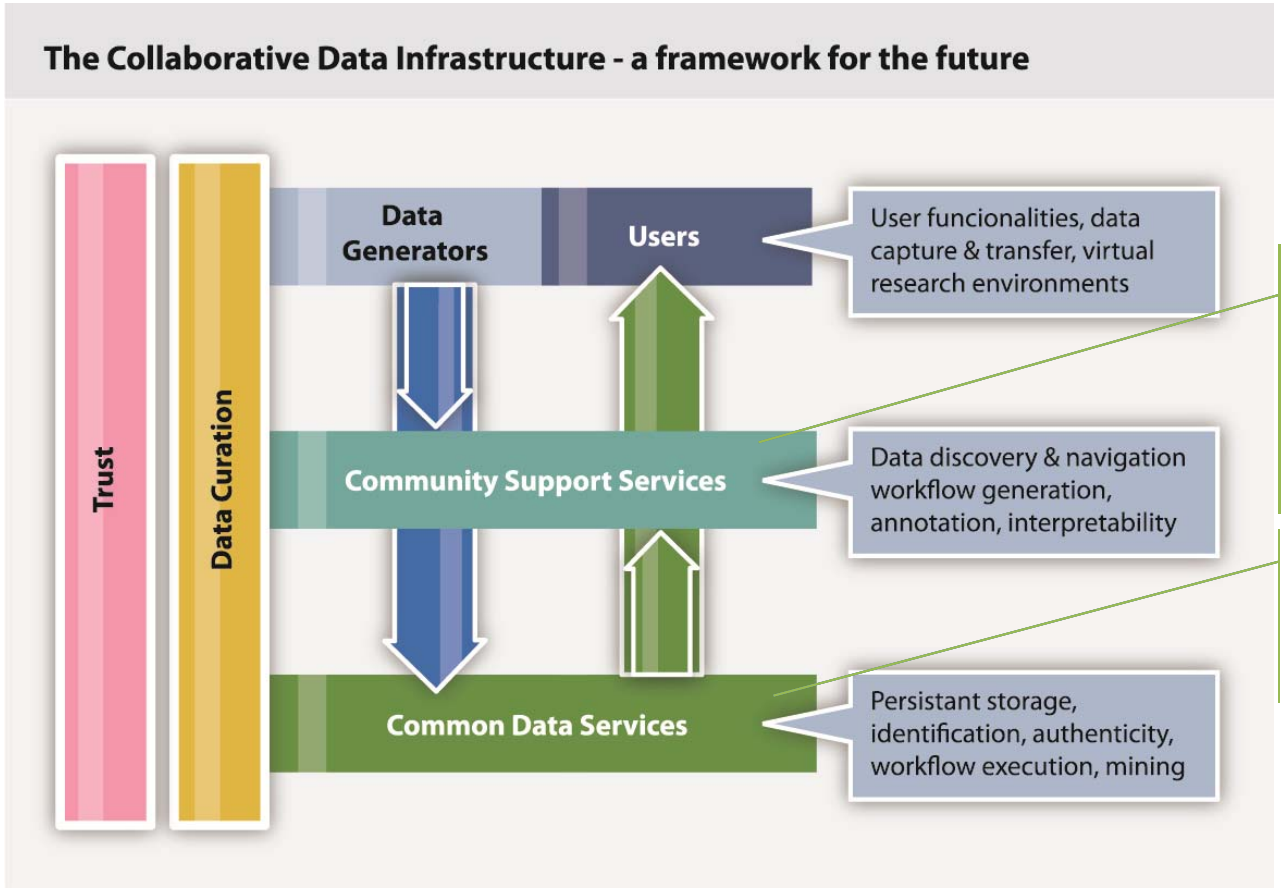
EUDAT: A new Cross-Disciplinary Data Infrastructure for Science

Peter Wittenburg
EUDAT Scientific Coordinator
The Language Archive, Max Planck Institute, Netherlands
D. Lecarpentier, W. Elbers, Alberto Michelini, R. Kanso, P. Coveney

