The Collaborative Data Infrastructure EUDAT

Dr. Morris Riedel

Federated Systems and Data

Juelich Supercomputing Centre (JSC)
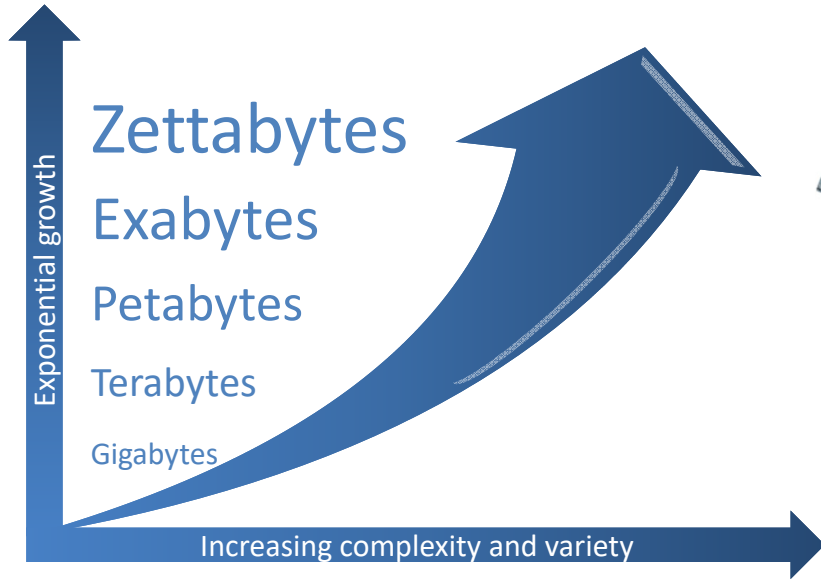
Adjunct Associate Professor
School of Engineering and Natural Sciences
University of Iceland

Data trends

Exponential growth → **Increasing complexity and variety**

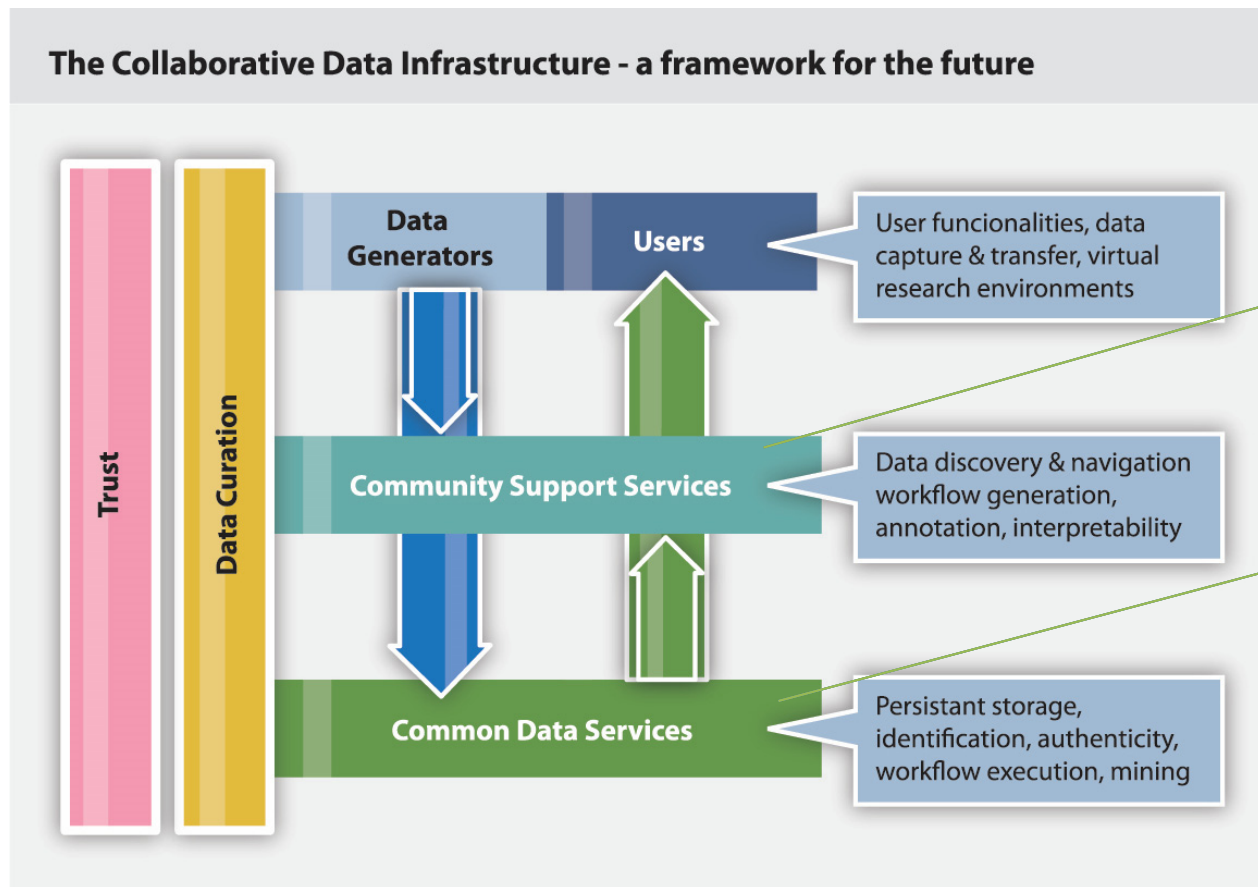- Zettabytes
- Exabytes
- Petabytes
- Terabytes
- Gigabytes

- Where to store it?

- How to find it?

- How to make the most of it?

- How to ensure interoperability?

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, etc.
5 Core Infrastructures more second round infrastructures

=> 12 EUDAT data centers

EUDAT

The EUDAT Case

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

EUDAT – real CDI Landscape

Five research communities on Board

- **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**

EUDAT
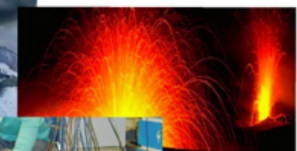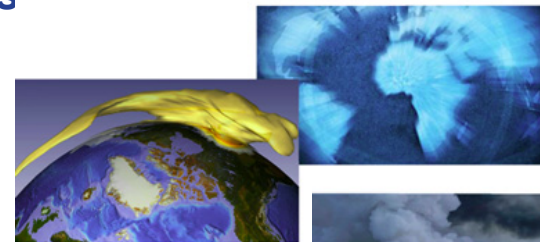
Working habits with Communities
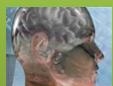
MINDSET

> *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?*
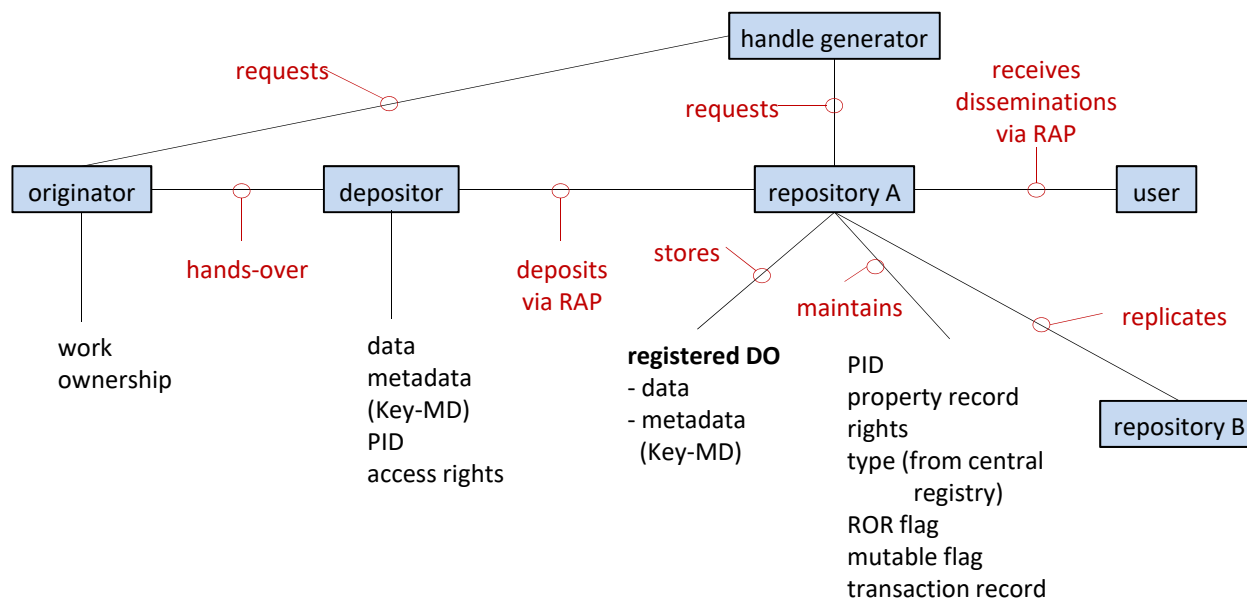
EUDAT

