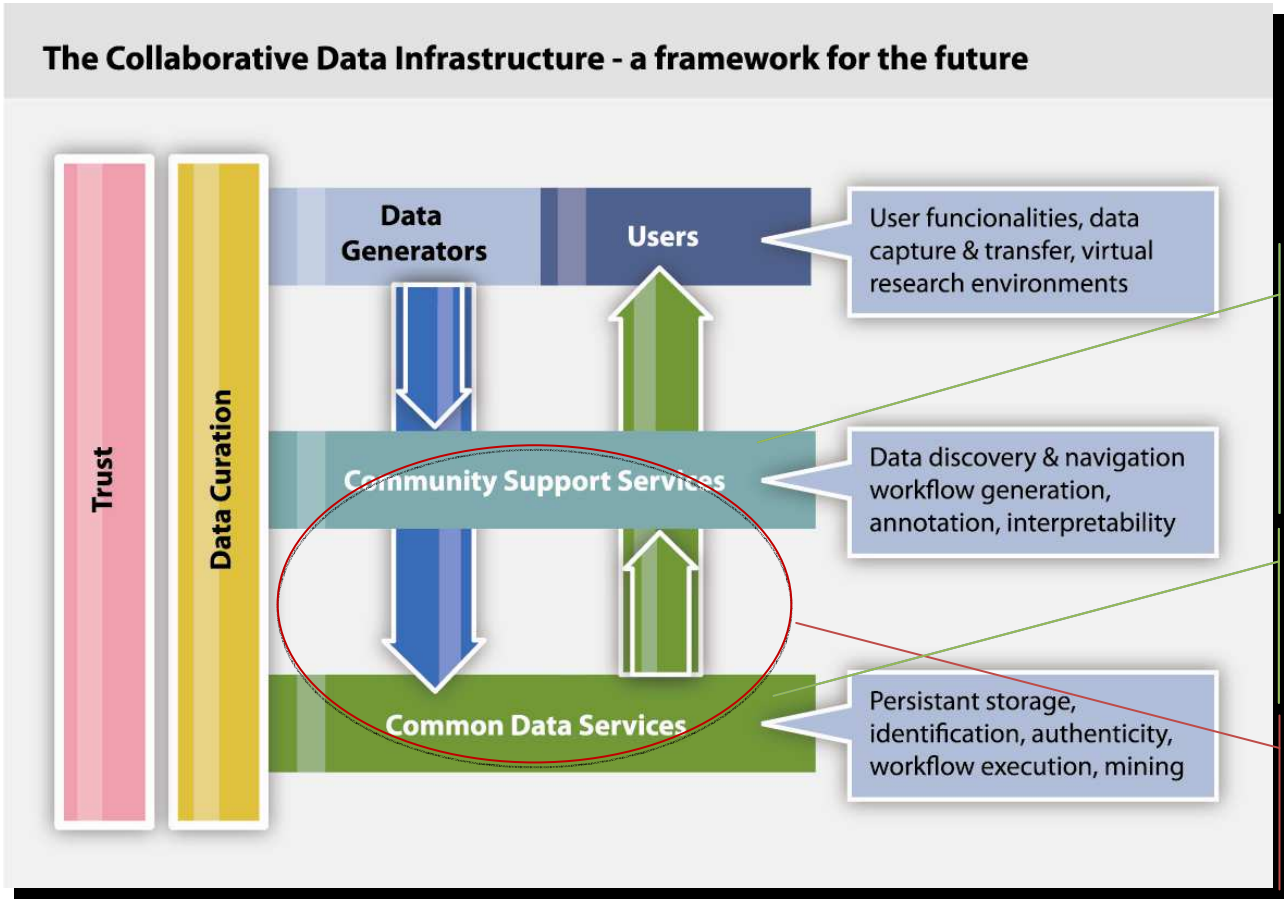
CDI and EUDAT

Peter Wittenburg, Daan Broeder
The Language Archive, Max Planck Institute, Netherlands
UF, Barcelona

