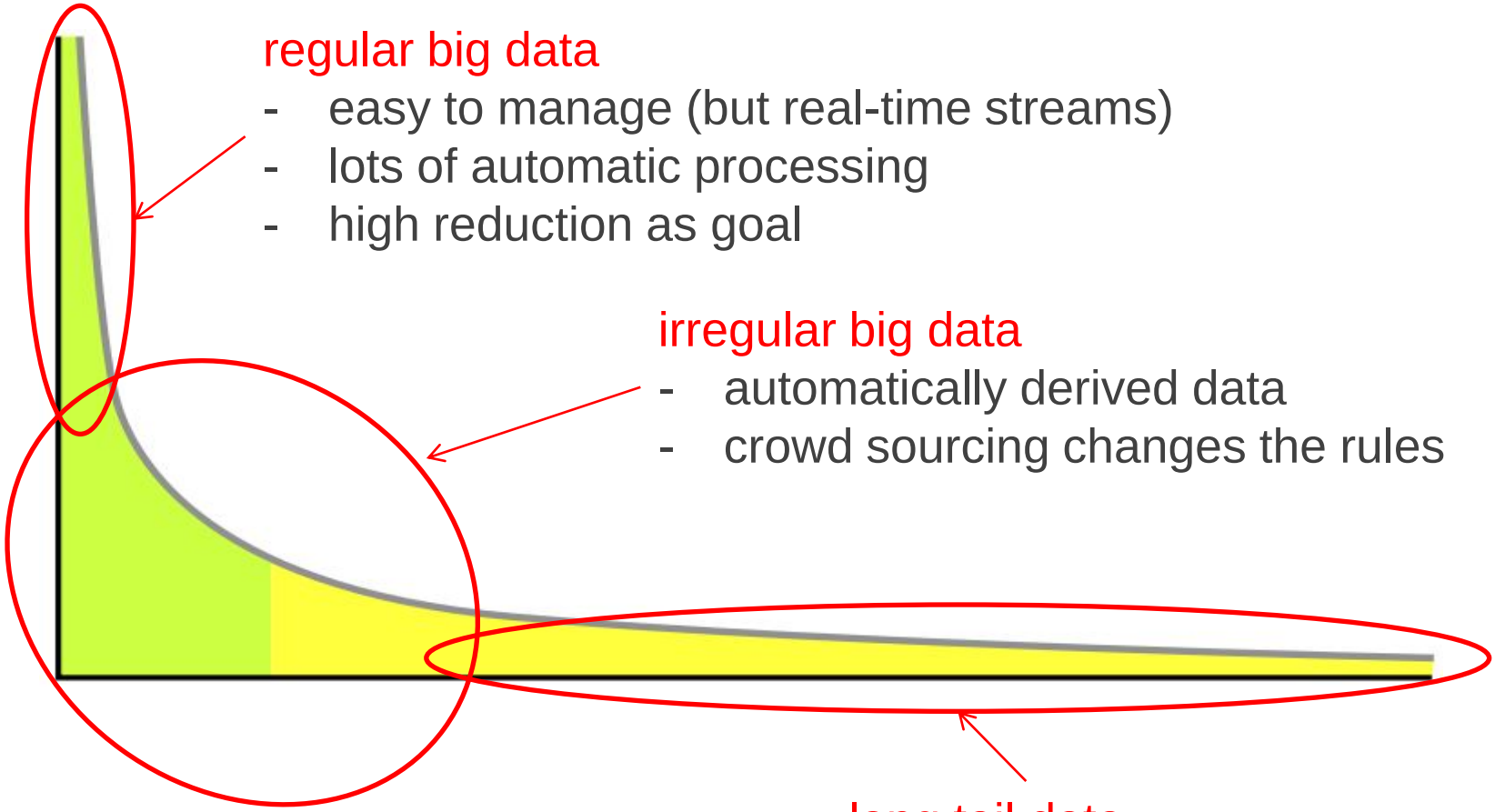
EUDAT
Common data infrastructure

Giuseppe Fiameni
SuperComputing Applications and Innovation
CINECA – Italy
Peter Wittenburg
Max Planck Institute for Psycholinguistics
Nijmegen, Netherlands

SEVENTH FRAMEWORK PROGRAMME

EUDAT

some major characteristics



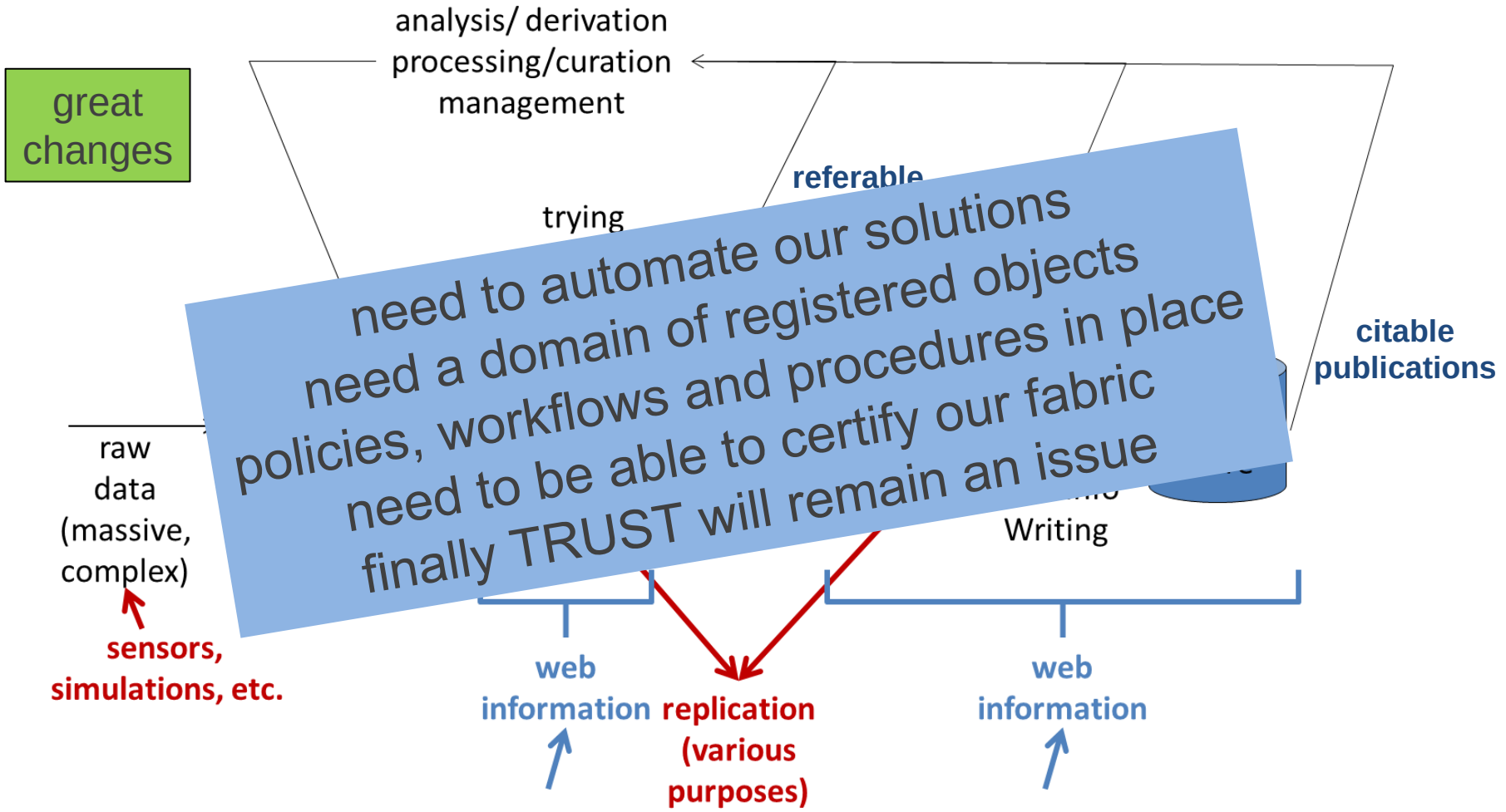
regular big data
- easy to manage (but real-time streams)
- lots of automatic processing
- high reduction as goal