EUDAT Centres involved in Operations

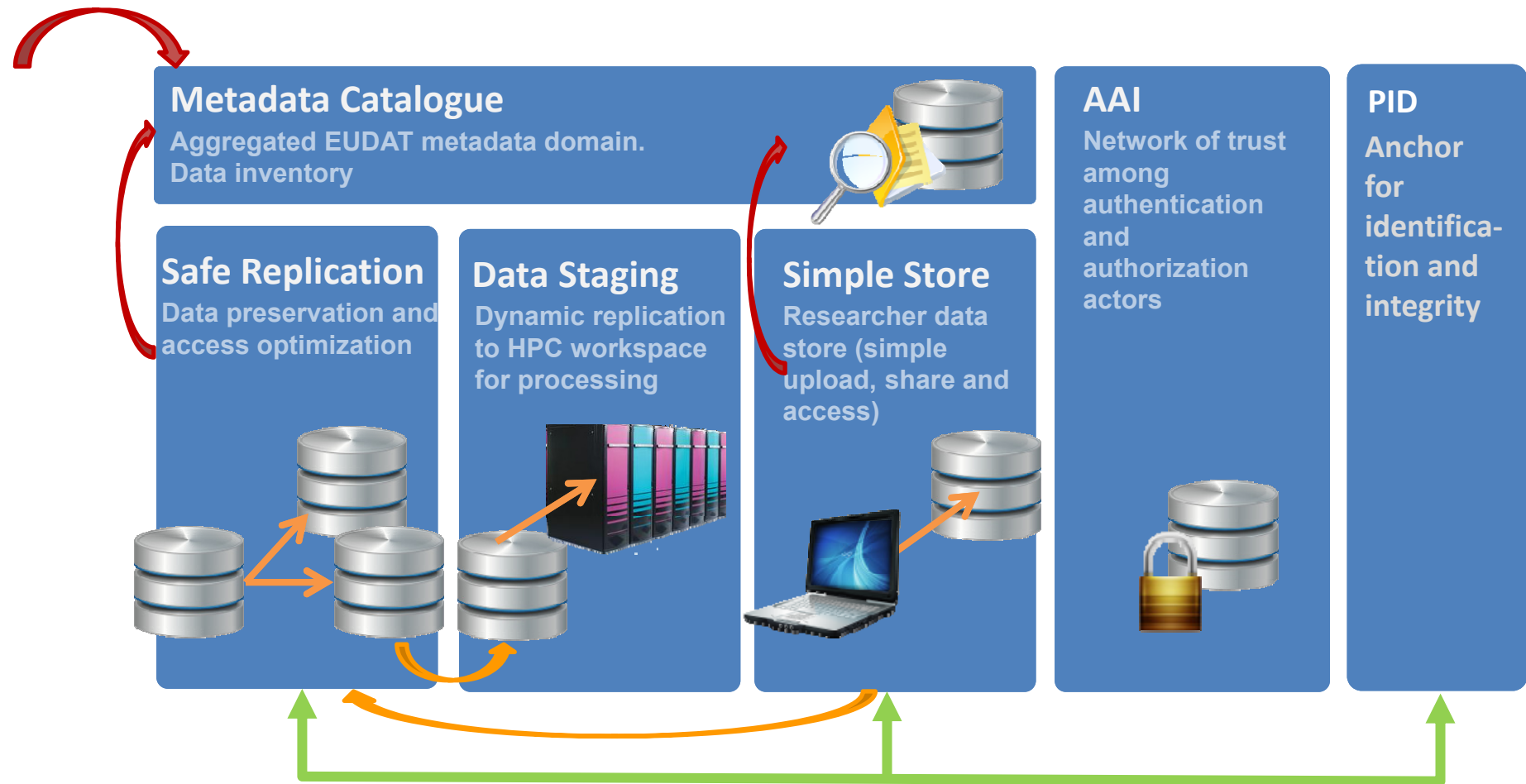# Community Centres

Analysis of Existing Data Management

❑ **community interactions based on abstract model (Kahn & Wilensky, 2006)**
   o **Data + Metadata + Handle (PID)**
❑ **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
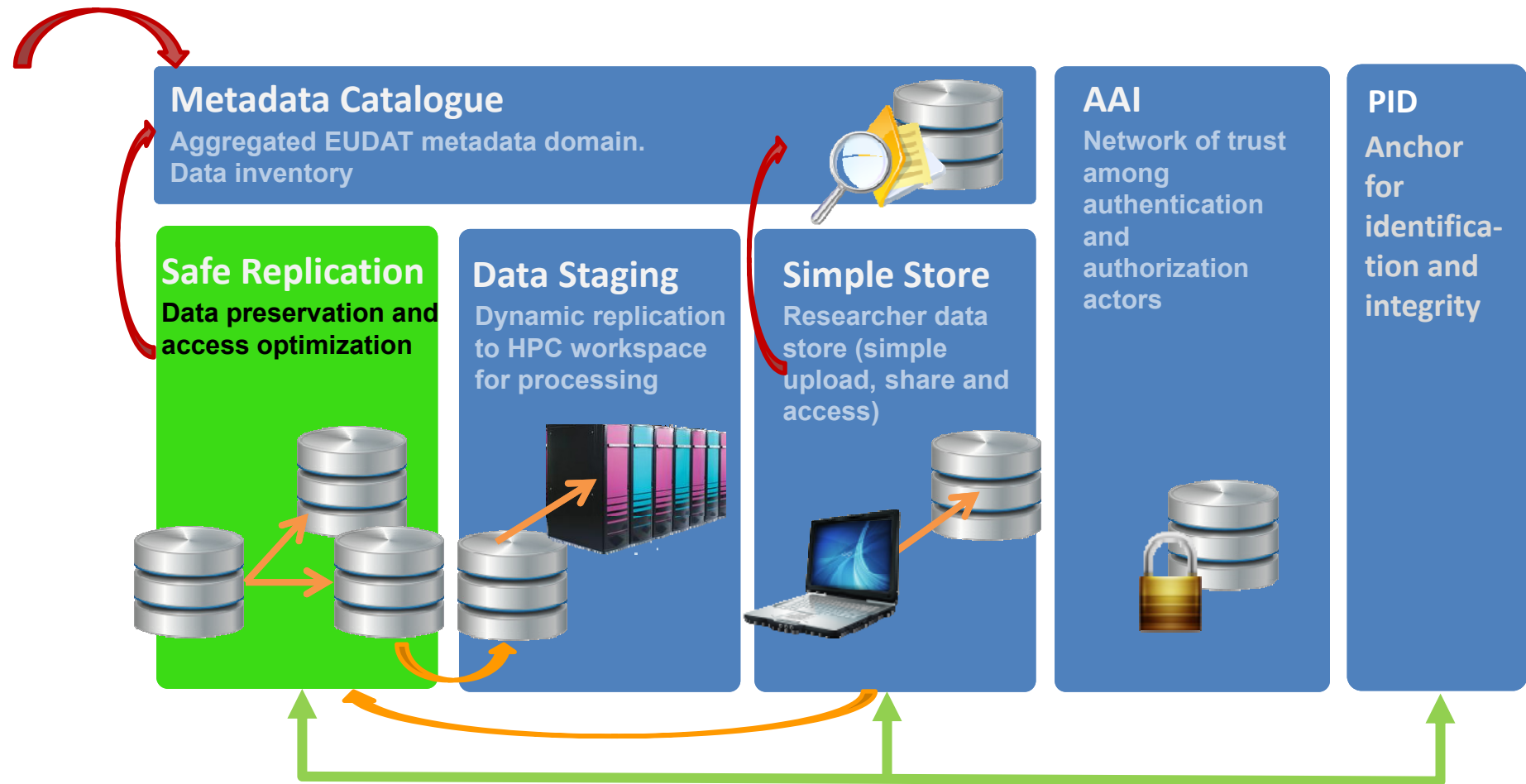**transaction record** = all disseminations of a DO

handle generator

requests

requests

receives
disseminations
via RAP

originator — depositor — repository A — user

hands-over

deposits
via RAP

stores

maintains

replicates

work
ownership

data
metadata
(Key-MD)
PID
access rights

registered DO
- data
- metadata
  (Key-MD)

PID
property record
rights
type (from central
   registry)
ROR flag
mutable flag
transaction record

repository B

First Services in Preparation

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 identification and integrity

Data PID Metadata

Services working on

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

Services in Discussion

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
Replica Access	X		X	X	X	X			x		x			
Semantic Annotation	o	X			x									
Web-service platform	X	o		X	o									
Real Time Data		x			X									