EUDAT

January, 2012

EUDAT's mission: common services in CDI



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

=> 12 EUDAT data centers

EUDAT

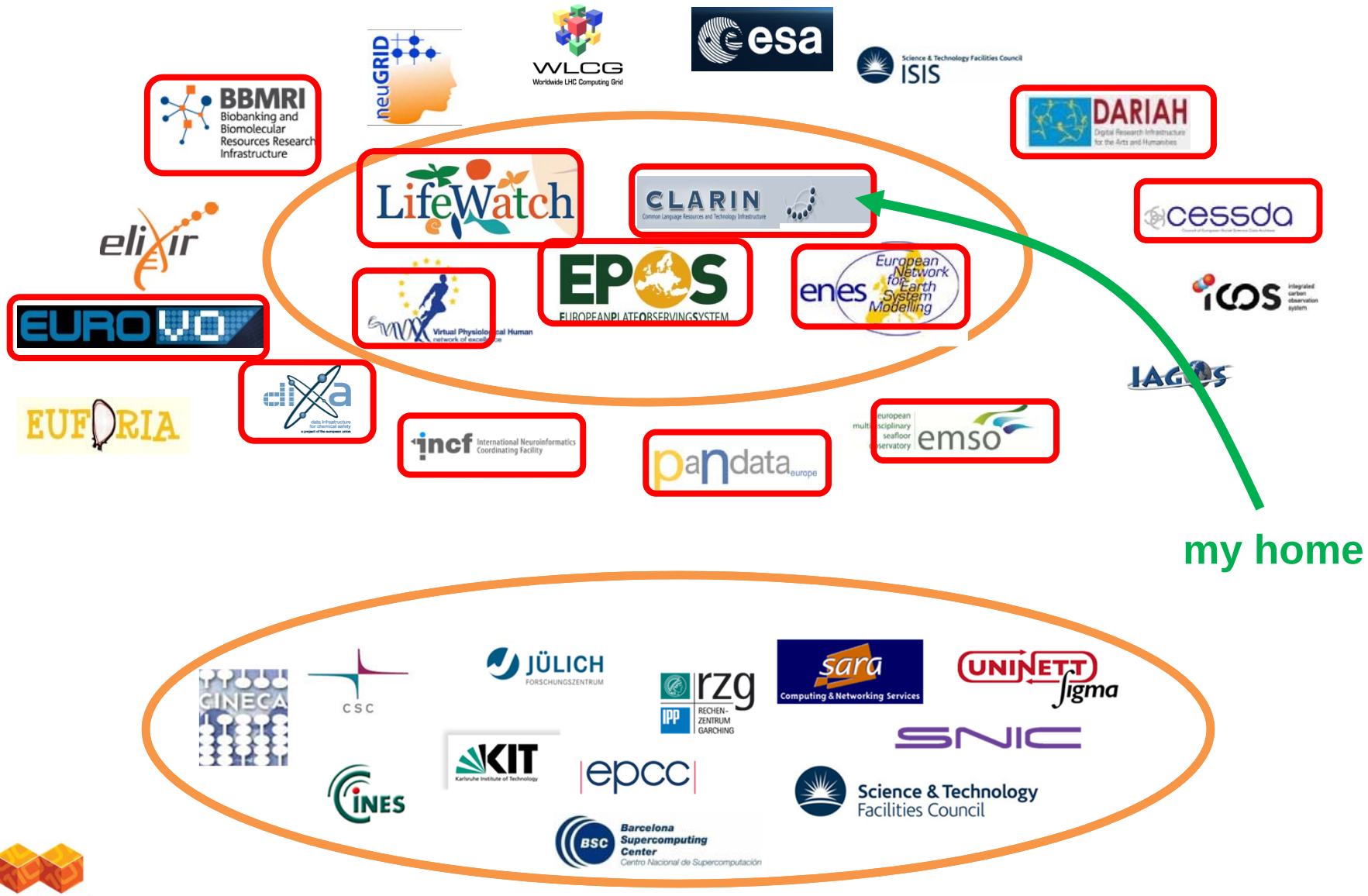
Understanding the Data Landscape in CDI

Why?

- ❑ **understanding how communities/departments organize their data**
- ❑ **building common services needs to build on the existing solutions – to a large extent**
- ❑ **in CDI you need to speak the same language**

EUDAT

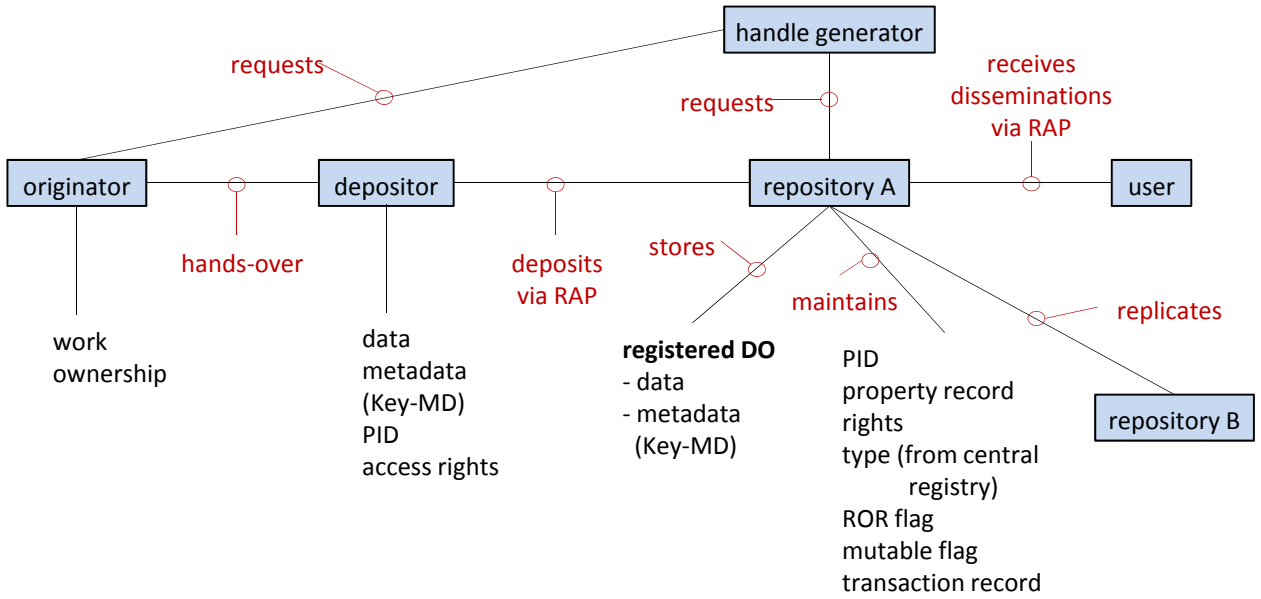
EUDAT – real CDI Landscape



EUDAT

Data Organization and Terminology

- community interactions based on abstract model (Kahn &Wilensky)
- used in many meetings and interactions - accepted quickly as reference model
- helped even in improving community organization plans
- helps contributing to RDA process