irregular big data
- automatically derived data
- crowd sourcing changes the rules

long tail data
- difficult to manage
- lots of relations

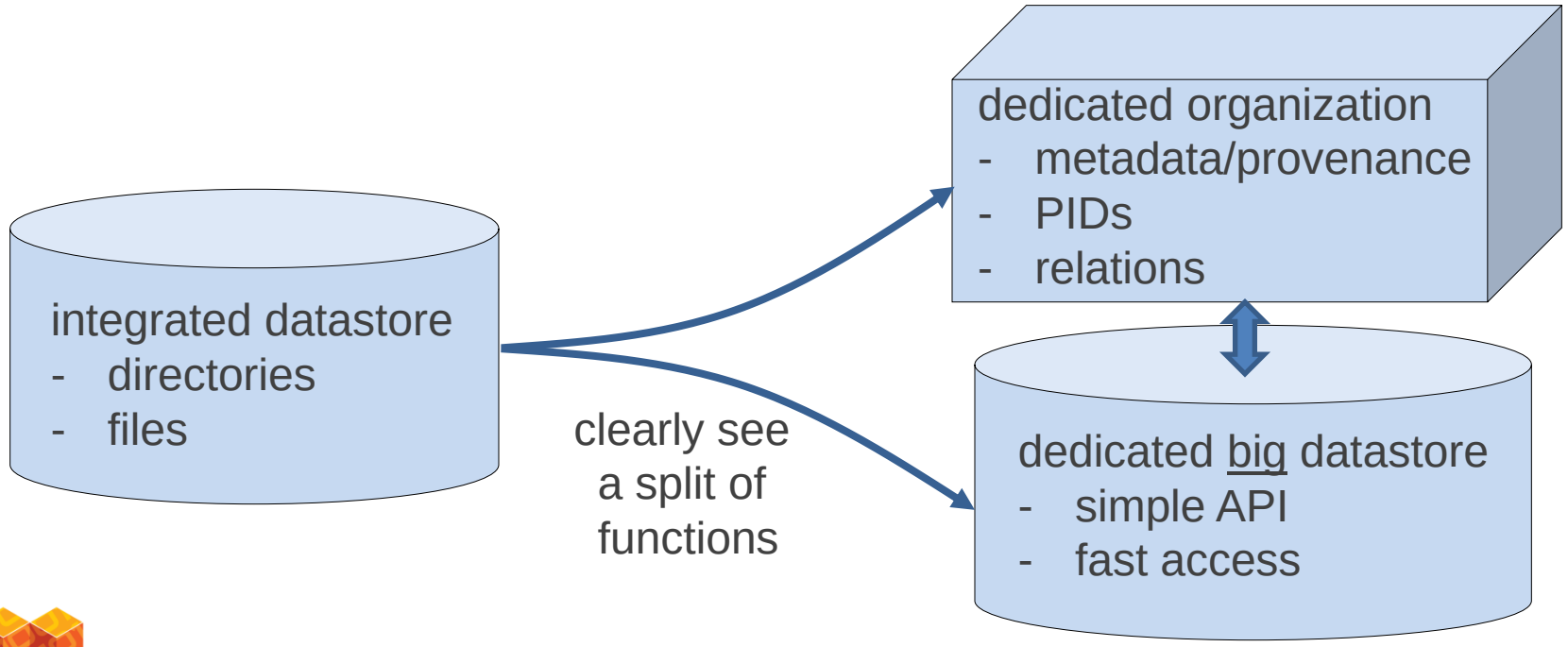
all the same for industry, government, public services, citizens, etc.