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

13

Concepts of the Safe Replication Service

Metadata Catalogue
Aggregated EUDAT metadata domain. Data inventory

AAI
Network of trust among authentication and authorization actors

PID
Anchor for identification and integrity

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)

Data   PID   Metadata

EUDAT

Safe Data Replication Vision

- Idea: Safe replication between 1 scientific community center and N data centers
  - Replication within a ‚registered domain of data' (i.e. PID assigment)

- 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

Safe Replication
Data preservation and access optimization

EUDAT
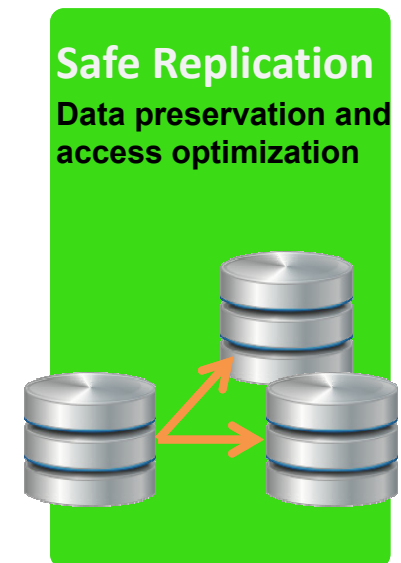
Safe Replication Service in a Nutshell

metadata   data

PID Domain

trusted data center A

trusted data center B

Federation Domain

replication

master collection

Safe Replication — Data curation and access optimization
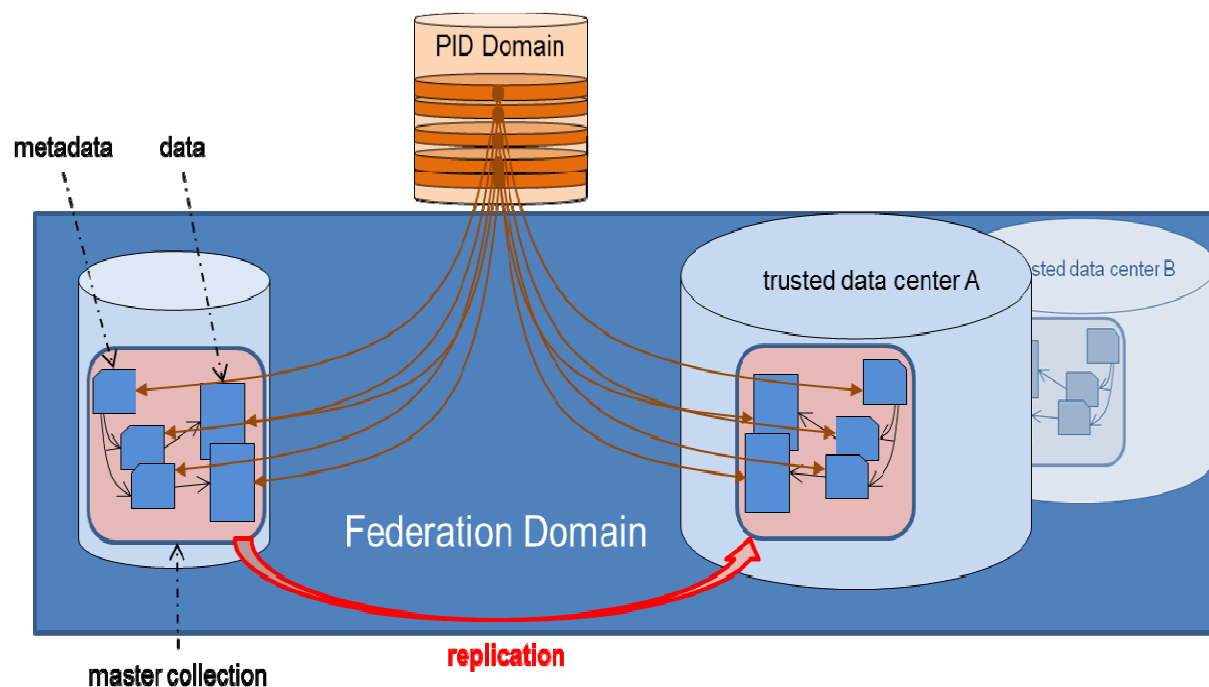
Better accessibility of scientific data

Make data referencable

High degrees of reliability and trust

More optimal data curation

16

Federated Approach for Use Cases

	Data Replication Use Case	Data Staging Use Case	Simple Store Use Case	MetaData Organization Use Case	Other Use Cases	
Seamless Client Technologies?						Community Virtual Research Environments
Community-specific data?						Community center Storage & Compute Resources
Community-specific services?						
						Firewalls
Data Center Storage?						Data center Storage & Compute Resources
Common and inter-discplinary services?						

Research Community — Research Community — Research Community — Research Community

Community specific services

Services needed by some

Services common to all

**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

17

Forming Strong EUDAT Collaborations

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

Scientific Community	Community Centers	Data Centers
CLARIN	MPI-PL	RZG, SARA
ENES	DKRZ	JSC, CSC
EPOS	INGV	CINECA, SARA

EUDAT

Use Cases Derived Reference Architecture
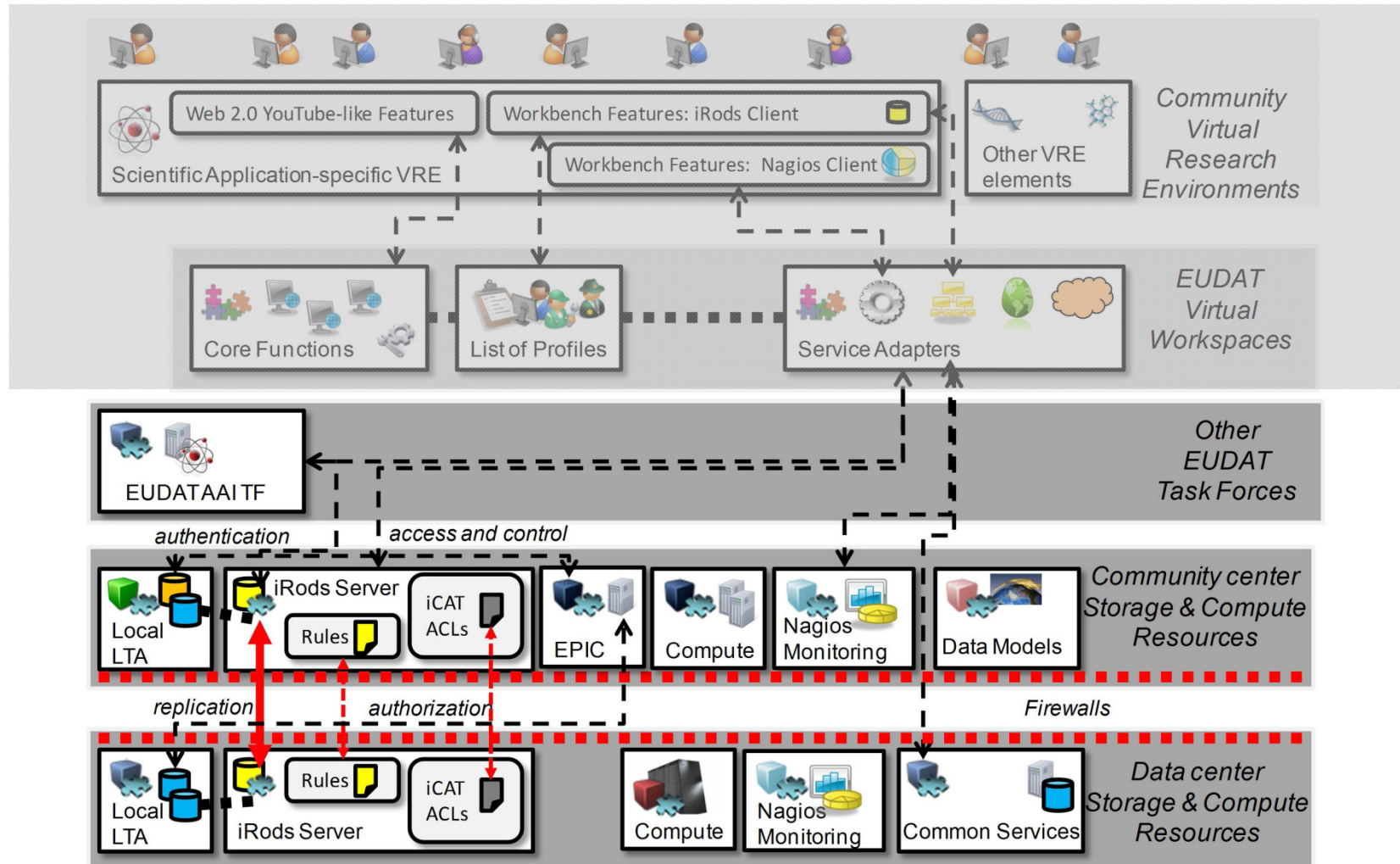
Overview: Manual Upload Replicated File