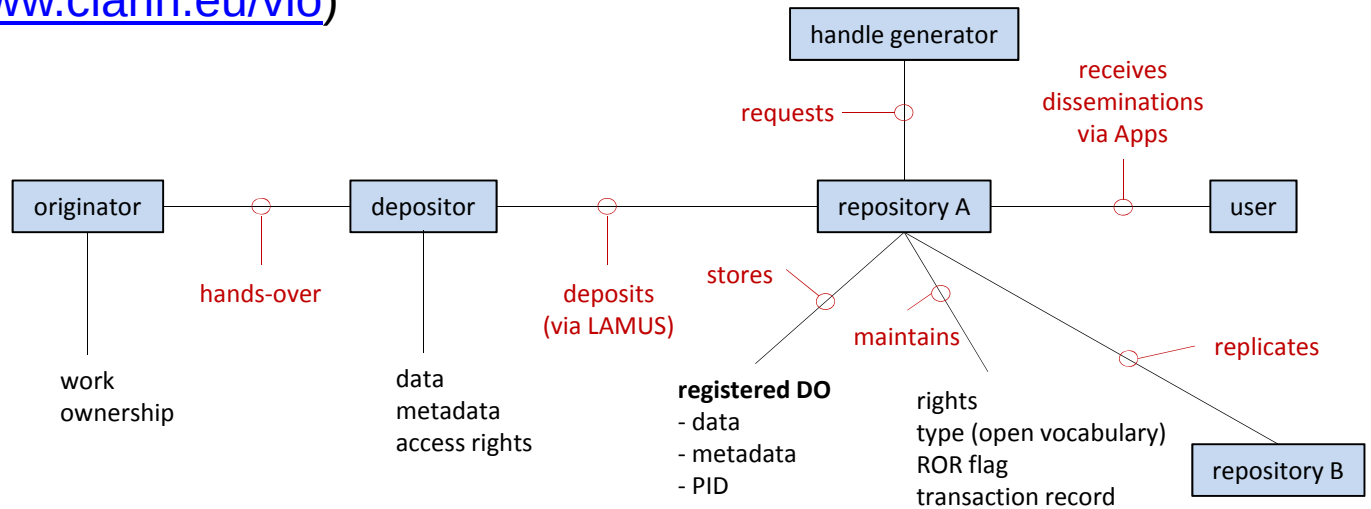
Definitions/Entities

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

EUDAT

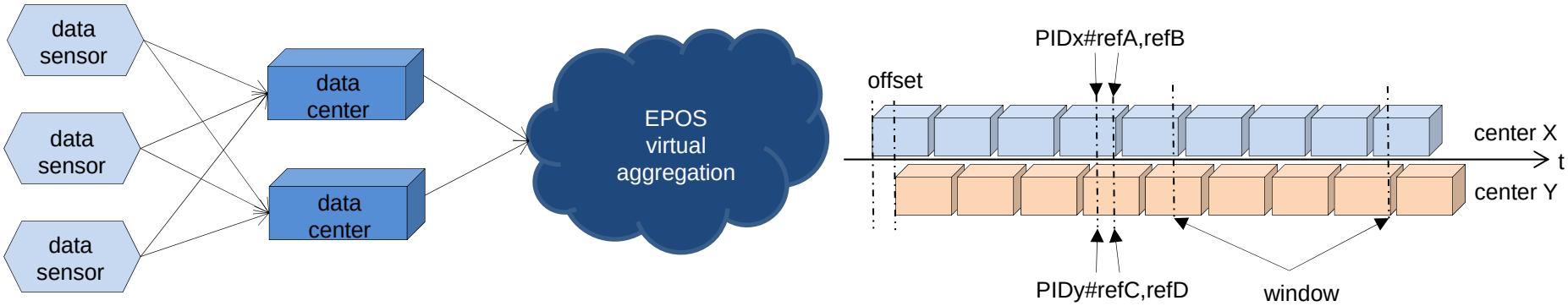
Data Landscape Analysis: CLARIN

- **CLARIN (Language Resource and Technology Community)**
  - about 200 centers in Europe with about 30 „community center“ candidates
  - have 4 types of centers (DataONE: tiers) from strong to weak requirements
  - requirements: rep. system, PIDs, CMDI based metadata, AAI
  - almost all busy with re-structuring - only few fulfill strong requirements
  - components/profiles and concepts registered (ISOcat, SCHEMcat)
  - Virtual Language Observatory: harvesting, mapping, indexing (www.clarin.eu/vlo)



