Humanities and social sciences join forces to link with e-Infrastructures

Peter Wittenburg
The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
CLARIN Research Infrastructure

where it all started …

Given our human capabilities to change our conditions of life in all aspects we cannot simply continue with the old paradigms in research.

John Taylor:

"e-Science is about global collaboration in key areas of science and the next generation of infrastructures that will enable it."

As for building new fast trains we need new tracks, new signaling options, etc.

were SSH considered …

SSH ESFRI Roadmap Projects

CLARIN is where my group is engaged in

all very much distributed domains

work on structuring domains with strong centers as backbone

these big challenges are known …

- how to come to a stable climate in which next generation can survive?
- how to solve our eminent energy problems given the enormous effects on the environment?
- how to maintain a stable health given all environmental changes and influences?
- etc.

why do we ignore these …

- how to maintain stable societies given the globalization affecting our cultures and languages?
- how to maintain stable minds given decreasing quality of cultural debates and increasing technological innovation?

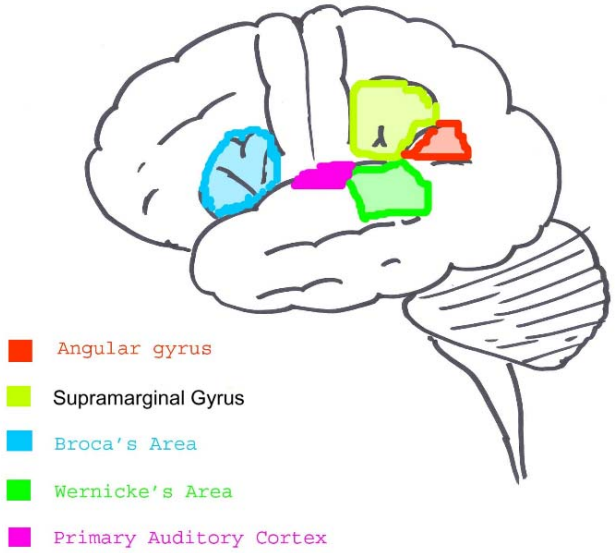
- we do so as if we can manage - but ...

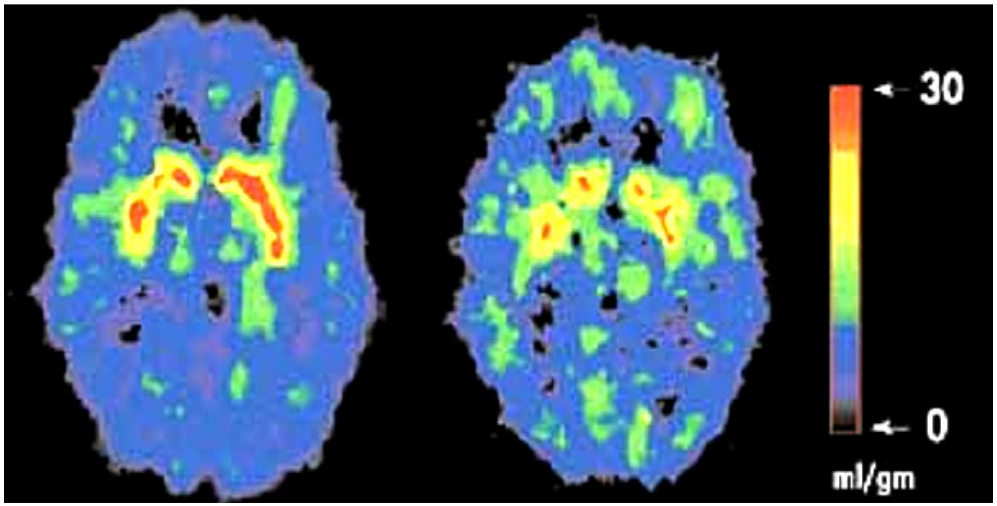
# of course the “small” challenges …

major scientific break-throughs were achieved by the small groups driven by scientific curiosity

- so let’s not forget these “small challenges”
- in our domain of languages and mind:
  - how does our human brain/mind processes language?

# of course the “small” challenges …





we gather many facts about behavior of mental language machine and use expensive equipment to collect colorful brain-images

but do we understand how it works?