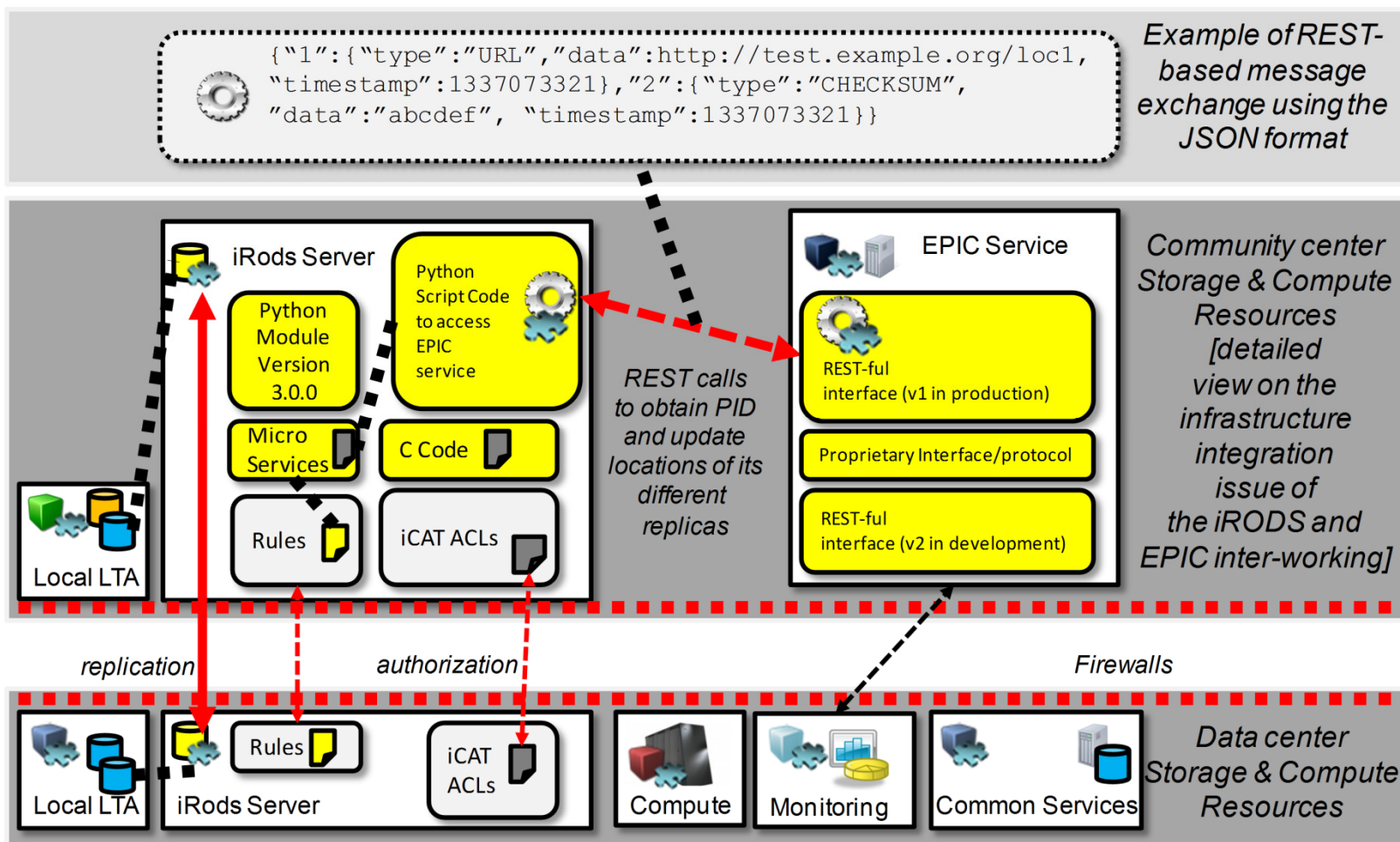- Need to understand federations and zones in iRods
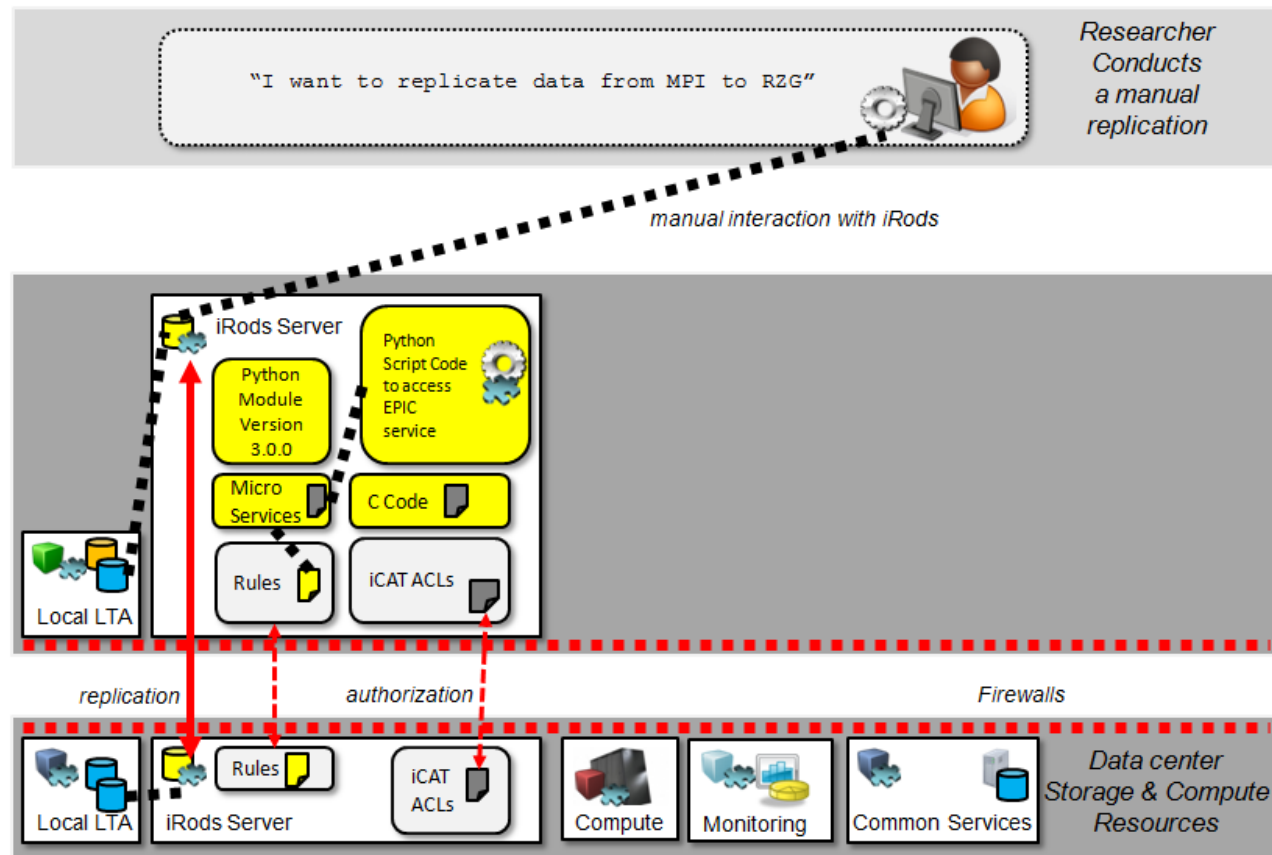
Overview: Rule-based data management

