PIDs and an associated PID service (EPIC)

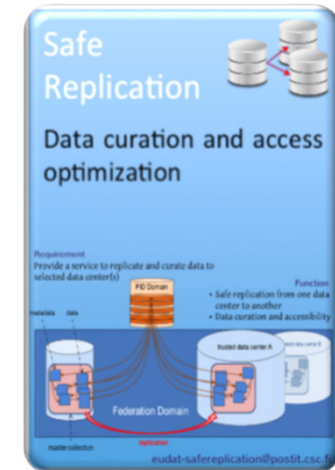
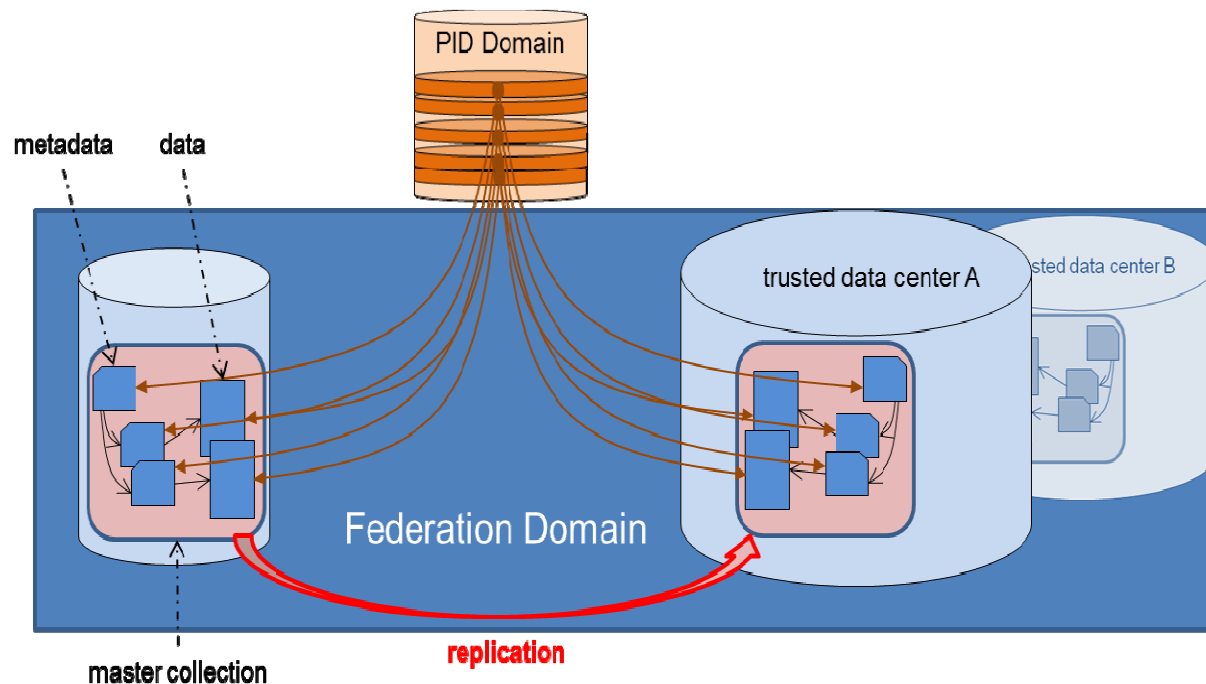


How to perform policy-based data replication



Training on ‘Skillset Level’

Blueprint for safe data replication



Better accessibility of scientific data

Make data referencable

High degrees of reliability and trust

More optimal data curation



Safe Data Replication Approach

- Idea: Safe replication between 1 scientific community center and N data centers
 - Replication within a 'registered domain of data' (i.e. PID assignment)
- Flexibility, scalability and management require policy-based data management (i.e. rule engine)
 - With local policies at centers and global policies for infrastructure(s)
- Islands (community + data centers) in parallel & close interaction → merge?
 - Enabling community as process for acknowledging existing data management plans of communities



Example EUDAT: Forming strong relationships

EPOS - European Plate Observatory System


- Distributed data sensors
- Large scale statistics
- Metadata schema
- Reference architecture



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

ENES - Service for Climate Modeling in Europe

- About 20 centers in EU
- CIM data model
- Using CDI @ German Climate Center
- Using DOIs and EPIC
- Metadata based on ISO 11179



ENES provides information and services to foster intricate simulations of the climate system using high performance computers as well as the distributions and dissemination of data produced by such simulations

CLARIN - Common Language Resources and Technology Infrastructure

- About 200 centers in EU
- Require PIDs, CMDI
- ISOcat, SCHEMcat
- Virtual Language Obs.