# of course the “small” challenges …

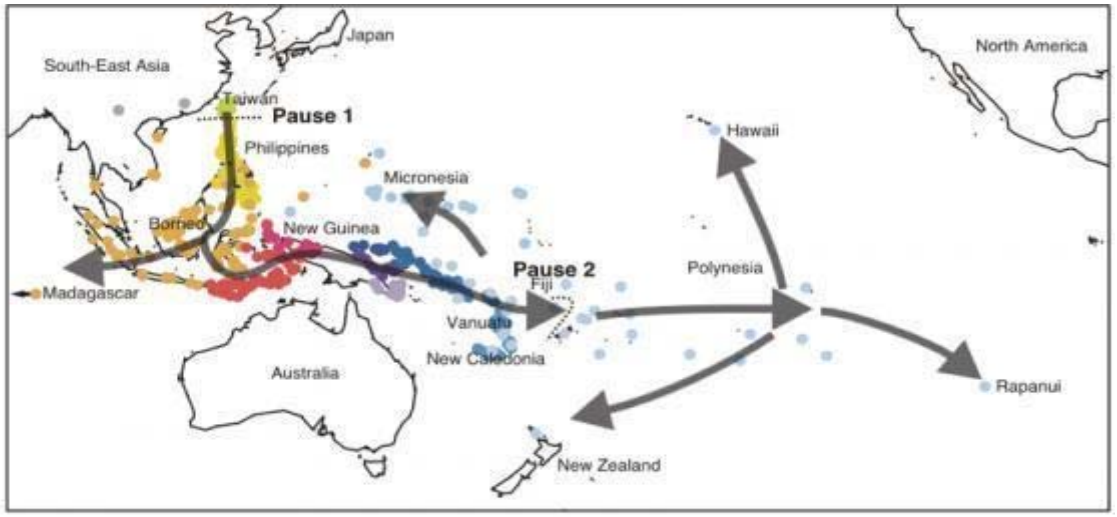
major scientific break-throughs were achieved by the small groups driven by scientific curiosity

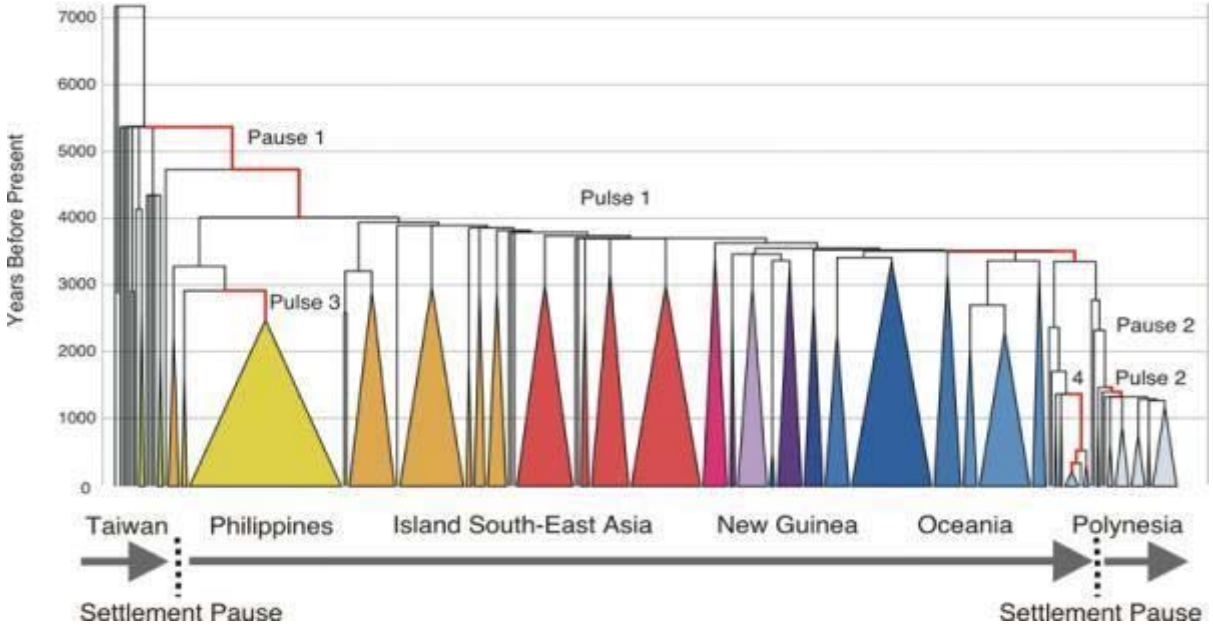
•so let’s not forget these “small challenges”