- Need to understand rules & micro-services in iRods
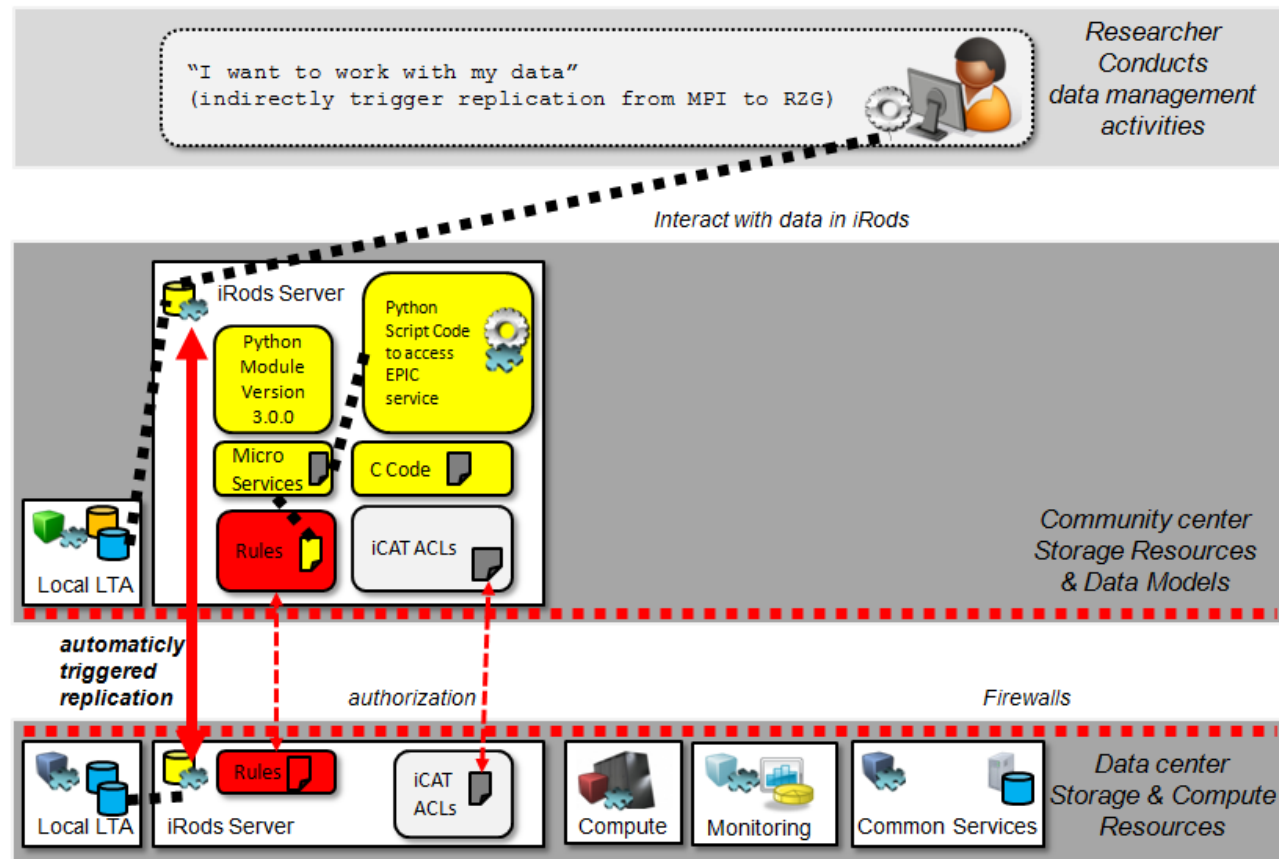
Use of Persistent Identifier (PID) Service

EUDAT: where are we?

- ❑ **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 infrastructure with three levels of thinking**
    - ❑ **Working habit of Mindset, Skillset, Toolset**

Thanks for the attention.

http://www.eudat.eu

Join the Research Data Alliance Meeting

http://forum.rd-alliance.org

**EUDAT**

Data Landscape Analysis: EPOS

- **EPOS (Seismologists, Vulcanologists, etc.)**

  - lots of distributed data sensors producing continuous package streams
  - due to various reasons data streams include gaps to be filled over time
  - data windows of interest (WoI) are defined „vulcano eruption X"
  - aggregations of such data are of relevance (large scale statistics etc)
  - work currently on a description of metadata schema for WoIs
  - work on a scheme of how to refer to packages and offsets (Handles, fragments)
  - one center is now implementing reference architecture
  - need to synchronize with US and other colleagues
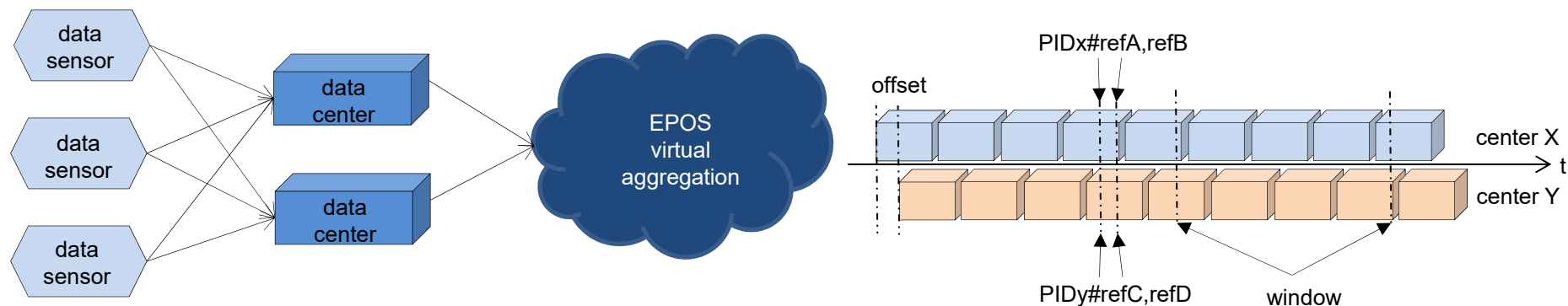
SAFE Data Replication

- safe replication between 1 community center and N data centers
- flexibility, scalability and management require policy rule based approach
- 3 islands (community + data center) in parallel & close interaction
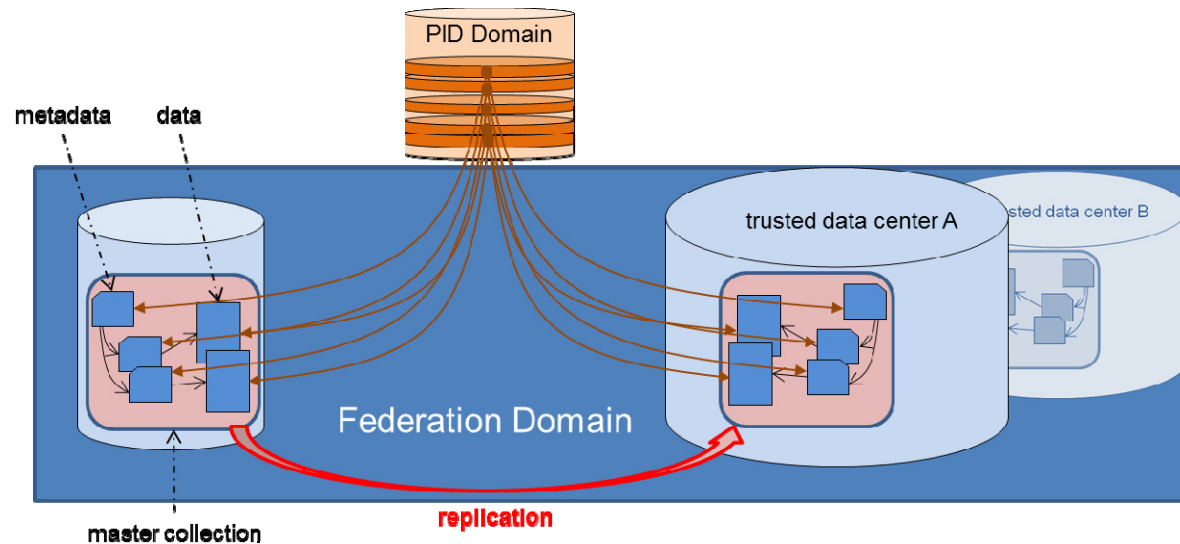
- basic technlogies: AAI, iRODS, Handles, community MD & OAI-PMH, center registry
- in June merging of 3 islands to one flexible replication domain
- REPLIX experience is basis

Staging to HPC Pipes

- intention is to make use of HPC machines for computations on stored data
- different configurations possible:
  - computations on a single HPC node where data already is
  - computations on multiple nodes - use of PRACE fast distributed file system