big scientific data –> the data fabric

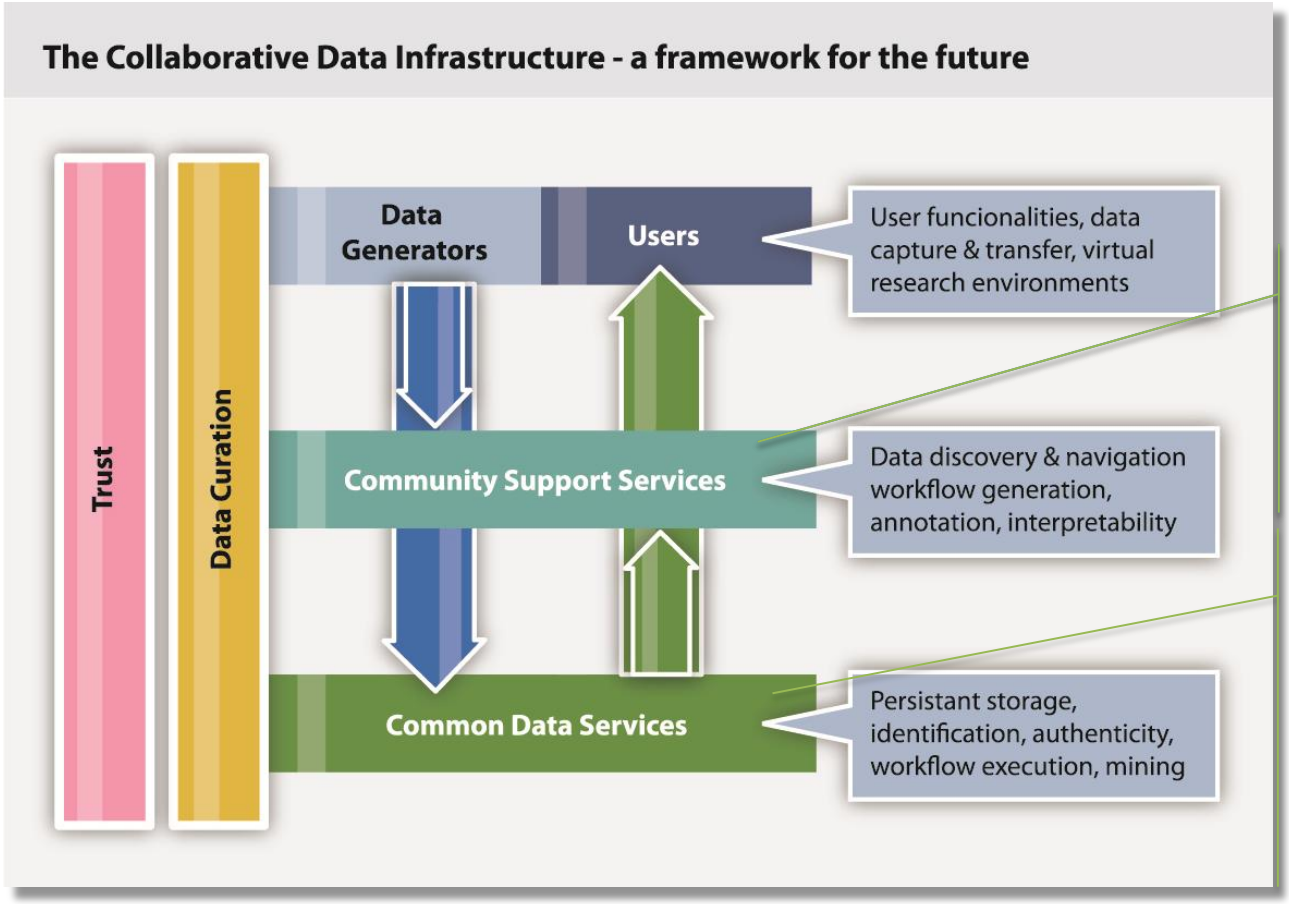


complexity is relevant

- filenames/directories are not sufficient anymore to memorize - even our experimentalists (brain images etc) start believing it
- lots of relationships (organization, content, provenance, etc.) to be stored
- many work on special aggregations (need to be named & stored)
- currently too much time lost with management



EUDAT's mission: common services in CDI



CLARIN, LifeWatch, ENES, EPOS, VPH, INFC etc.
6 Core Infrastructures about 20 infrastructures

⇒ 12 EUDAT data centers

⇒ **and/or cross-disciplinary initiatives**

diagram taken from EC's HLEG report "Riding the Wave"

common services EUDAT is working on

Metadata Catalogue
Aggregated EUDAT metadata domain.
Data inventory

AAI
Network of trust among authenti

PID
Identity
rity
enticity
ons

Safe Replication
Data curation and access optimi
Various

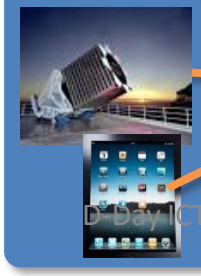
Data Staging

who is the driving force behind this?
it's the communities
(or at least the community representatives)
it's not the big centers
but they are important as a kind of glue
AND in some areas they have great IT knowledge
in some not so much (MD, SemAnn)

services to come

dropbox-like service
easy sharing
local synching

Dynamic Data
immediate handling



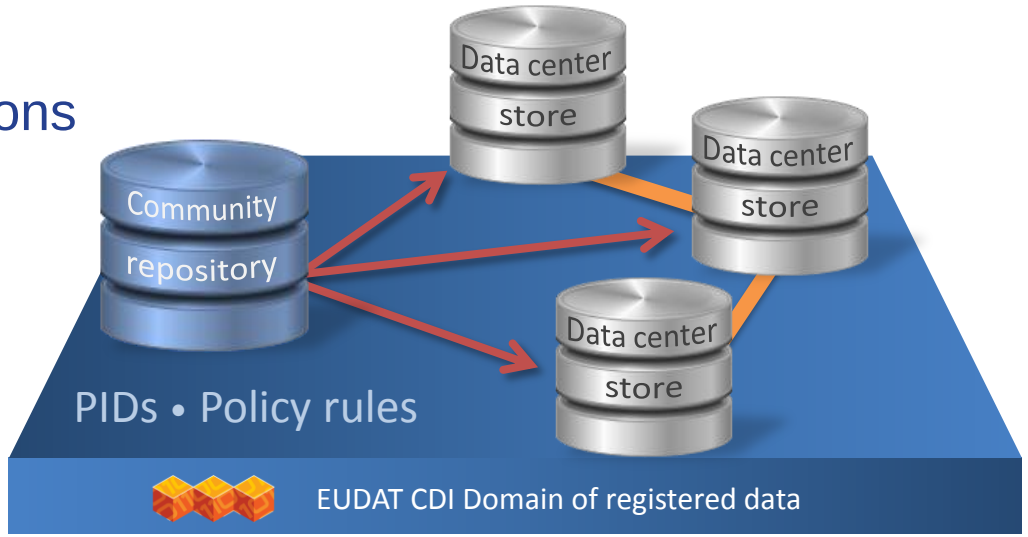
Semantic Anno
checking & referencing

what next ?