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



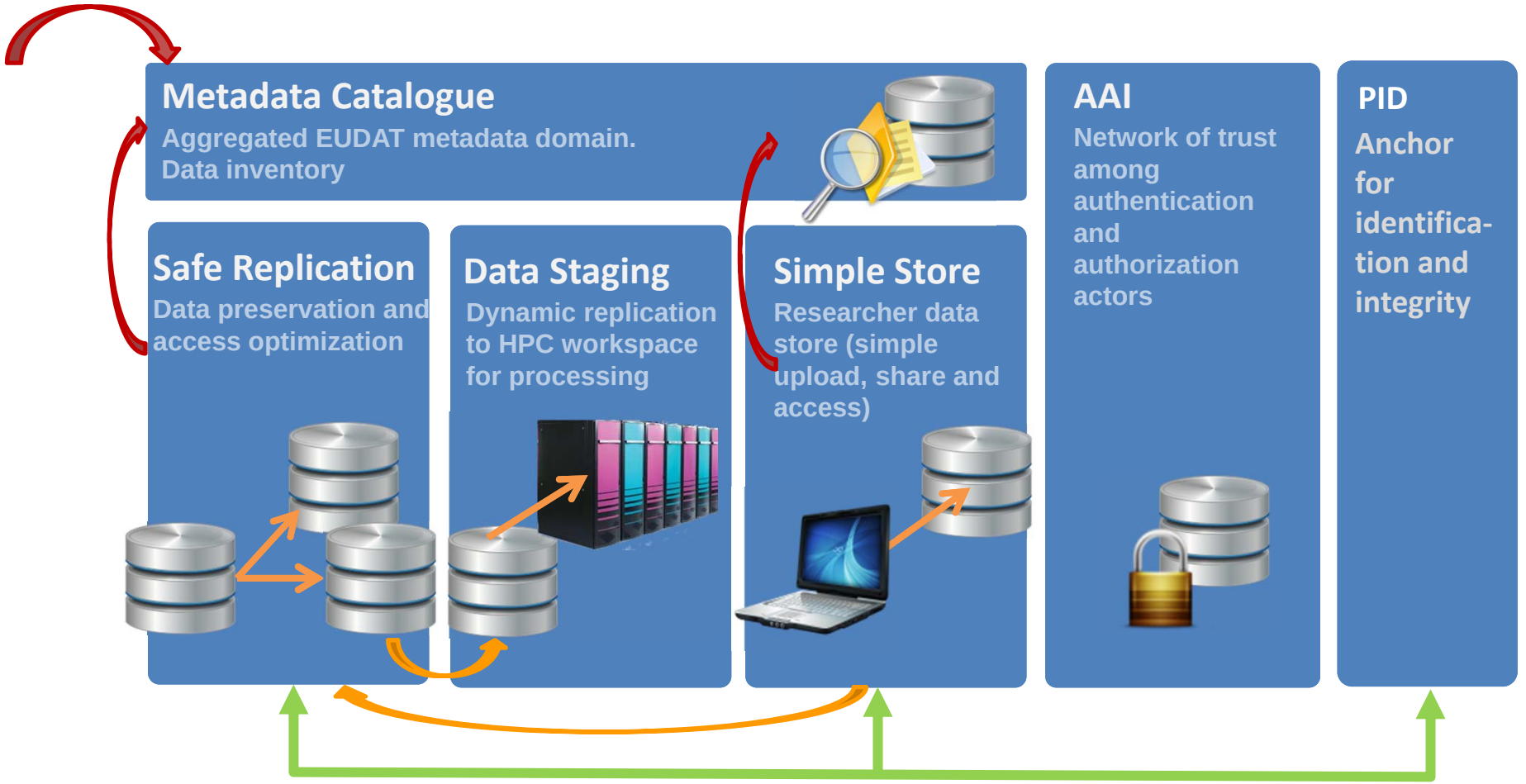
Common Data Services in CDI

Why high priority on fast delivery?

- ❑ **communities need concrete service examples to understand potential and impact**
- ❑ **you need to do it to understand the many practical problems, to get a sense about the nature of collaboration and shared responsibility, to get an idea about costs, etc.**
- ❑ **finally the requirements are emerging while doing**

EUDAT

First Services in Preparation



EUDAT

Data PID Metadata

Services working on

Common Services	CLARIN	LW	VPH	ENES	EPOS	INCF	ECRIN	Bio Vel	Dixa	CESSDA	DARIAH	Pan Data	BBMRI	EMSO
Safe Replication	X	o	X	X	X	X			x		x			
Replica Access	X		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	o	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	CLARIN	LW	VPH	ENES	EPOS	INCF	ECRIN	Bio Vel	Dixa	CESSDA	DARIAH	Pan Data	BBMRI	EMSO
Replica Access	X		X	X	X	X			x		x			
Semantic Annotation	o	X												
Web-service platform	X	o		X	o									
Real Time Data					X									
Memento Service														

Why Memento service?

- **still a split between ”big” data practices and web practices**
- **bridging the two worlds**

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

Enabling services required

- some machinery required below the surface
- also partly defined by community specifications
  - PID Service
  - distributed AAI
  - site registry
  - monitoring
  - hosting
  - workspaces
  - ticket system
  - etc

EUDAT: where are we?