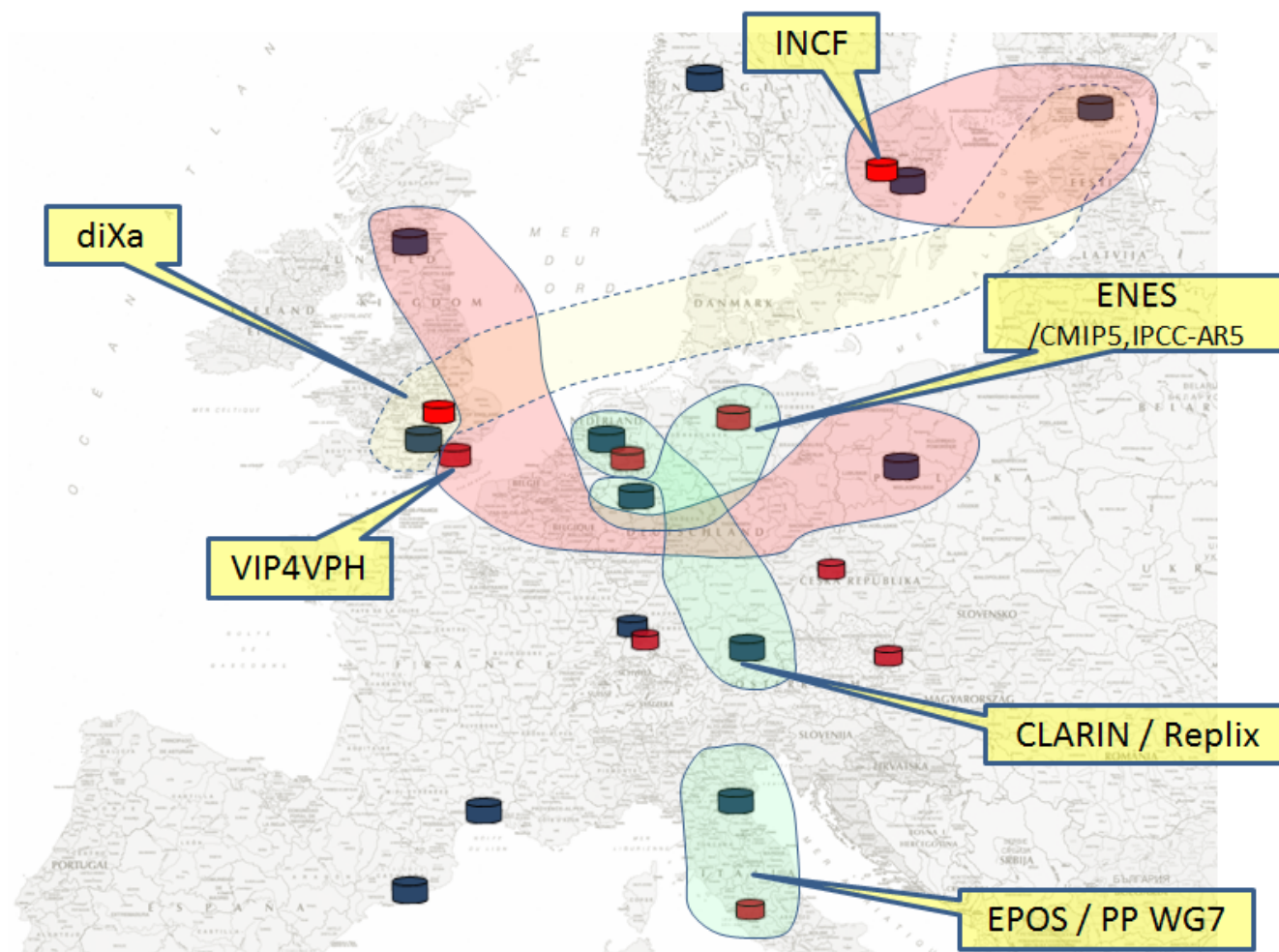
<http://www.clarin.eu/vlo/>



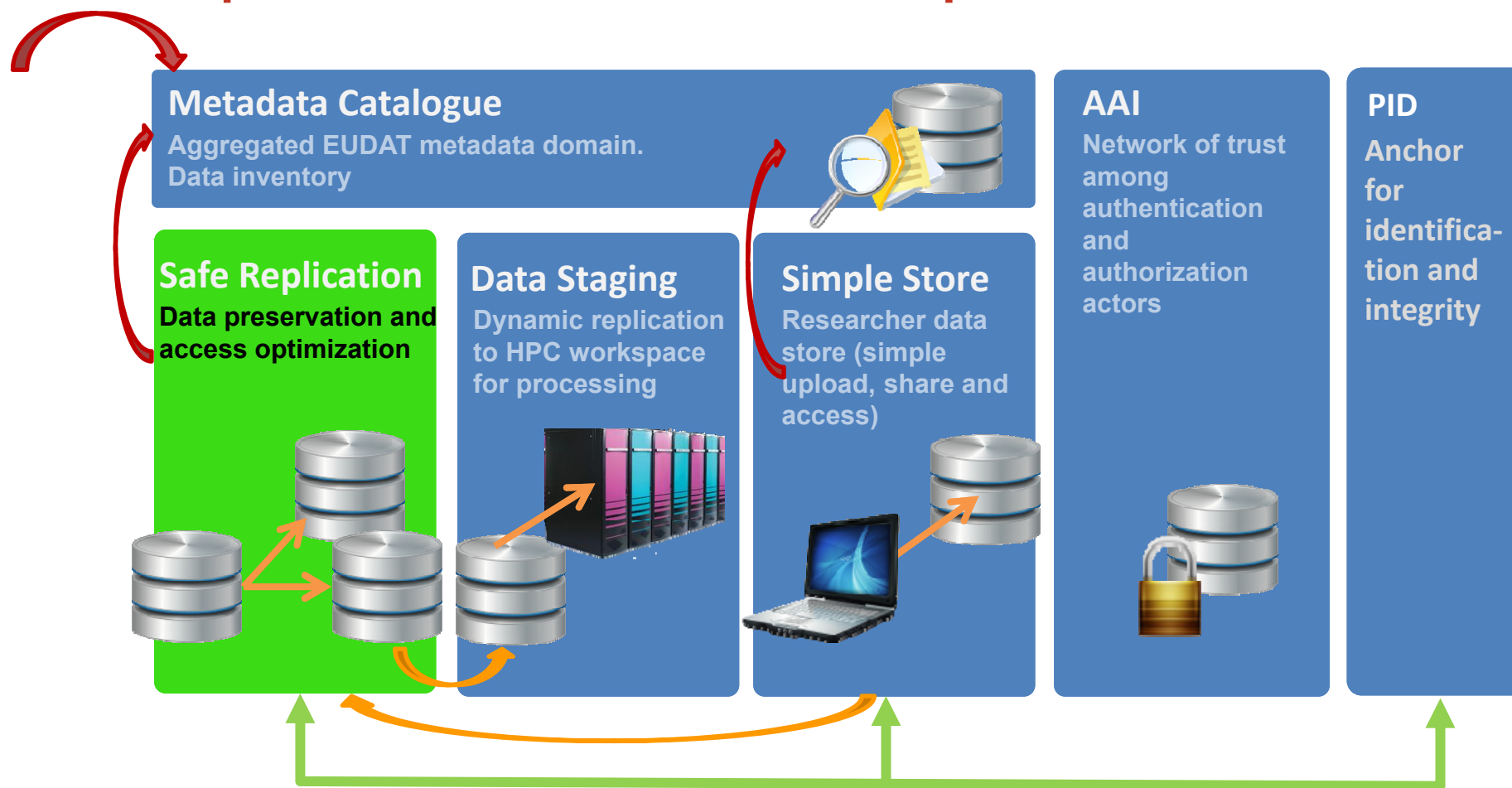
The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable



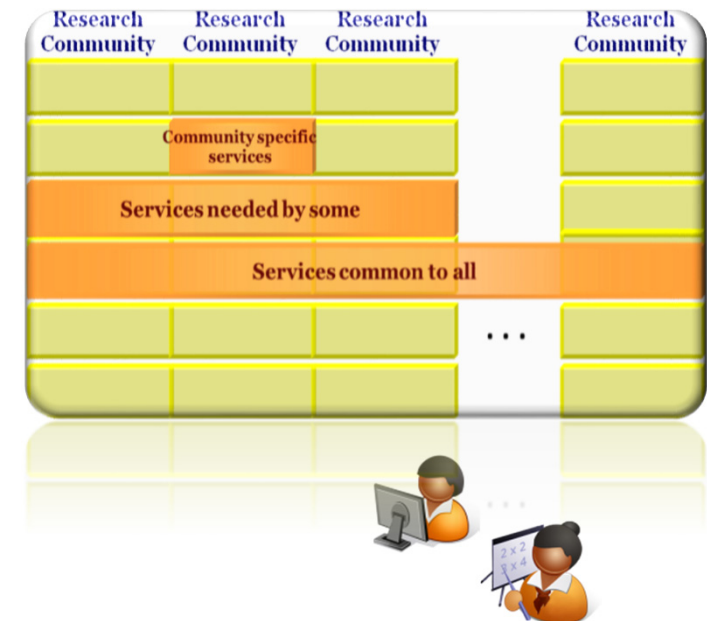
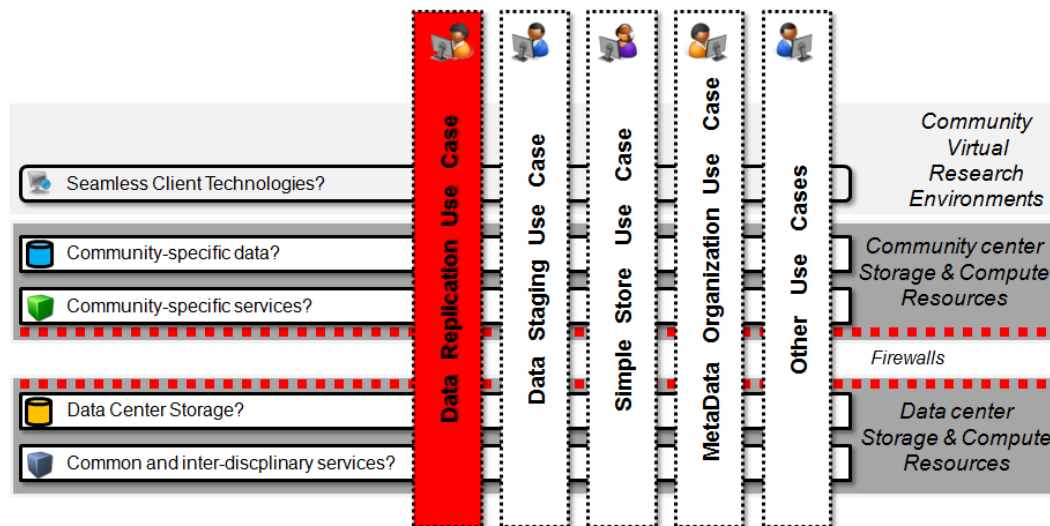
Example: EUDAT Science Relationships