Expert Task Force
built, to be ready
in summer
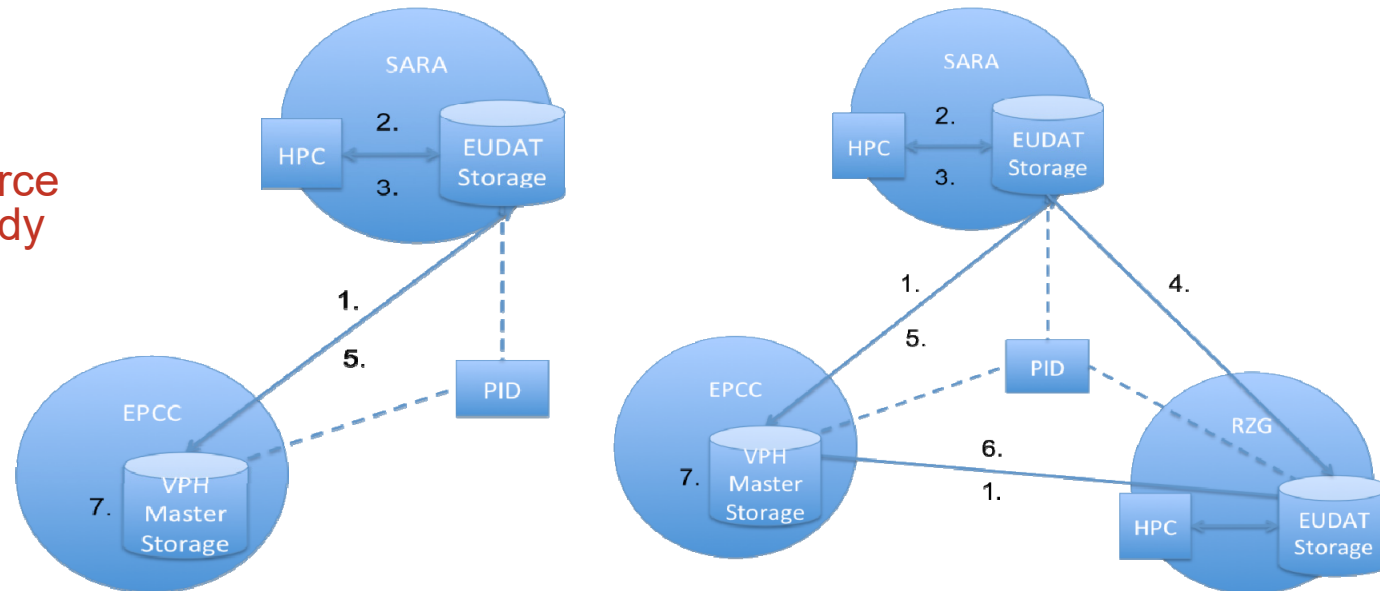
- principles:
  - user issues a compute command
  - script pushes data into the HPC workspace, results go into workspace
  - input data is discarded after job end, user needs to store the results

EUDAT

Aggregated Metadata Domain

- not yet fully specified
- question: for what ???
    - probably loss of specific information - thus interdisciplinary research
    - should show what is stored in the EUDAT data centers
    - one stop shop for virtual collection building
    - making PR for collections (ANDS model)
- general index with some faceted browsing machine probably not sufficient
    - element semantics probably too different
- therefore currently analysis of semantics and simple mapping schemes

- enabling technologies:
    - OAI-PMH, refs via PIDs, SOLR/Lucene for indexing/browsing
    - when and how semantic expansion
    - do we need higher performance technology?

- decision about criteria in February
- technology watch in March

Researchers Simple Store

- not yet fully specified
- question: for what ???
  - researchers need/want Simple Store for all their „secondary" data
  - trust is an important issue - owner/copyright must be (with) the researcher
  - data should be part of the EUDAT data domain (thus Metadata, PIDs)
  - ingest via community control to prevent misuse
- Simple Store must have simple access component (like YouTube) and perhaps easy ‚promotion' of data into community center collections

- enabling technologies:
  - AAI, PIDs, MD Indexing

- decision about criteria in February
- technology watch in April (what about Mercury etc.)

EUDAT

need to agree on layers: access

Typical Access Workflow

Enabling Technologies

Scientists, Data Curators, End Users, Applications

Discovery

Access
(ref. resolution, protocols, AAI)

Interpretation

Reuse

Datasets
Accessed via Repositories

all in collaboration with CNRI
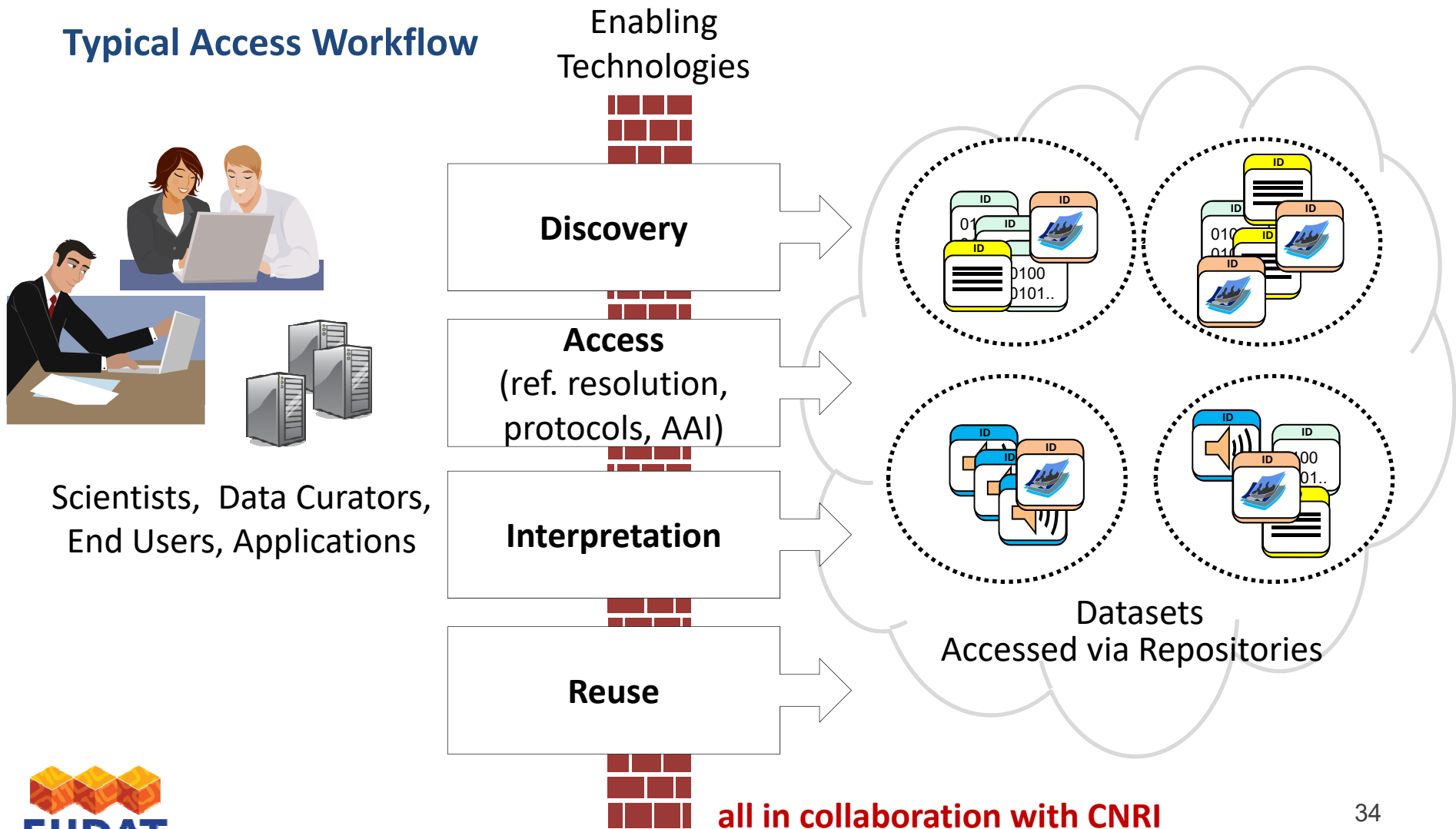
need to agree on layers: managment