EUDAT

March 2012

First - need to understand CDI



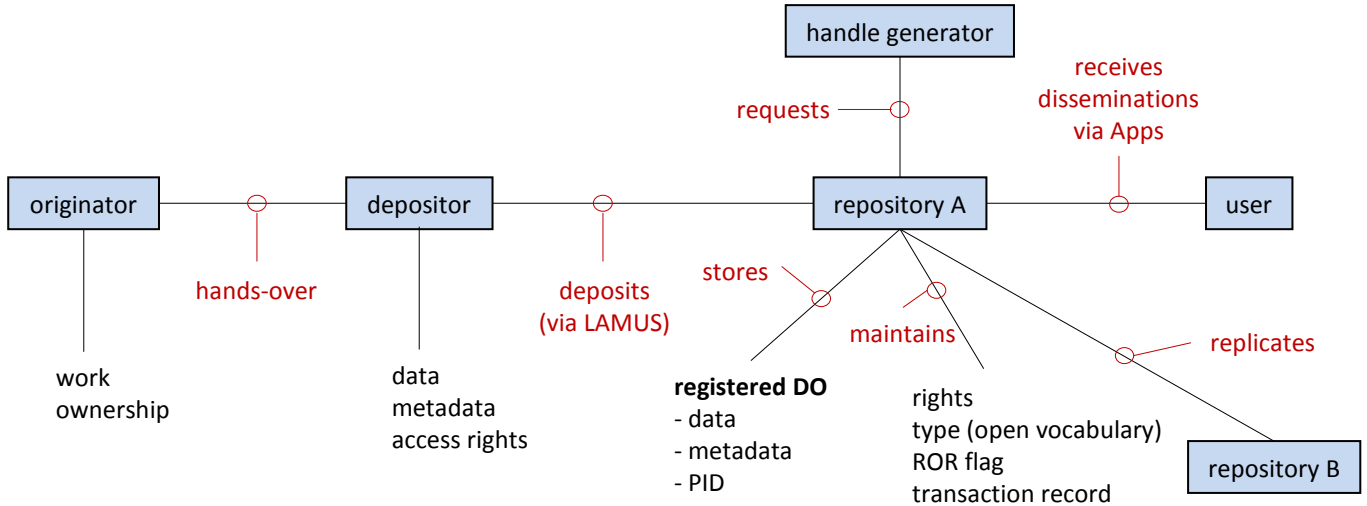
CLARIN, LifeWatch, ENES, EPOS, VPH, etc.
5 Core Infrastructures
~15 second round infrastructures

=> 10 EUDAT data centers

indeed some heterogeneity at both levels

Data Landscape Analysis: CLARIN

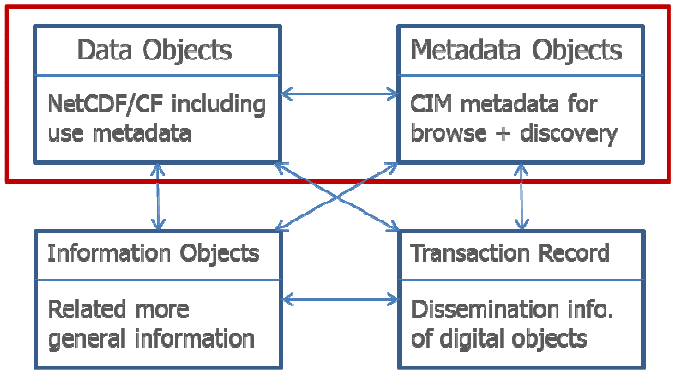
- **CLARIN (Language Resource and Technology Community)**
  - about 200 centers in Europe with about 30 „community center“ candidates
    - requirements: rep. system, PIDs, CMDI based metadata, AAI
    - almost all busy with re-structuring - only few fulfill strong requirements
  - components/profiles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