Example: EUDAT Safe Replication Service



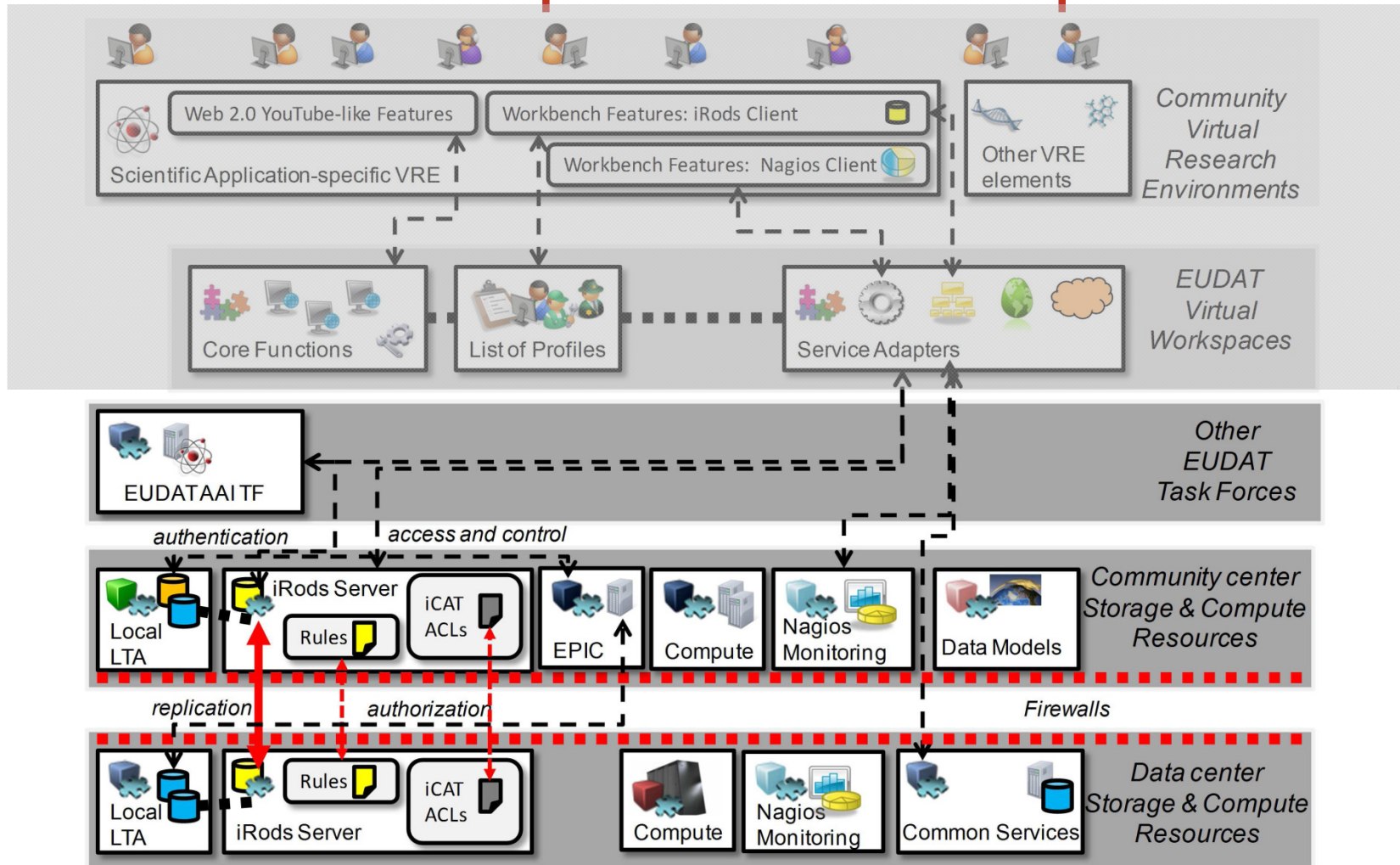
Federated Approach for Use Cases



Create M replications at different data centers for N years,
exclude data centers X to data centers Z from the replication scheme
and make them all accessible by maintaining the given access permissions.