- Prototype Services are in progress after about 1 year of work
  - SR and DS in operation for a few data centers of core communities
  - SS and MD will come in Q1
- 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
- do we know whether EUDAT can become a sustainable organization in Europe
  - is technology a problem? – well hard to solve but we can get there
  - Funders such as EC don’t want to spend money for long term – is this ok?
  - No – too many national aspects
  - No - there are these administrative restrictions for example
  - Panta Rhei – data organizations are changing almost everywhere

EUDAT

Thanks for the attention.

http://www.eudat.eu

Join the Research Data Alliance Meeting

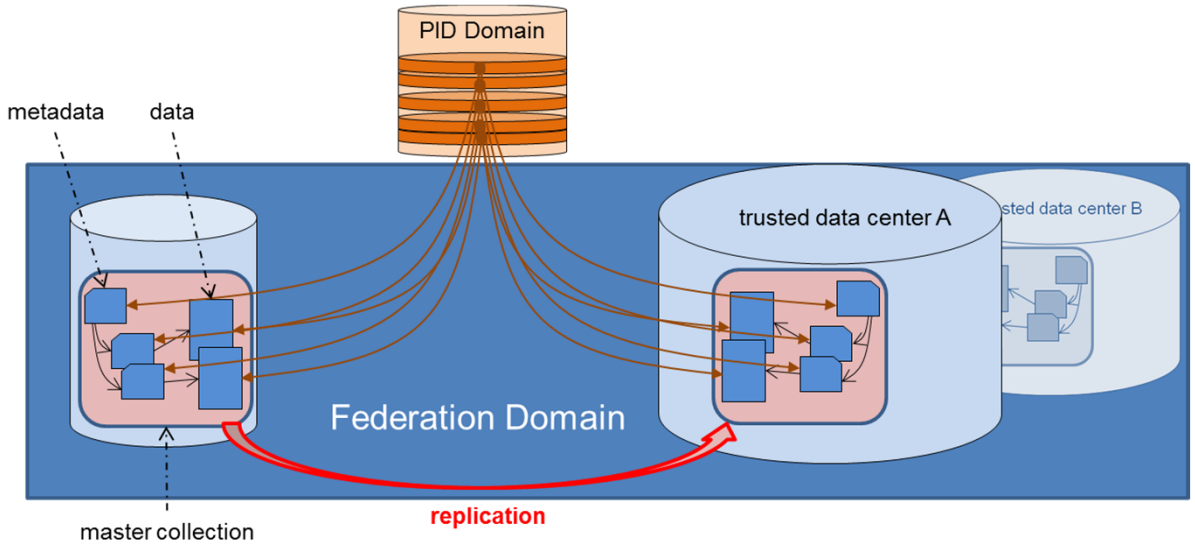
http://rd-alliance.org/invitation/

http://forum.rd-alliance.org

EUDAT

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



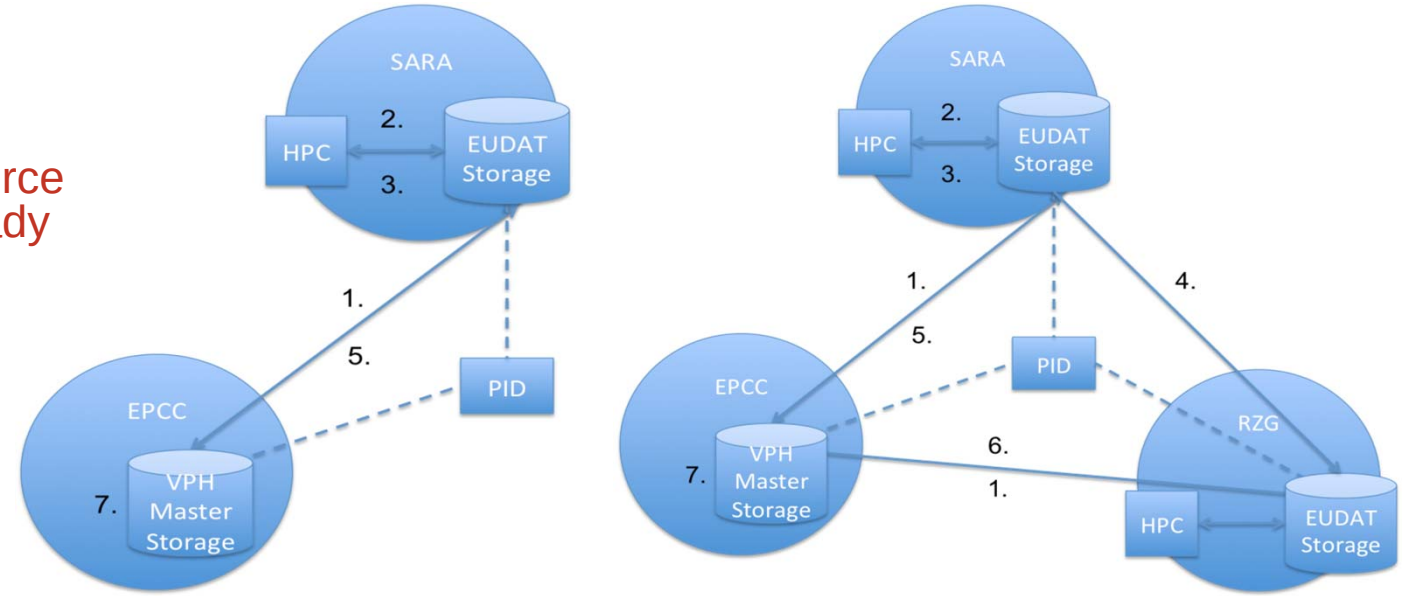
- basic technlogies: AAI, iRODS, Handles, community MD & OAI-PMH, center registry
- in June merging of 3 islands to one flexible replication domain
- REPLIX experience is basis