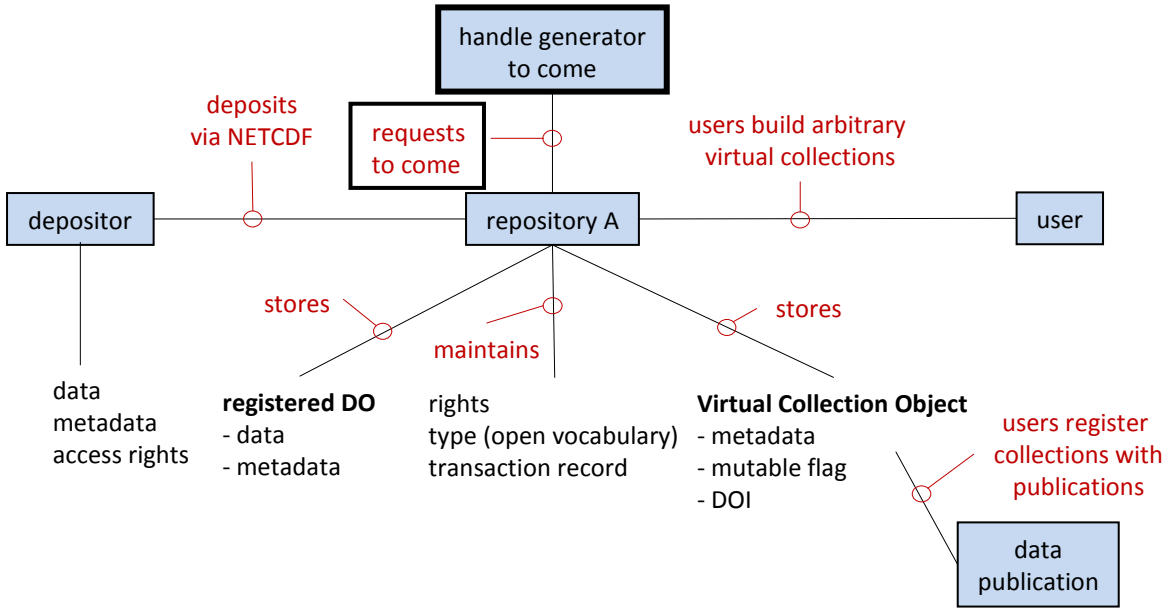
Data Landscape Analysis: ENES

•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

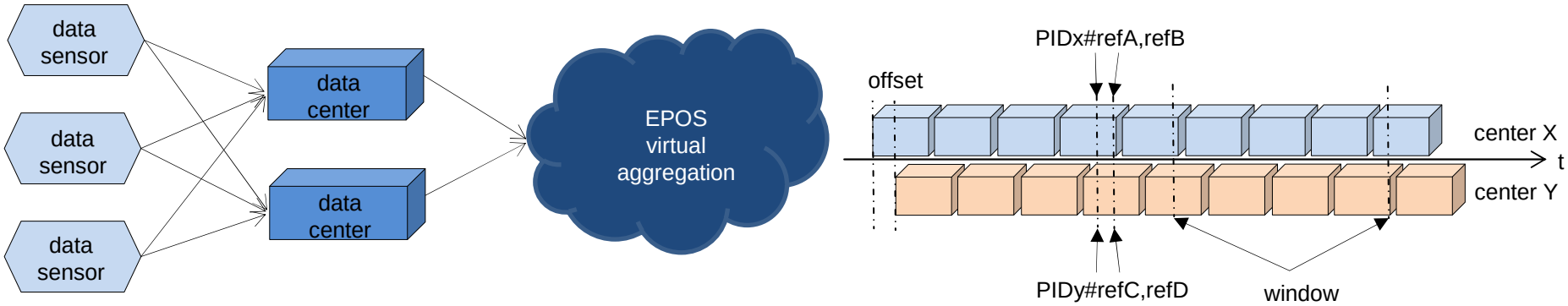


Identification of distinct data objects and P2P infrastructure



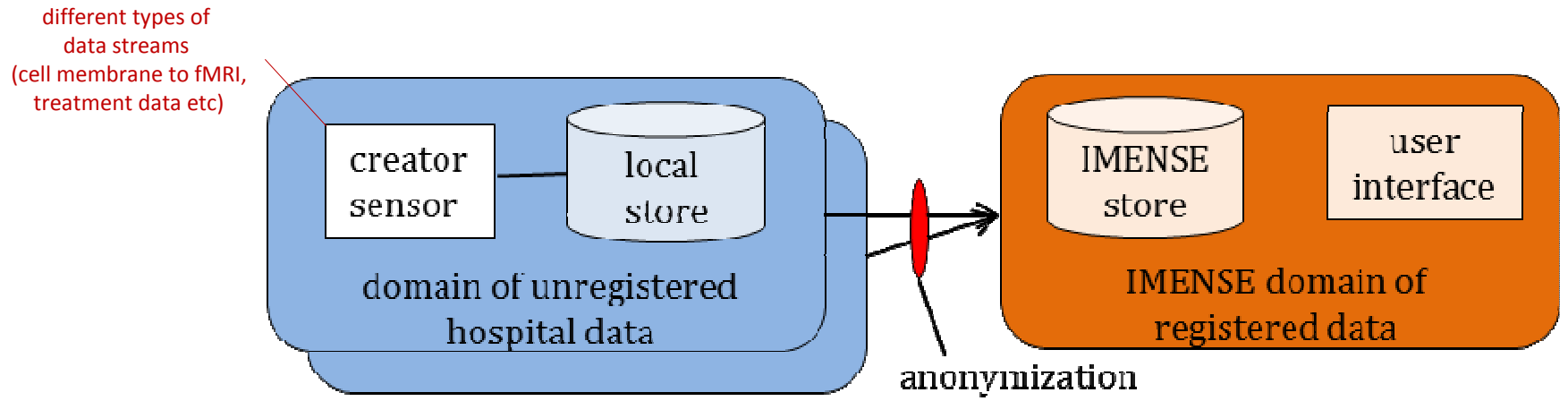
Data Landscape Analysis: EPOS

- **EPOS (Seismologists, Volcanologists, 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 „volcano 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



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
  - metadata not yet standardized & formalized (DICOM, JPEG headers, etc.)
  - nothing done with PIDs, AAI and OAI-PMH yet



Data Landscape Analysis: 2nd Round

- **second round of interviews to come in March**

Environmental Science	**ENES, EPOS,** Lifewatch, EMSO, IAGOS-ERI, ICOS, Euro-Argo, …
Social Sciences and Humanities	**CLARIN,** CESSDA, DARIAH, …
Biological and Medical Science	**VPH**, ELIXIR, BBRMI, ECRIN, DiXA, …
Physical Sciences and Engineering	WLCG, ISIS, DESY, PanData, …
Material Science	ESS, …

EUDAT

Data Landscape Analysis: Summary

- **panta rei - all is moving**
  - data infrastructures are shooting on a moving target
    - from core communities only 2 have a ready made architecture
  - process of discussion is rather fruitful
    - forces explicitness and fosters harmonization
    - discussions and moderation roles are highly appreciated
  - data volumes ready to be contributed range from Petabytes to Terabytes

EUDAT

Community Service Wishes

In Progress as Services (Task Forces set up)