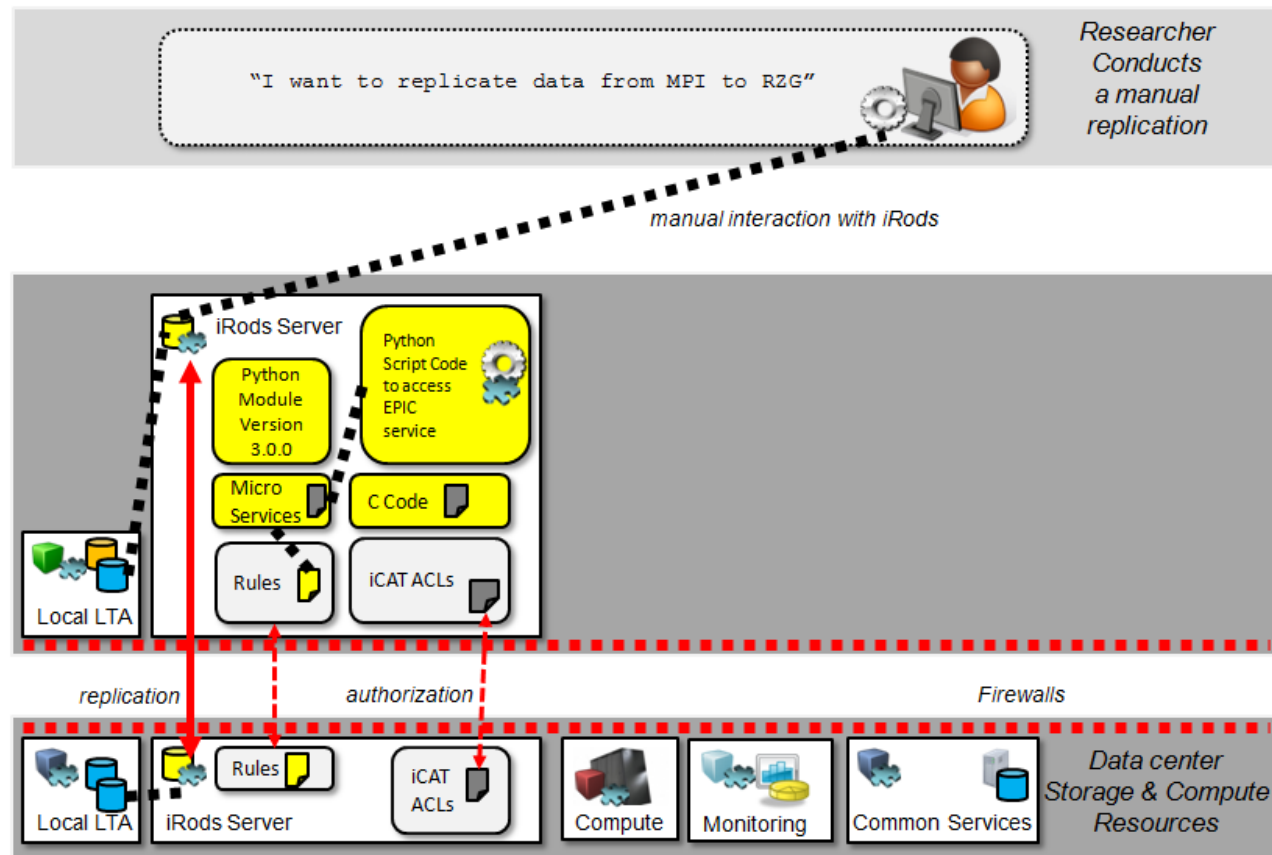


EUDAT: Example of Safe Replication



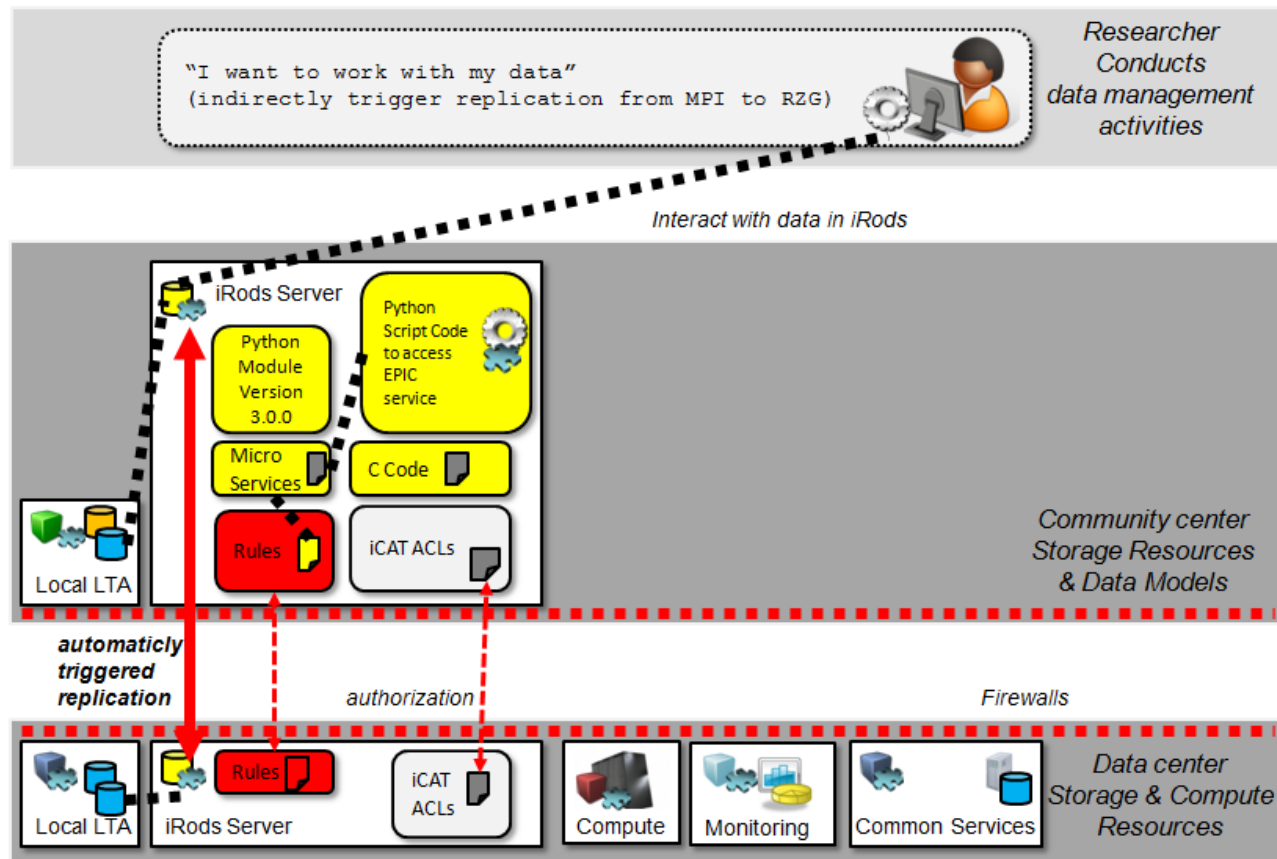
Overview: Manual Upload Replicated File

- Need to understand federations and zones in iRods

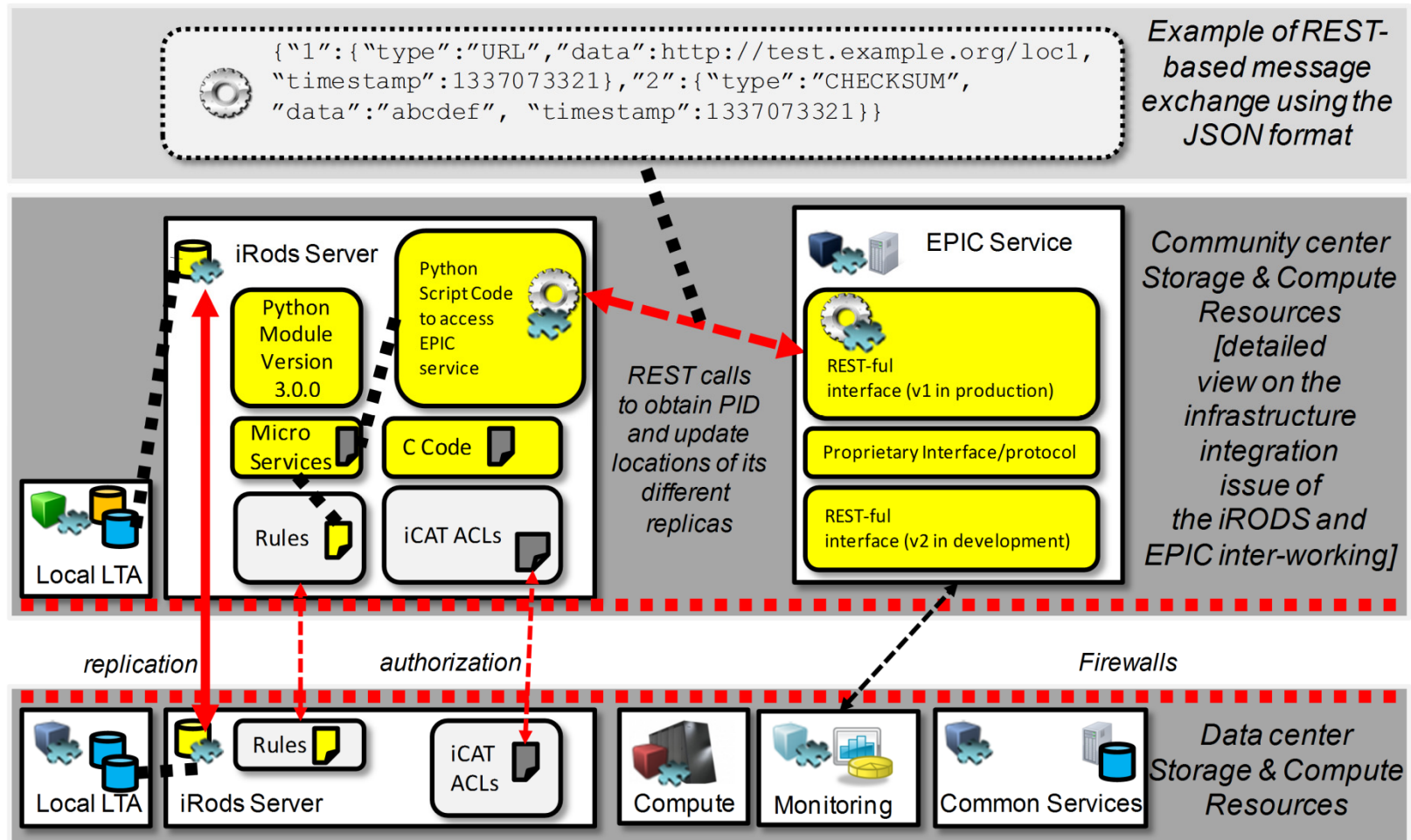


Overview: Rule-based data management