EUDAT

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

EUDAT

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

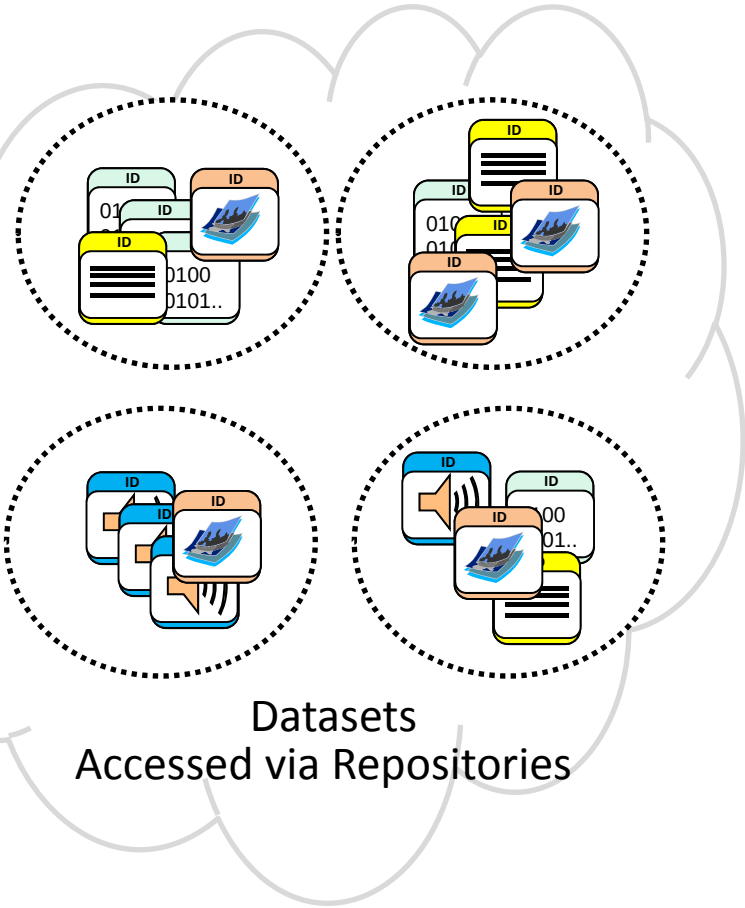
Scientists, Data Curators, End Users, Applications

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

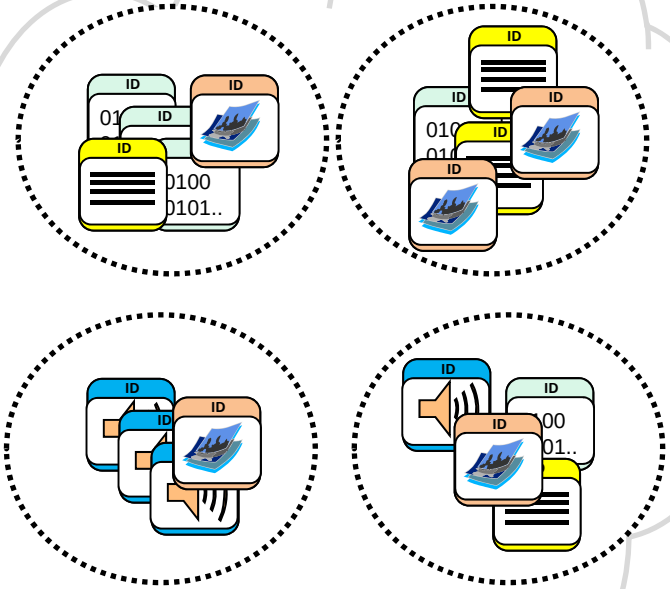
Enabling Technologies

Collections + Properties

Access (ref. resolution, protocols, AAI)