Typical Management Workflow

Data Managers
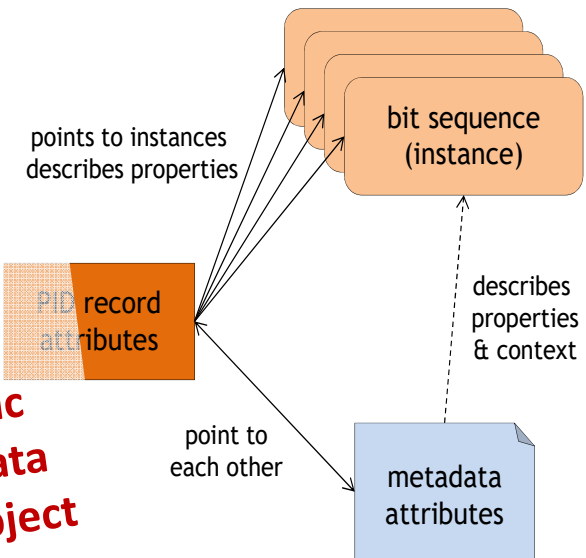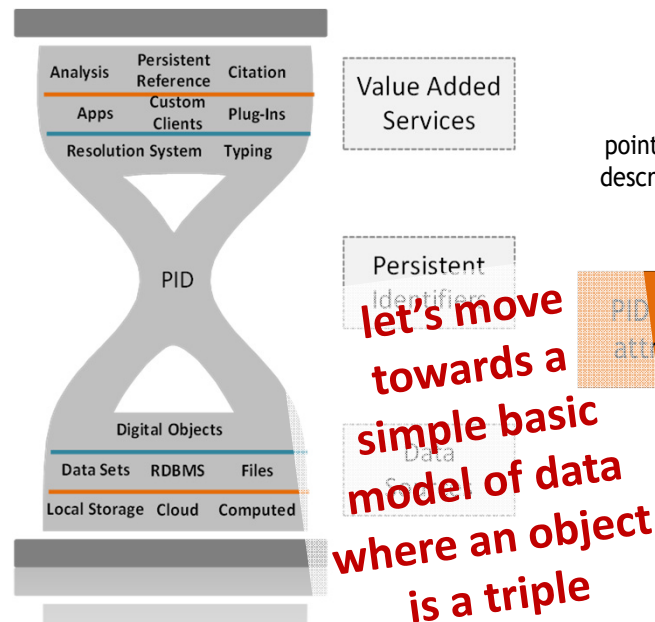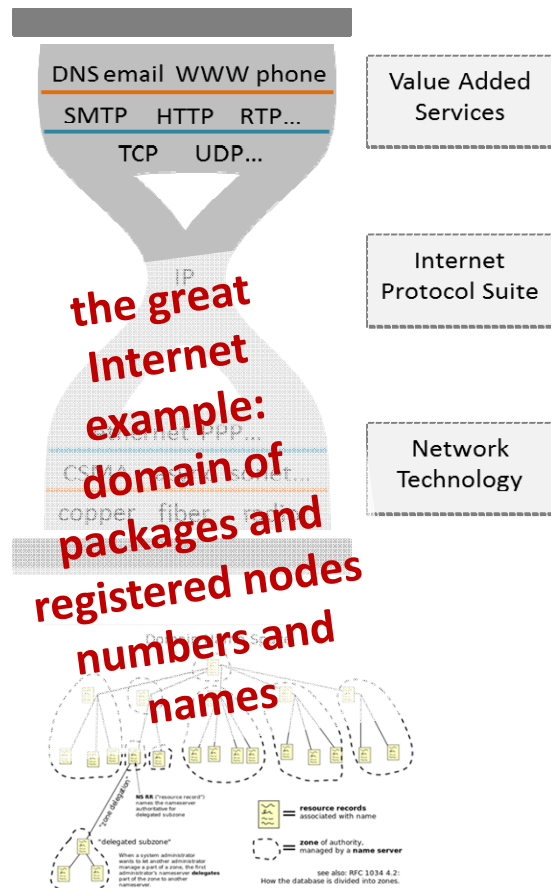Data Scientists

Enabling Technologies

Collections + Properties

Access
(ref. resolution, protocols, AAI)

formalized policies workflow engine

Assessment

Datasets
Accessed via Repositories

all in collaboration with CNRI
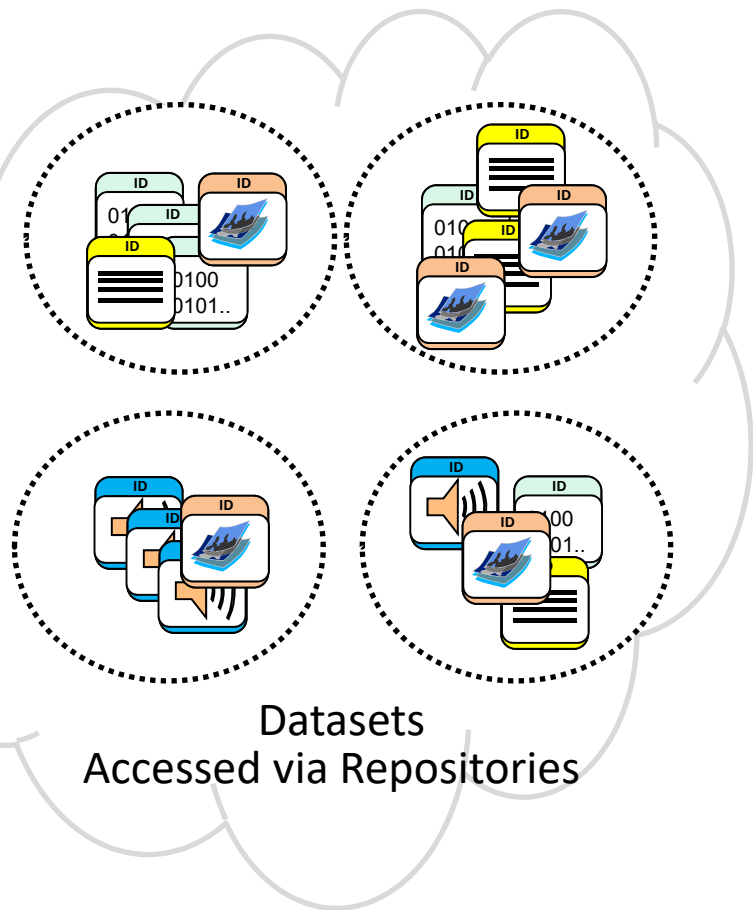
EUDAT

need to agree on basic models & terms

the great Internet example: domain of packages and registered nodes numbers and names

**let's move towards a simple basic model of data where an object is a triple**

❑ **let's come to a common object model with PID as anchors – like IP numbers in networks**

❑ **PID and MD store properties of objects and collections, policy rules manipulate properties**

❑ **EUDAT is a domain domain of registered data objects**
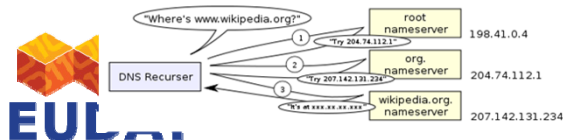
all in collaboration with CNRI

Reality

taken from R. Koskela (DataOne)
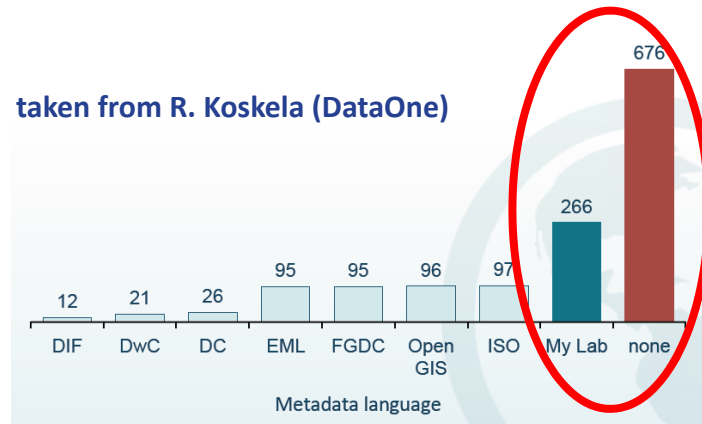
- ❑ in the labs there is no agreed metadata
    - ❑ if so no registered schemas and category sets (semantics)

- ❑ externally registered PIDs are not used
- ❑ many encapsulate and do not have an idea what an object is that can be reproduced

❑ in EUDAT interviews/analysis with/of about 15 communities, in Radieschen interviews with about 12 departments

❑ thus first results of systematic analysis of data organizations – some surprises

❑ all communities are busy with their data organizations in some way - Panta Rhei
- ❑ they are at differerent stages – organization and broad deployment
- ❑ departments are often lost in data management and lack offers
- ❑ don't believe people who claim to have solved the issue

❑ greatest success in EUDAT/DASISH etc: several communities seem to speak one language

What is RDA working on

❏ **Data Foundation and Terminology (implies some agreed conceptualization)**

❏ **PID Information Type Harmonization**

❏ **Data Type Registry**

❏ **Practical Policy**

❏ **Metadata Normalization**

❏ **Pub/Data Citation/Linking**

❏ **Legal Interoperability**

❏ **Repository Audit and Certification**

❏ **The Engagement Group**

❏ **Marine Data Harmonization**

❏ **Defining Urban Data Exchange for Science**

almost all group results would have an impact on EUDAT and simplify a lot

EUDAT

EUDAT – RDA

❑ **RDA will have a great impact on cross-disciplinary enterprises as EUDAT**

    ❑ **it is bottom-up and driven by "data practitioners"**

    ❑ **it's focus is on removing concrete barriers on the way of sharing and interoperability – so it's not another policy group**

❑ **I hope that RDA will also have implications on data organizations of communities**