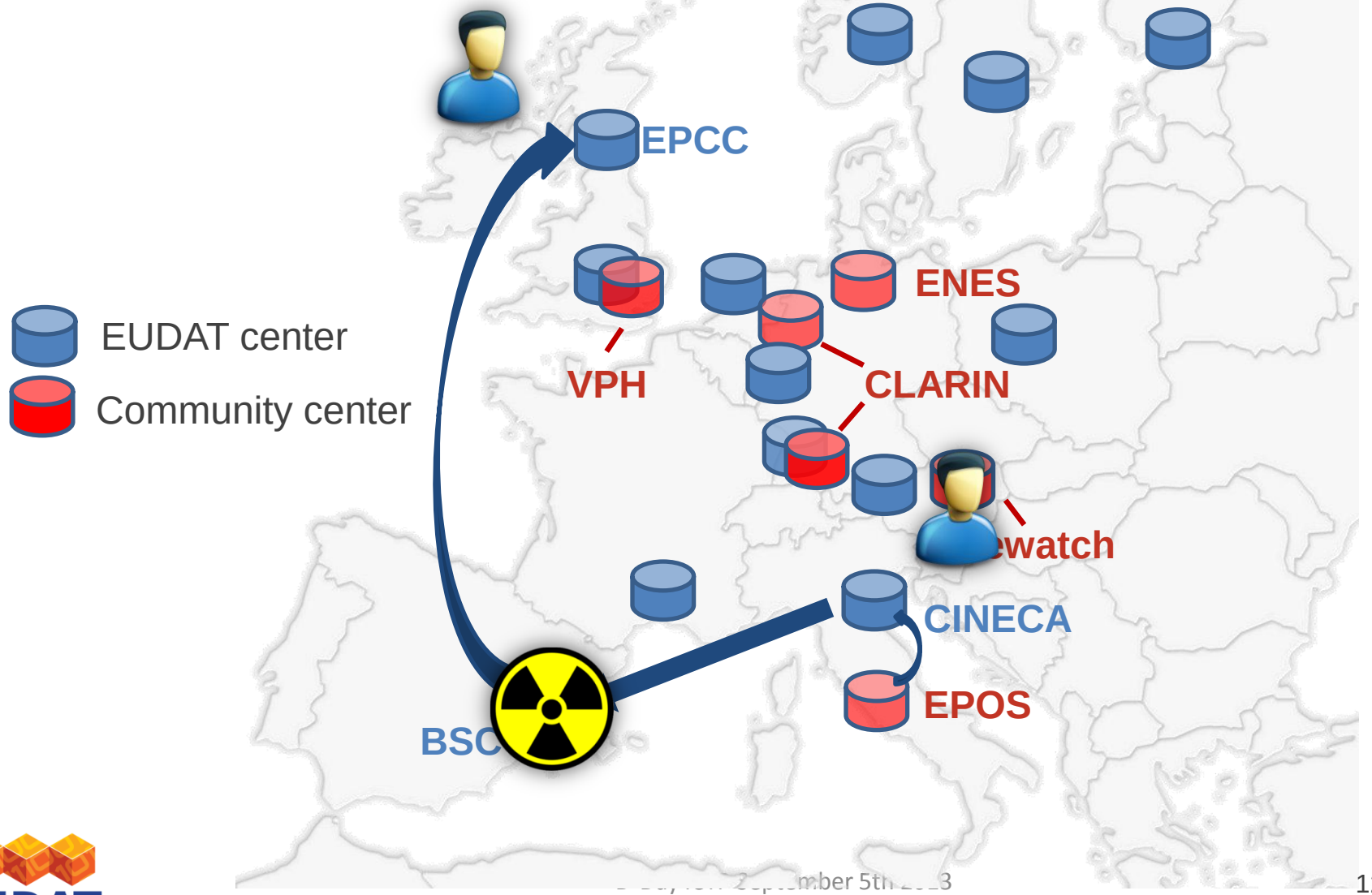
EUDAT

Safe Replication Service

- Robust, safe and highly available data replication service for small- and medium- sized repositories
  - To guard against data loss in long-term archiving and preservation
  - To optimize access for user from different regions
  - To bring data closer to powerful computers for compute-intensive analysis

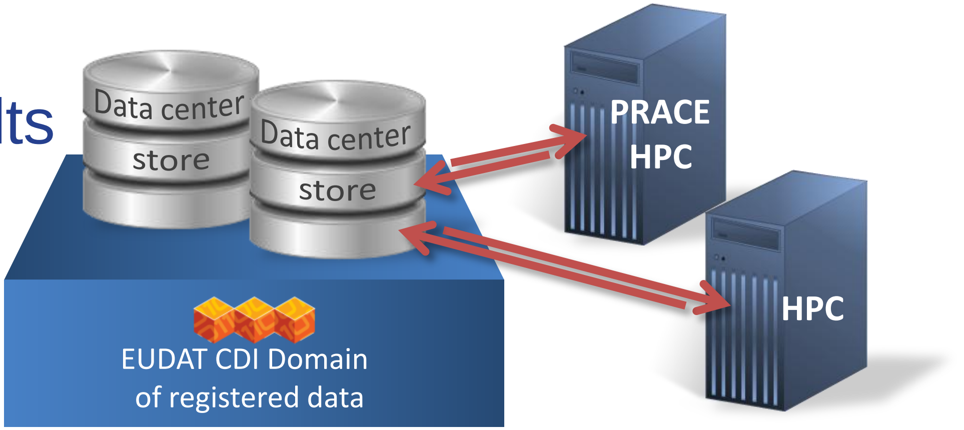


replicate my collection X to three data centres



Data Staging Service

- Support researchers in transferring large data collections from EUDAT storage to HPC facilities
- Reliable, efficient, and easy-to-use tools to manage data transfers
- Provide the means to re-ingest computational results back into the EUDAT infrastructure

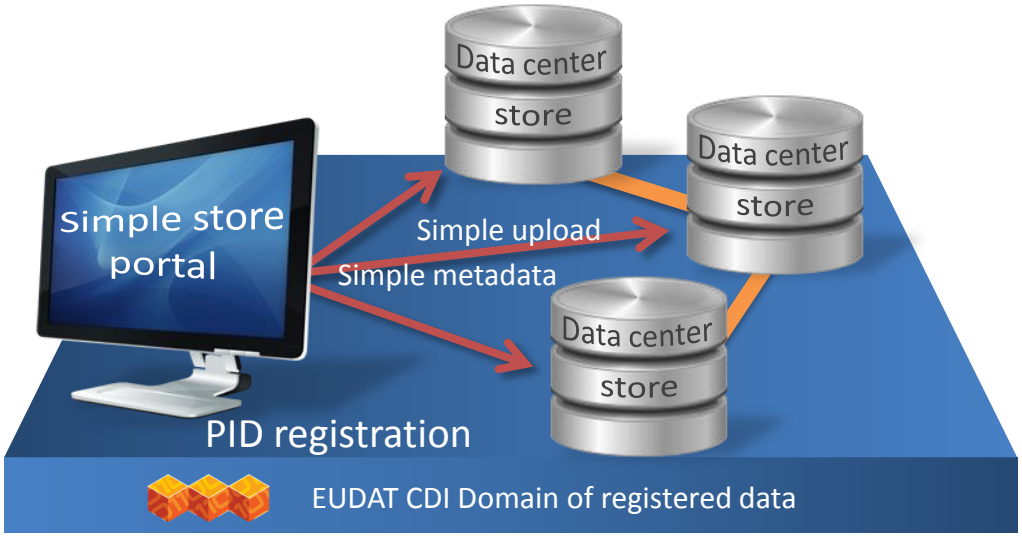


- not a simple service!
- politics involved (access to HPC)

http://eudat.eu/datastaging | eudat-datastaging@postit.csc.fi

Simple Store Service

- Allow registered users to upload ”long tail” data into the EUDAT store
- Enable sharing objects and collections with other researchers
- Utilise other EUDAT services to provide reliability