•in our domain of languages and mind:

- how does our human brain/mind process language?
- how have the 6500 languages still spoken developed over time?

# of course the “small” challenges …





according to this dependency tree Taiwan is at the root of Polynesian languages - it’s about objective identity and mechanisms of how languages can evolve

calculated on large feature matrices extracted from lots of data using phylogenetic algorithms

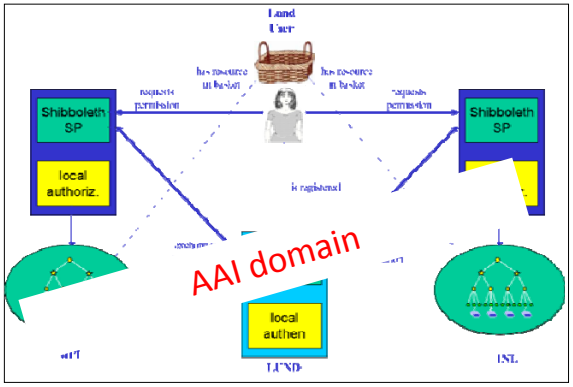
# of course the “small” challenges …

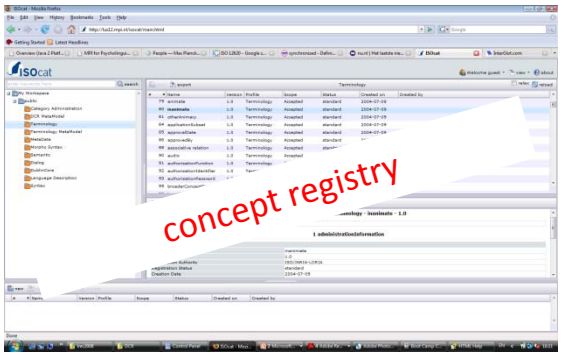
of course there are many more of these challenges in the domains of ...

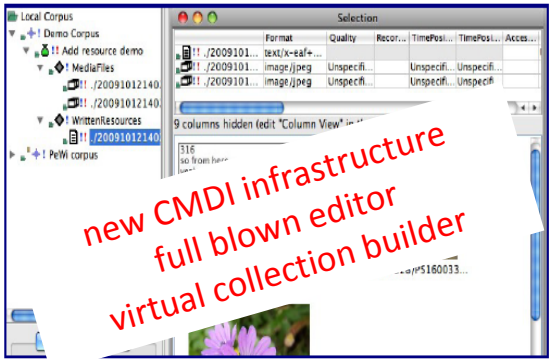
- DARIAH
- CESSDA
- SHARE
- ESS
- CLARIN

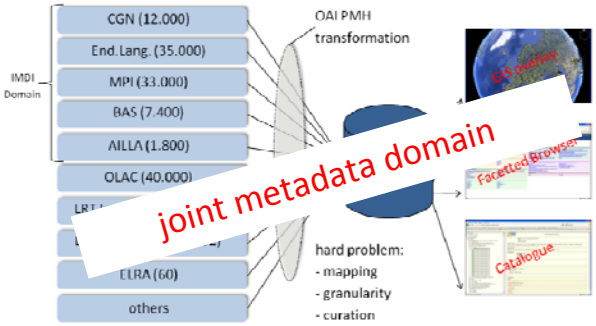
all working on integration and interoperability …