- Need to understand rules & micro-services in iRods



Use of Persistent Identifier (PID) Service



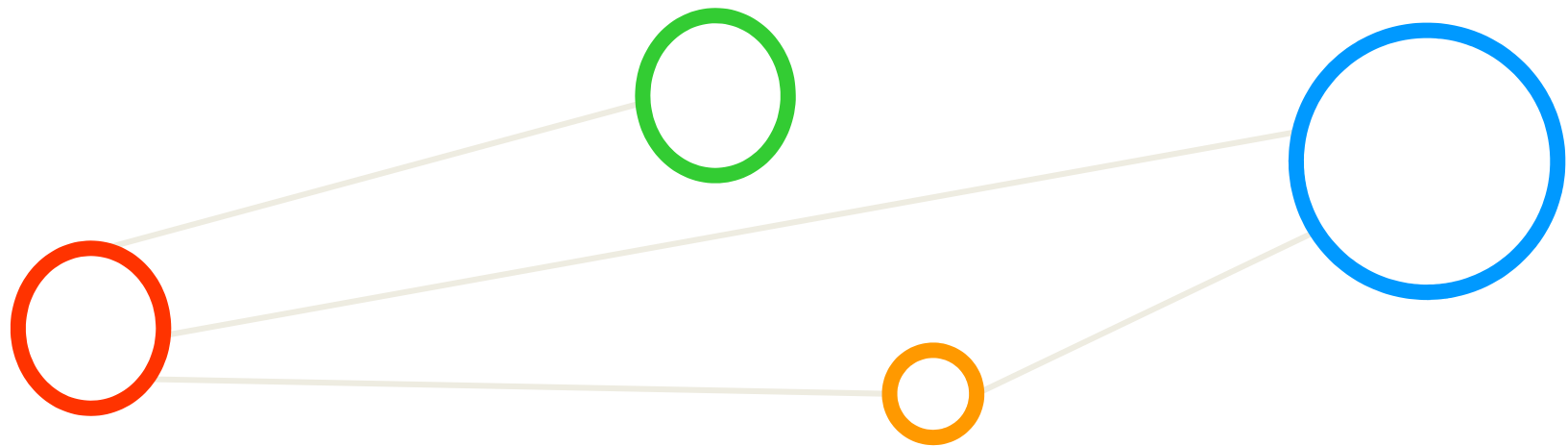


Lessons Learned in this Training Section

- ✓ Understand that long-term relationships matter
- ✓ Knowing the difference between simple backups and safe data replication
- ✓ Understand key aspects of policy-based replication by defining policies on different levels (i.e. rules global/local/infrastructure)
- ✓ Having an idea for what rules can be used in the context of the registered domain of data