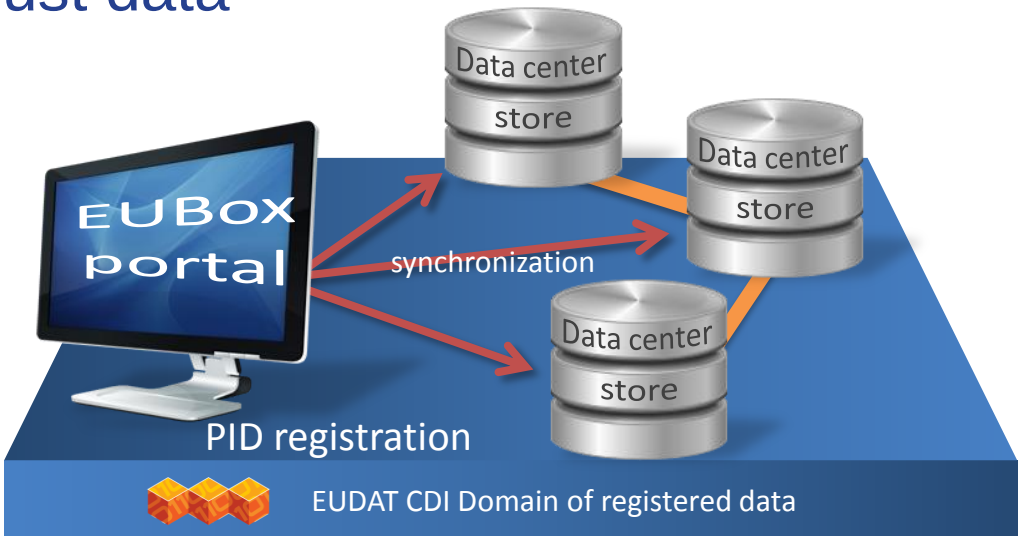


- much competition
- see it as complementary – finally it is about trust

EUDAT Box Service

- some similarity to SimpleStore of course
- just similar to Dropbox incl. load balancing and replication
- there is no metadata – just data

- how to integrate into registered domain of data?

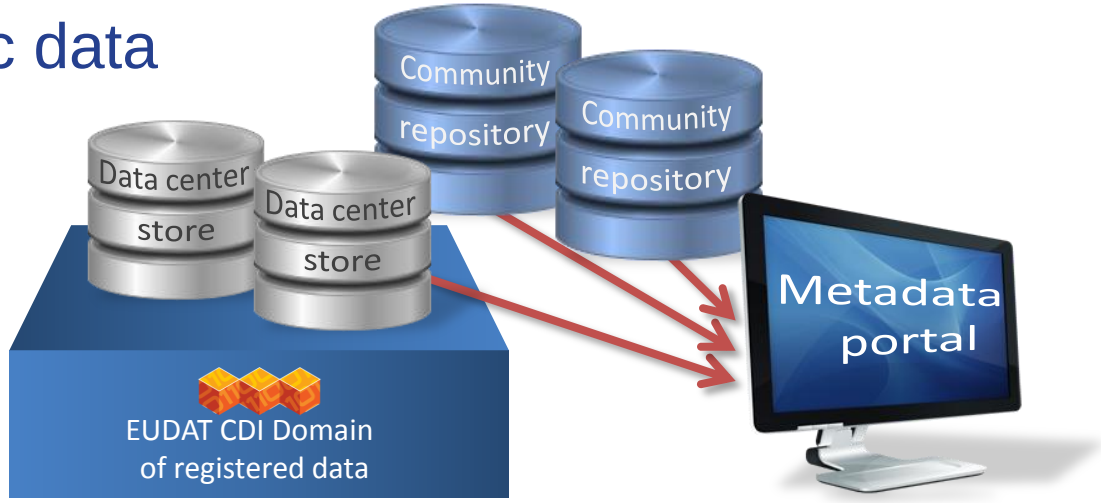


- much competition
- see it as complementary – finally it is about trust

Metadata Service

- Easily find collections of scientific data – generated either by various communities or via EUDAT services
- Access those data collections through the given references in the metadata to the relevant data stores
- Europeana of scientific data

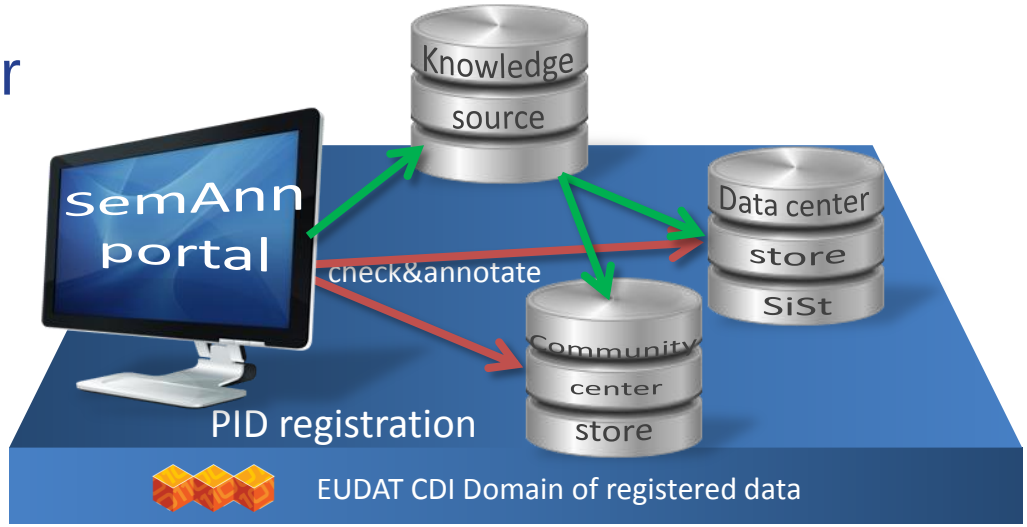
- how to offer metadata in a cross-disciplinary space?
- scalability issue?



Semantic Annotation Service

- acts as a plugin component to be executed before uploading a resource with tags (crowd sourcing etc.)
- check tags against Knowledge Source & correct/refer/etc.