formalized policies workflow engine

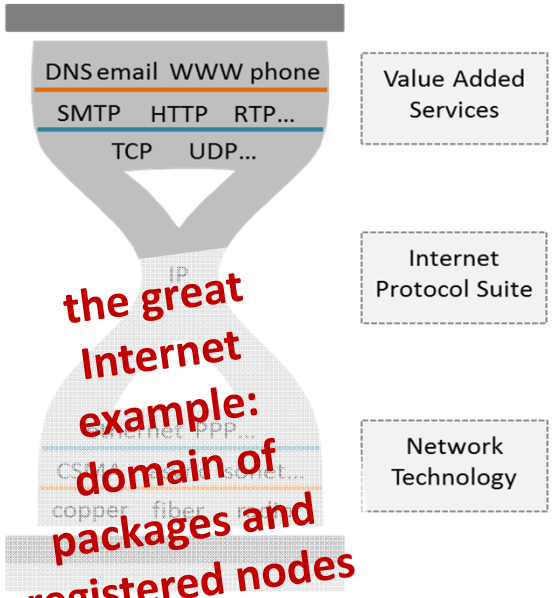
Assessment



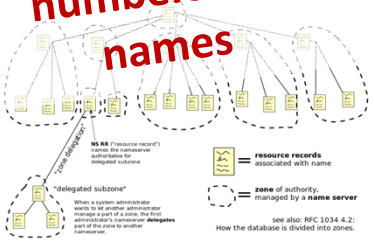
Datasets
Accessed via Repositories

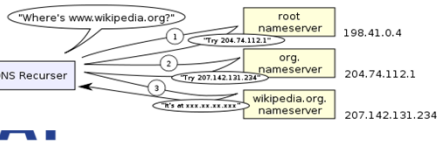
all in collaboration with CNRI

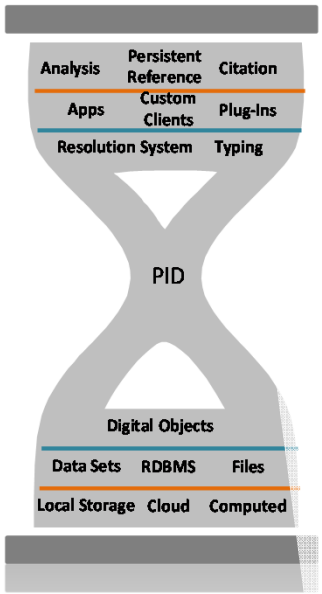
need to agree on basic models & terms



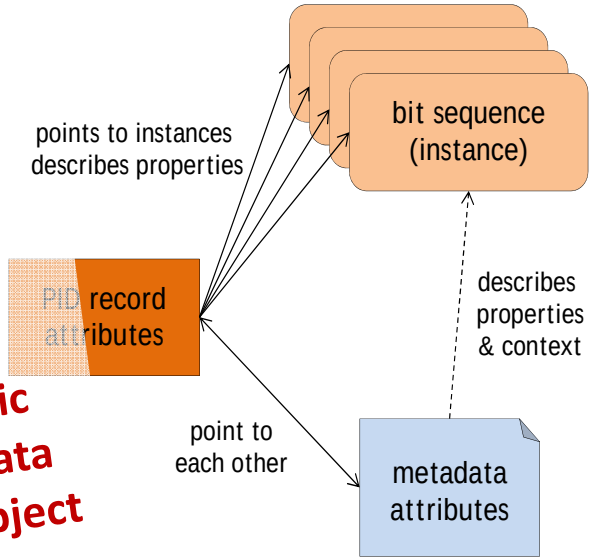
the great Internet example: domain of packages and registered nodes numbers and names







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

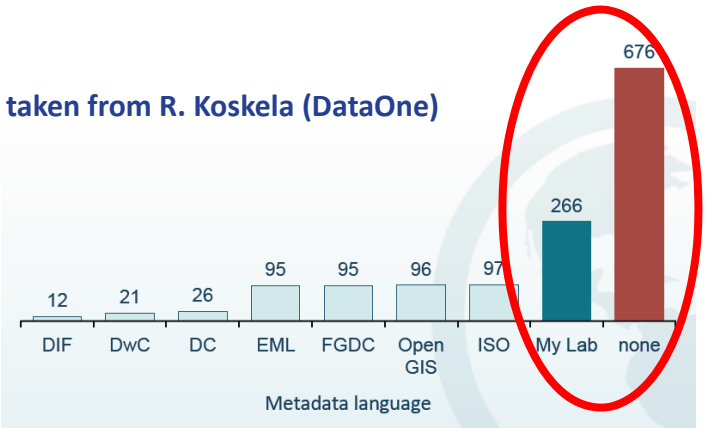


- let's come to a common object model with PID as anchors – like IP numbers in networks
- PID and MD store properties of objects and collections, policy rules manipulate properties
- EUDAT is a domain domain of registered data objects

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

EUDAT

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:

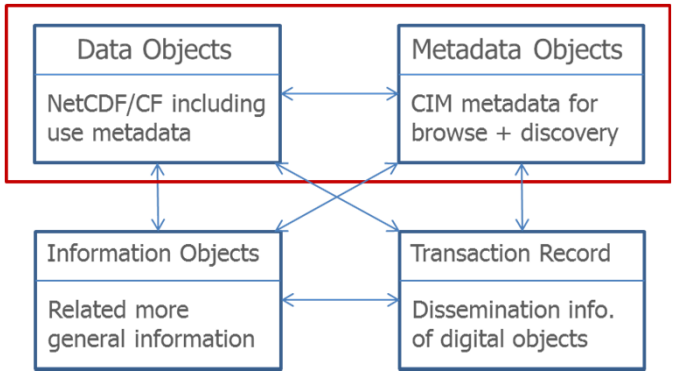
IETF	focus on networking
W3C	focus on the Web and its mechanisms
CODATA	focus on policies in area of data
World Data Systems	focus on proper data centers
G8+O5 Data Group	also focus on policies in area of data