Summary & Actions





Summary and Actions

- ❑ Prototype Services are in progress after about 1 year of work
 - ❑ Safe Replication and Data Staging in operation for a few data centers of core communities
 - ❑ Simple Store and MetaData will come soon
 - ❑ **Production means enabling 'the services' together with user communities**
- ❑ Worked hard to get this done and to understand how to interface with communities
- ❑ Needed to chose for some technologies – but take care of technology lock-in
 - ❑ iRODS just as a thin layer for example and not as a system doing all
- ❑ There is a far way between **"we know how it works"** and having a **"real service"**
 - ❑ Communities & researchers are interested in operational services
- ❑ Go ahead and extend the collaborative infrastructure with three levels of thinking
 - ❑ **Working habit of Mindset, Skillset, Toolset**



Thanks for the attention.

Get in contact with us:

<http://www.eudat.eu>

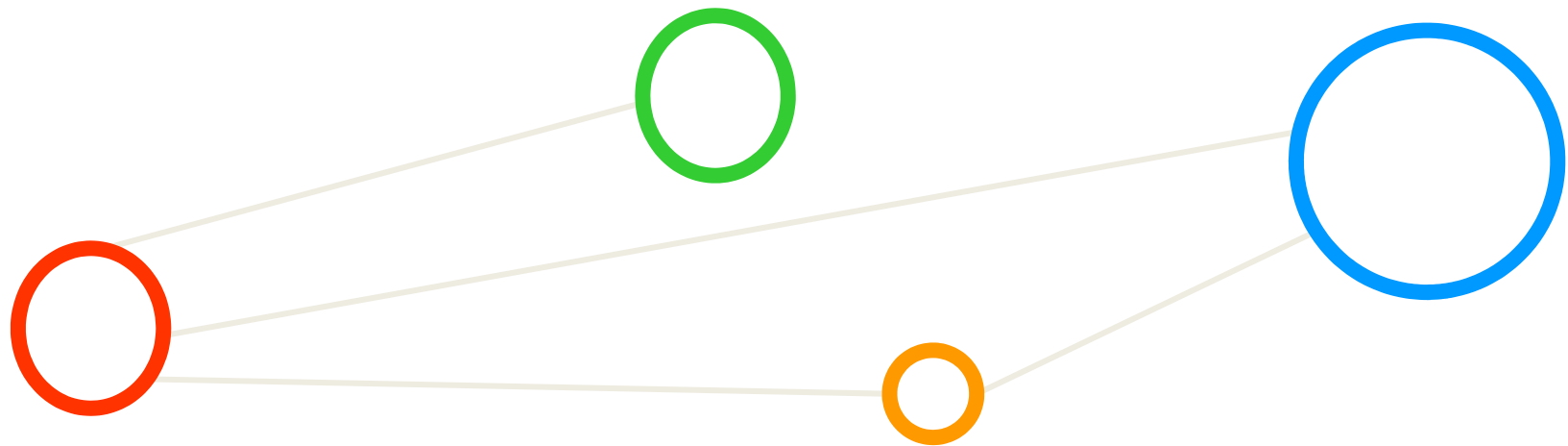


Join the Research Data Alliance

<http://rd-alliance.org/>



References





References

- [1] M. Riedel, P. Wittenburg, J. Reetz, M. van de Sanden, J. Rybicki, B. von St. Vieth, G. Fiameni, G. Mariani, A. Michelini, C. Cacciari, W. Elbers, D. Broeder, R. Verkerk, E. Erastova, M. Lautenschlaeger, R. Budig, H. Thielmann, P. Coveney, S. Zasada, A. Haidar, O. Buechner, C. Manzano, S. Memon, S. Memon, H. Helin, J. Suhonen, D. Lecarpentier, K. Koski and Th. Lippert, ***A Data Infrastructure Reference Model with Applications: Towards Realization of a ScienceTube Vision with a Data Replication Service***, Journal of Internet Services and Applications, Volume 4, Issue 1
- [2] EUDAT Web Page, Available online: <http://www.eudat.eu>
- [3] RDA Web Page, Available online: <http://rd-alliance.org>
- [4] High Level Expert Group on Scientific Data, ***Riding The Wave – How Europe can gain from the rising tide of scientific data***, Submission to the European Commission, October 2010, Available online: <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>
- [5] Knowledge Exchange Partners, ***A Surfboard for Riding the Wave – Towards a four country action programme on research data***, published 2011, updated 2012, Available online: <http://www.knowledge-exchange.info/surfboard>