- could be used as trigger in Simple Store
- plugin available to everyone



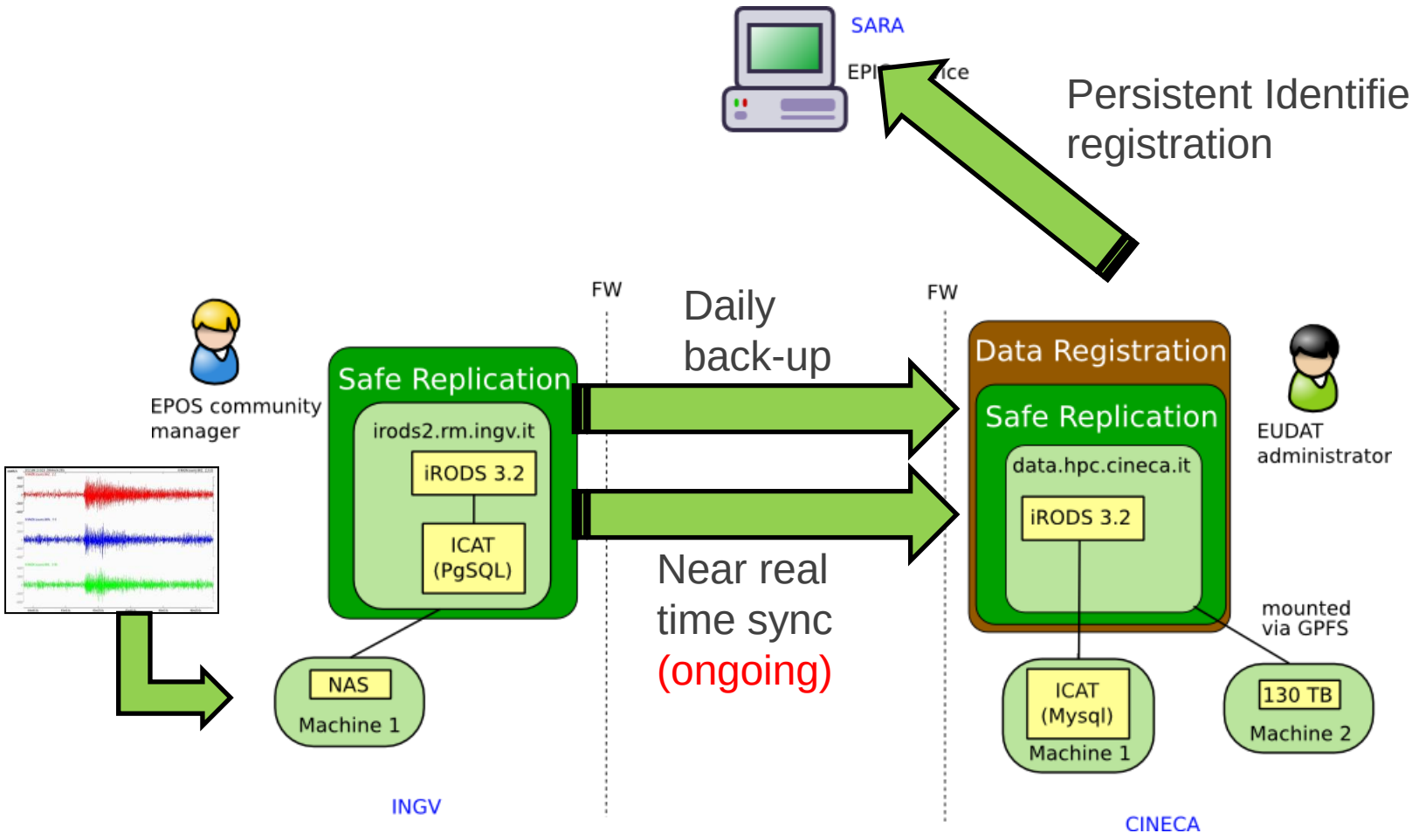
- not center dependent

service targeting

- Replication: targeted at data managers/archivists/ projects/departments without facilities
- Data Staging: same plus “easy” access to HPC
- SimpleStore: place for individuals/projects/groups to store & exchange data
- EUBox: share data via synchronization
- Metadata: EUDAT data & everyone interested
- SemAnn: individual/projects working with massive amounts of human created data

data stored in domain of registered data is not EUDAT’s data!
how to make this visible? – in SiSt community branding etc.

EPOS service implementation



is there a global challenge?

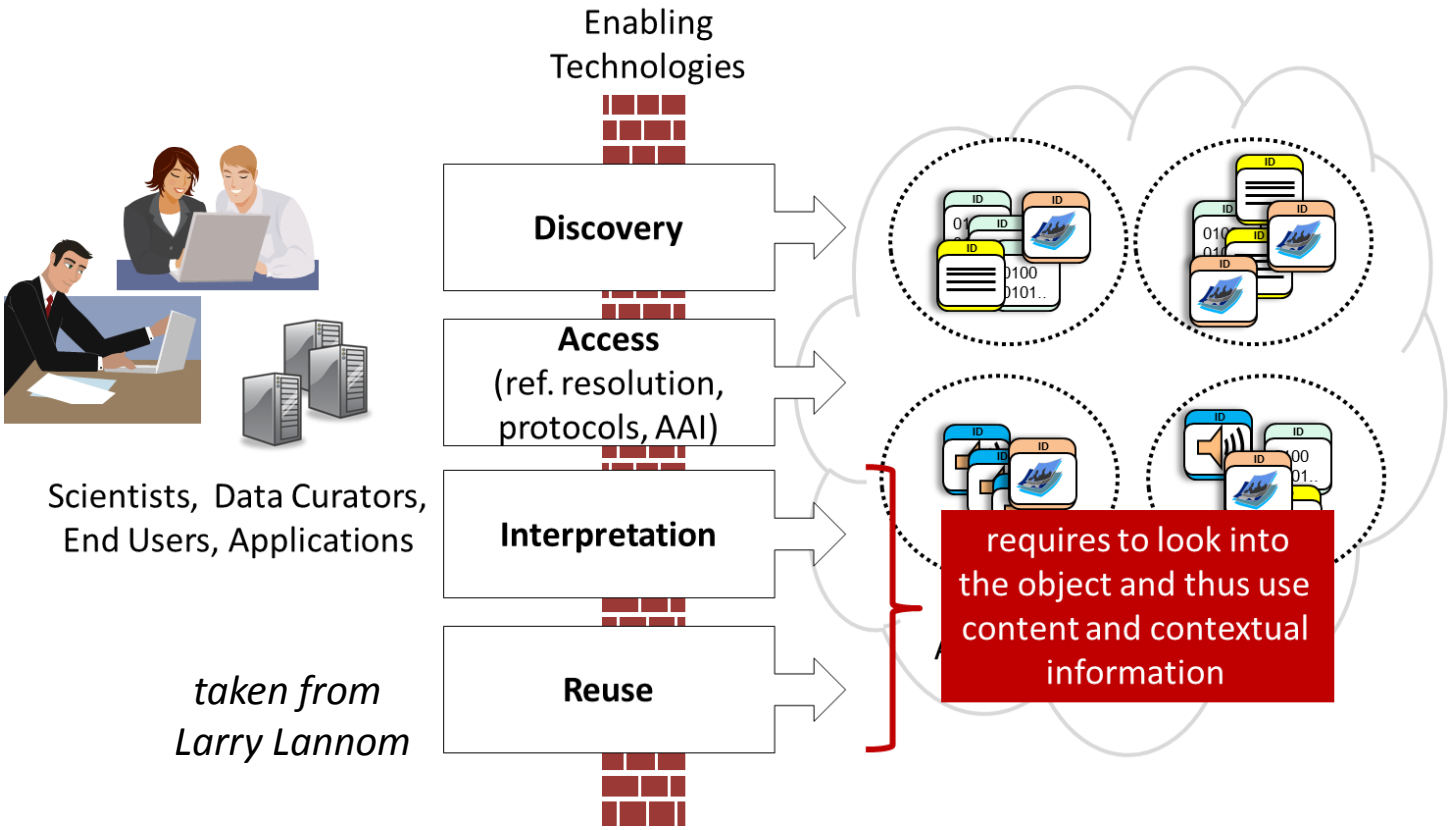
- EUDAT interfaces with many different data providers as do comparable initiatives such as DataONE, etc.
- currently little is compatible at various layers
  - infrastructure layer: no agreed components, no agreed APIs
  - content layer: formats, semantics (concept registration & bridging)
  - logical layer: PID + attributes, metadata principles + attributes, concept/schema registration, policies
- **something to be done – to be accelarated?**

who is working on it?