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

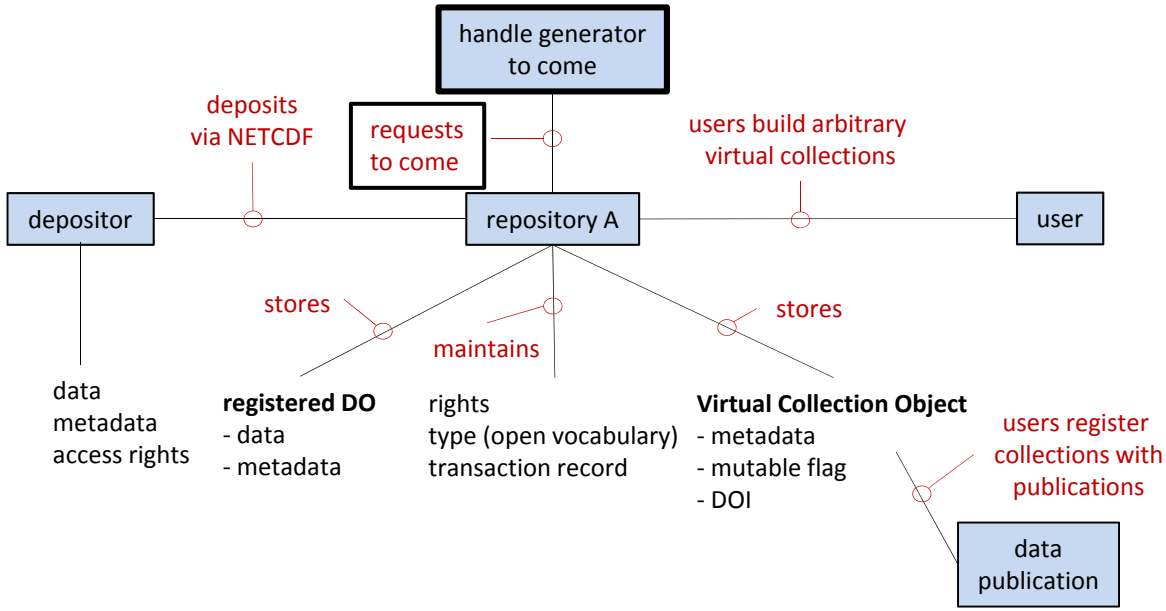
EUDAT

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

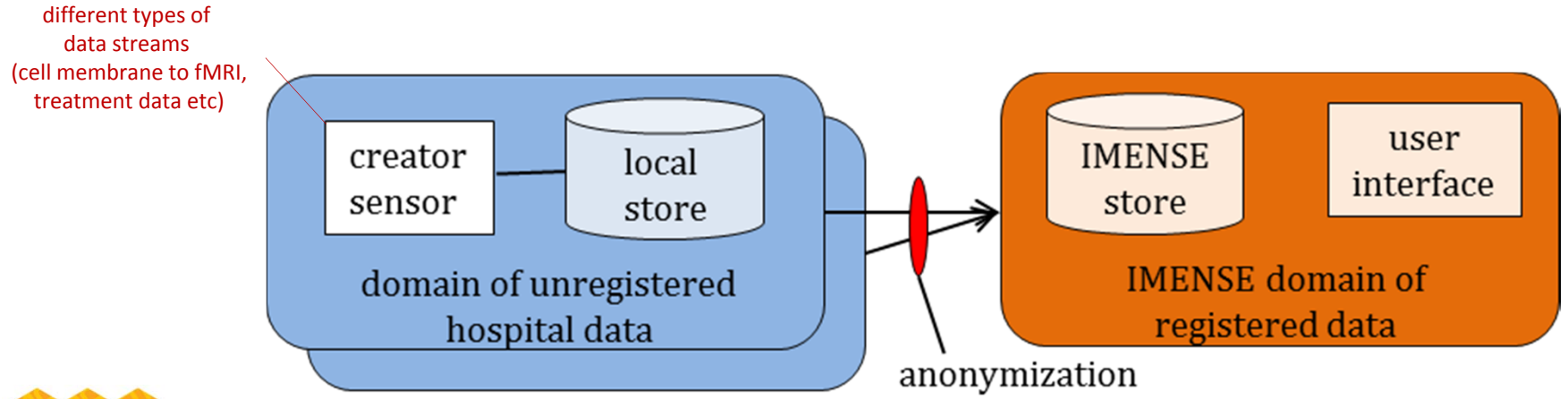


Identification of distinct data objects and P2P infrastructure



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



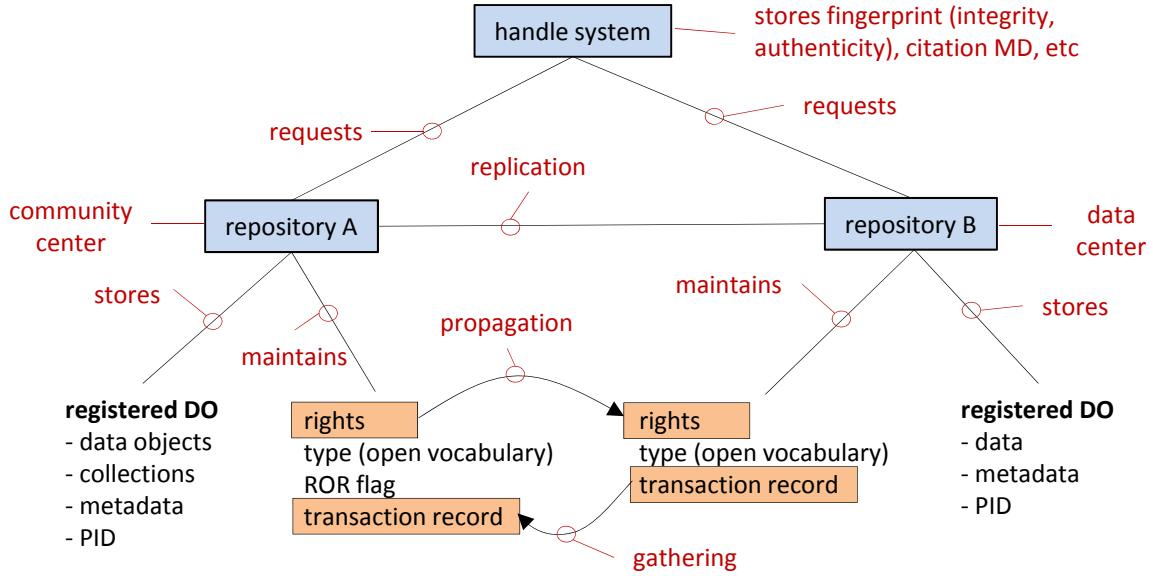
Data Landscape Analysis: LifeWatch

- **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)