- Safe Data Replication (for Bit-stream Preservation & Access Optimization)
- Dynamic Data Replication into HPC Workspace

In Specification/Discussion as Services

- Aggregated EUDAT Metadata Domain
- Researcher Data Store (Simple Upload, Share and Access)

In Progress as Research Issues (WP7)

- more elaborate policy rules and federation scalability
- generic workflow execution framework
  (automatic annotation, data mining, etc.)

EUDAT

Enabling Technologies

- **Building robust and available persistent identifier service (is in place based on Handles)** — ready to go
  - EPIC: millions of objetcs, DataCite: published collections
  - EPIC offers registration/resolution service for all data centers in Europe
  - EUDAT: all objects need to be registered, all policy operations will use PIDs
- **Federated AAI service** — not yet ready
  - Shib/SAML based world - still a mess due to fragmentation
  - can we rely on harmonized EU wide Identity Federation?
  - will individual indentity providers offer needed attributes?
- **Shared Workspaces** — to be done
  - obviously for different purposes (storing data, automatic annotations, etc)
- **Monitoring and accounting**
  - all participating servers/services need to show stability, availability
- **Network Services** **(of course)**

EUDAT

EUDAT CDI Summary

- understand data organizations as bottom-up exercise
- determine „common“ functions needed
- determine essential independent components with chance of wide acceptance
  - PID system, center registry, metadata landscape
- define agreed APIs for different components
- rely on policy-rule based approach