- different initiatives working on a variety of aspects (just a few)
  - ESFRI initiatives working on discipline interoperability and improving/harmonizing data landscapes – need harmonization
  - EUDAT working on common data services – need harmonization
  - OpenAIRE working on specific data service – need harmonization
  - Europeana working on aggregating metadata – need harmonization
  - etc.
- a variety of standardization and policy organizations
  - standards: ISO, IETC, IETF, W3C, OAI, OASIS, DONA, etc.
  - hl policies: CODATA, WDS, etc.
- some thought: we need a fast acting, bottom-up initiative focusing on removing barriers for sharing data

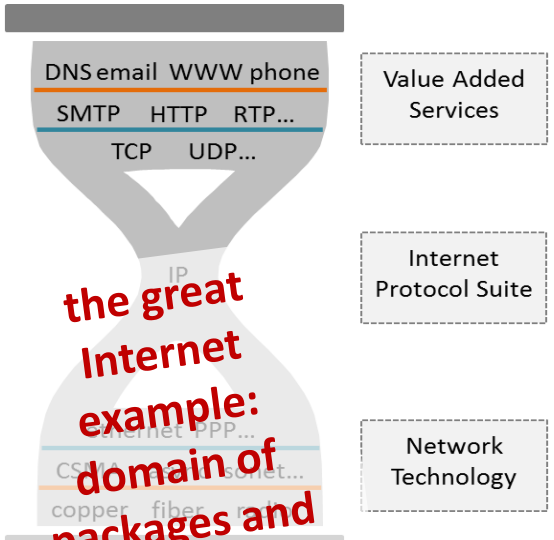
  → Research Data Alliance

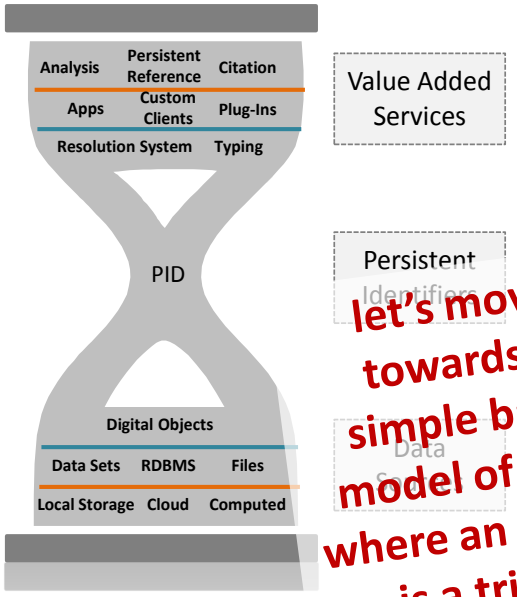
share canonical access procedure

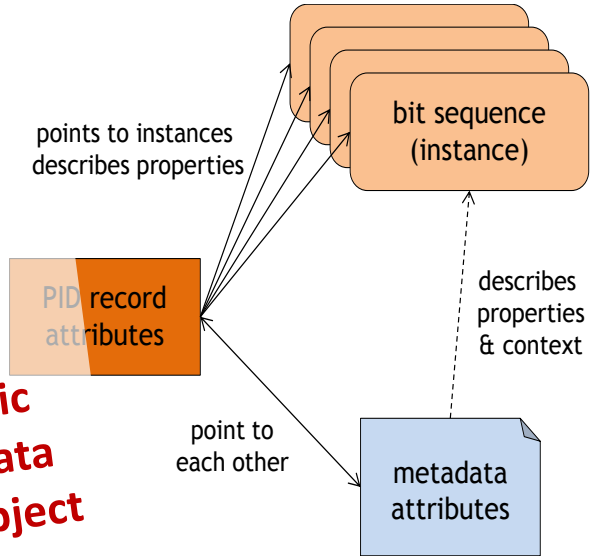


- need agreed ways to store and manipulate ext/int properties
- need agreed ways to do reference resolution (URIs vs. PIDs)
- need agreed ways to build common components or to rely on principles

learning from Internet

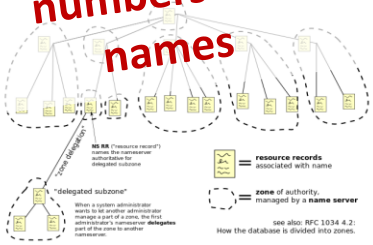


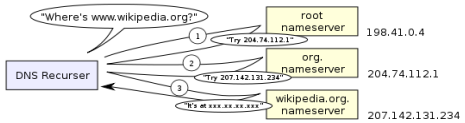




the great Internet example: domain of packages and registered nodes numbers and names

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





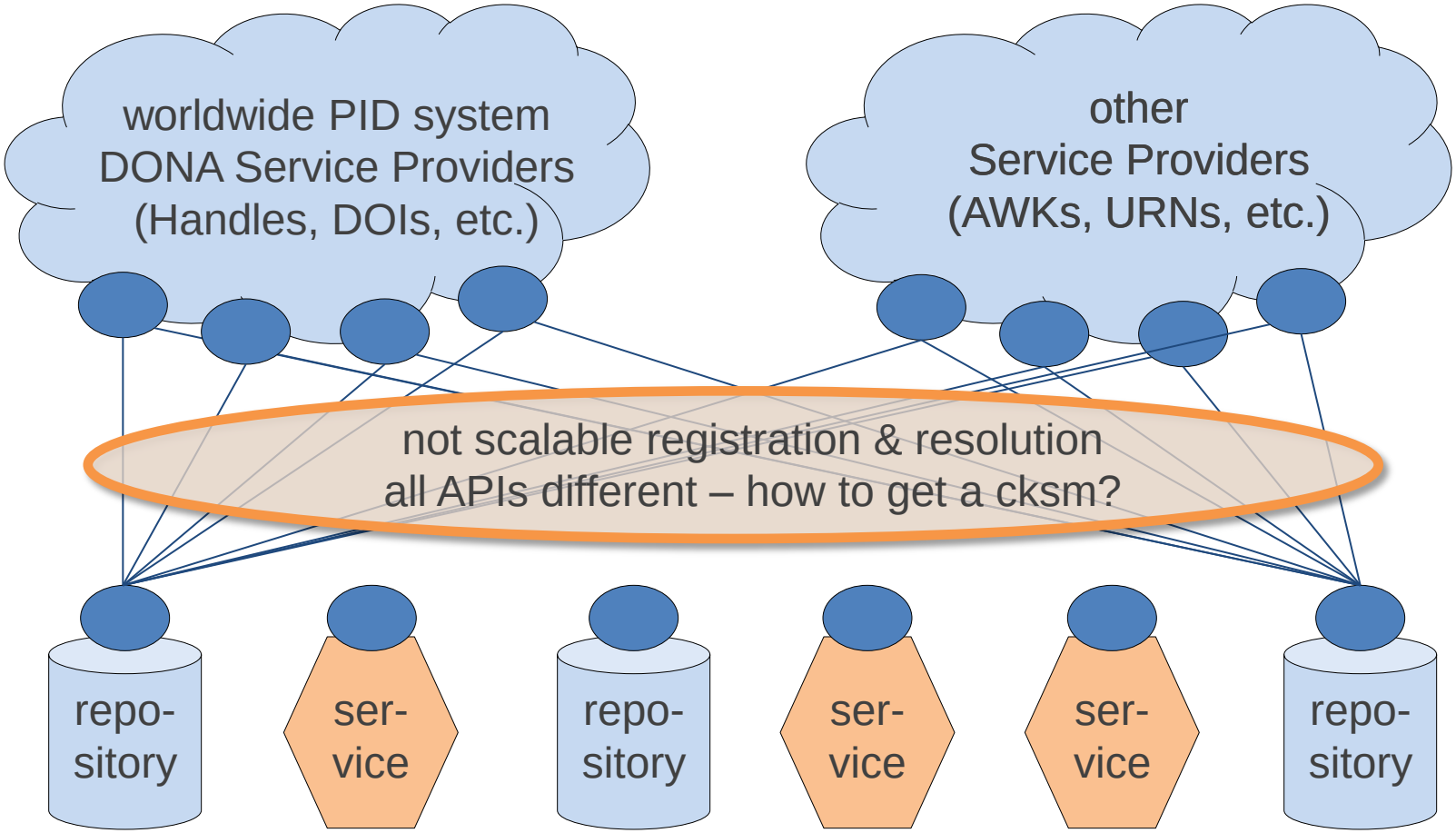
- let's come to a common object model with PIDs as anchors – like IP numbers in networks
- PID and MD records store properties of objects and collections, policy rules manipulate properties
- EUDAT is a domain of registered data objects