    ❑ **as usual – some argue that they solved the problems**

❑ **of course there are other important organizations we need to look at:**

    ❑ **IETF**   **focus on networking**

    ❑ **W3C**   **focus on the Web and its mechanisms**

    ❑ **CODATA**   **focus on policies in area of data**

    ❑ **World Data Systems**   **focus on proper data centers**

    ❑ **G8+O5 Data Group**   **also focus on policies in area of data**

❑ **come to the RDA Launch and Plenary: 18-20. March 2013 Gothenburg, Sweden**

EUDAT

Data Landscape Analysis: ENES

- **ENES (Climate Modeling Research)**

  - about 20 centers in Europe -
  - have CIM data model - but this is still in a prototype state, not deployed broadly
  - but CDI as operating at German Climate Center is taken as basis
  - CIM has kind of „canonical" design using DOIs and EPIC Handles
 - Metadata based on ISO 11179 etc.; OAI-PMH in place
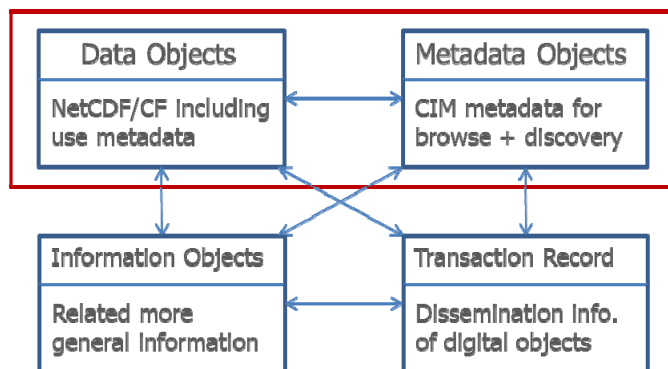
Identification of distinct data objects and P2P infrastructure
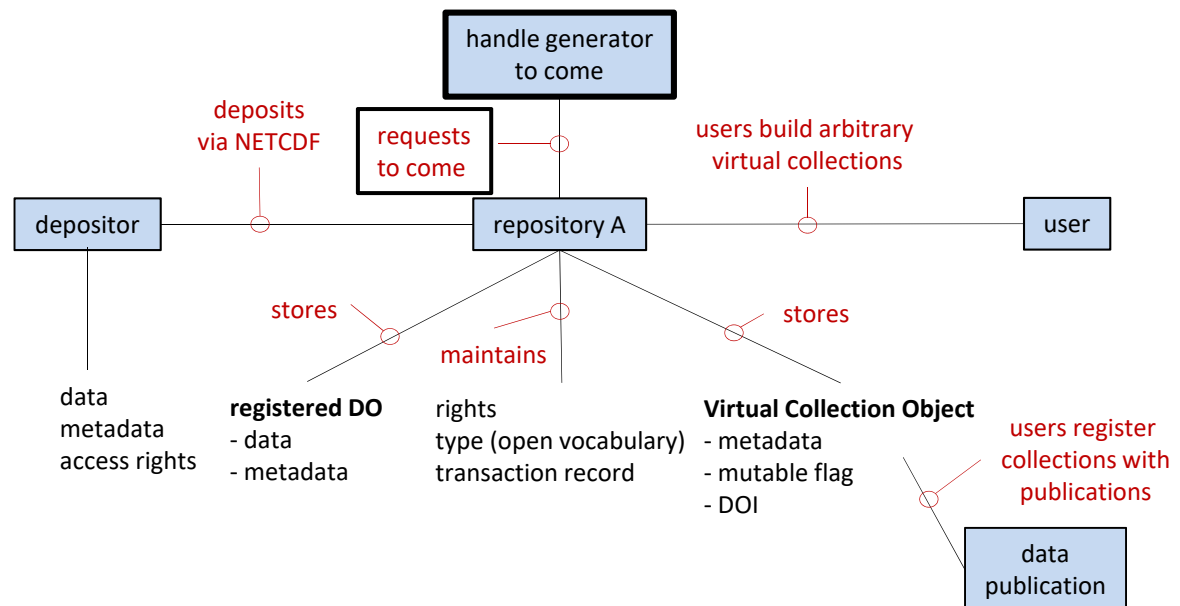
Data Landscape Analysis: VPH

- **VPH (Virtual Physiology of Humans)**

  - currently pilot project with about 5 hospitals in different countries
  - one centralized data center - in next phase distributed system
  - focus was on metadata aggregation
  - IMENSE stores all textual data and Metadata in a DBMS and gives access
  - data aggregation is planned together with a large data center in EUDAT
  - metadata not yet standardized & formalized (DICOM, JPEG headers, etc.)
  - nothing done with PIDs, AAI and OAI-PMH yet

different types of
data streams
(cell membrane to fMRI,
treatment data etc)
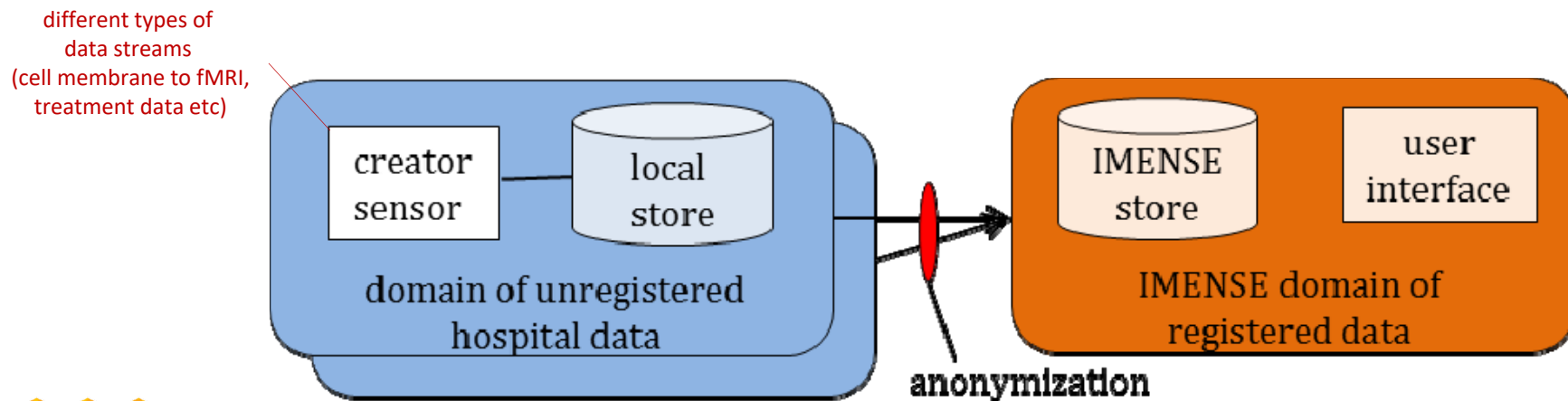
Data Landscape Analysis: LifeWatch

- **Biodiversity (much based on GBIF)**

    - yet no chance of qualified interaction due to time restrictions
    - different contributors and actors
    - very heterogeneous domain

    - first requirements & implementations without LifeWatch
    - need to be flexible enough anyhow

REPLIX

- safe replication between CLARIN center and RZG data center
- purpose: preservation, computation (AV Recognition) and access optimization
- total amount: 80 Terabytes
- requires policy rule based approach due to quality assessment (Data Seal)
- iRODS, Handles, CMDI Metadata
- deployment of Archive/Access software stack as well
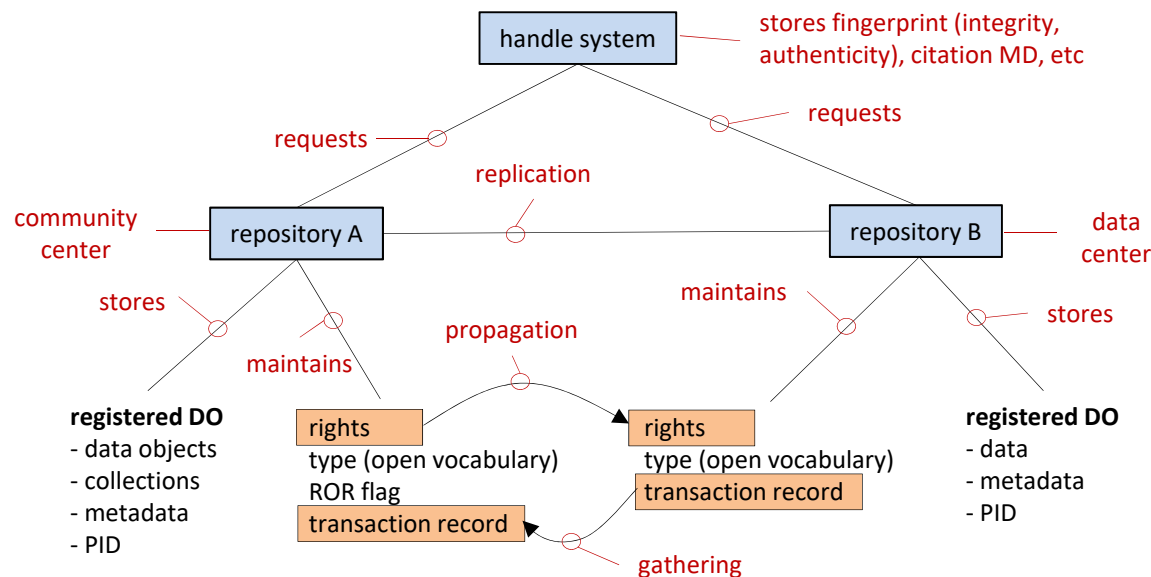
replication at logical collection level basis for demos at ASIST and ICRI conferences both in March
(MPI - RENCI)