work in RDA

- Data Foundation and Terminology
- PID Information Type Harmonization
- Data Type Registry
- UPC for Data
- Practical Policy
- Metadata Normalization
- Contextual Metadata
- Pub/Data Citation/Linking
- Scientists Engagement
- Community Capability Model
- Preservation Infrastructure
- Legal Interoperability
- Repository Audit and Certification
- Marine Data Harmonization
- Defining Urban Data Exchange for Science

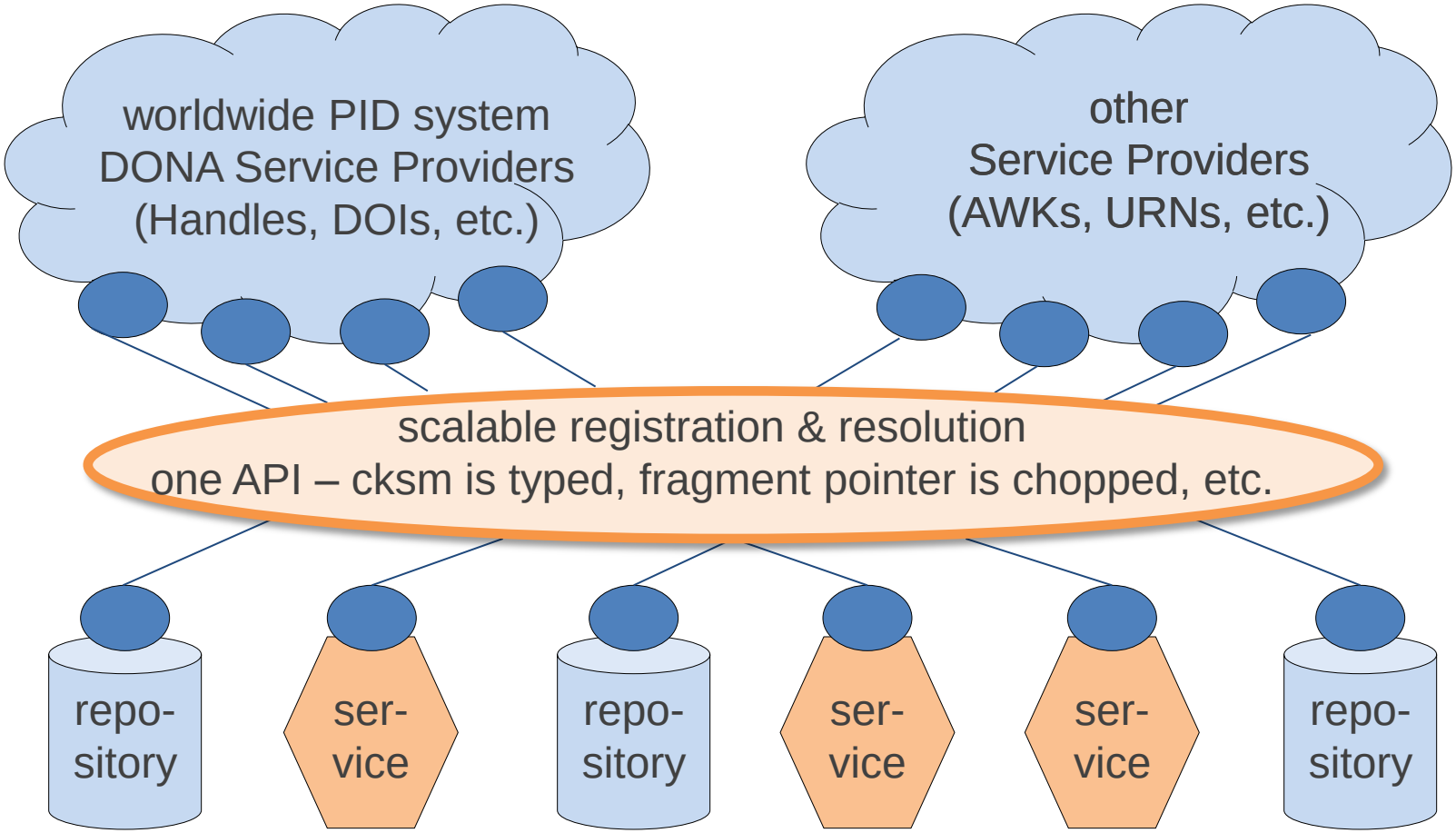
2. RDA Plenary, 16-18 September 2013, Washington, US
3. RDA Plenary, 26-28 March 2014, Dublin, AU/Europe
4. RDA Plenary, ? October 2014, ?, Europe (bid is open)
5. RDA Plenary, ? March 2015, ?, US (bid is open)

example: PID Information Types



chairs: Tobias Weigelt (DKRZ), Timothy Dilauro (JHU)

example: PID Information Types



EUDAT/RDA – lessons learned?

- some RDA lessons
  - too early really
  - but “domain of registered data” and “data fabric” will be essential
  - some enthusiastic people – but little time left for RDA work
  - much top-down activity (EC, NSF, AU ministry, etc.)
  - many new group initiatives – will they survive?

  - give one more year and we will see
  - to me it is THE chance to make progress, depends on all of us

http://rd-alliance.org

Thanks for the attention.

Questions?