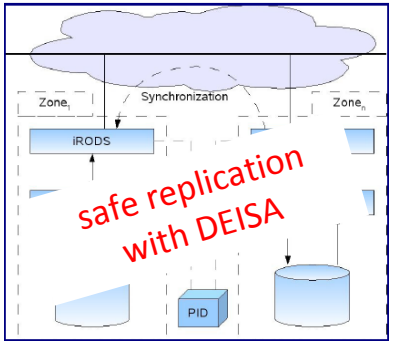
a quick look into CLARIN as one example











DASISH cluster proposal

currently all preparing ERICs (SHARE has one already)

CLARIN
Common Language Resources and Technology Infrastructure

DARIAH

cessda

European Social Survey

SHARE
Survey of Health, Ageing and Retirement in Europe
50+ in Europe

what did we suggest to do after several intensive meetings

seems that we will get funded ☺

the DASISH cluster plans …

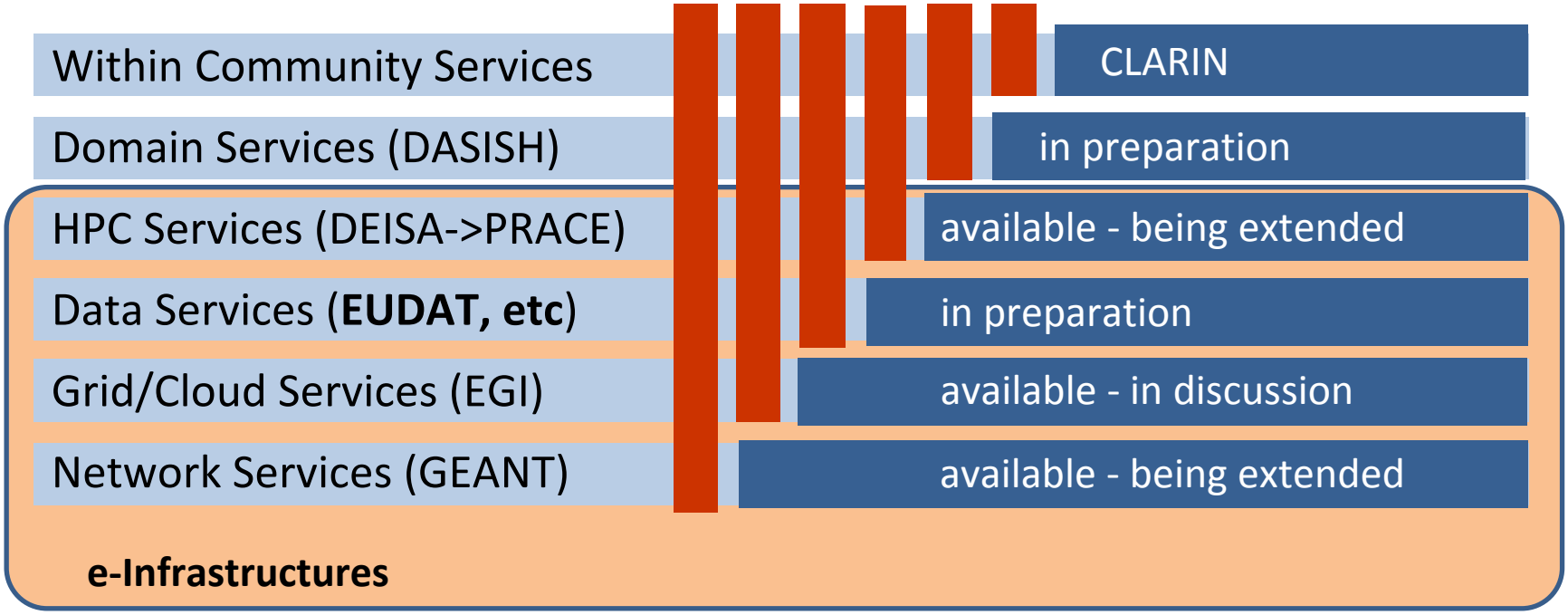
which topics do we all share?

- How to achieve integration and interoperability beyond the borders of the individual projects given different data organizations & languages?
- How can we manage to preserve our cultural and scientific memory and keep the records of science accessible?
- How to come from a down-load first scenario to a truly web-based usage scenario to optimally access and enrich the stored data to tackle the many big and small research challenges?
- How can we improve the quality of our data to enable advanced and cross-disciplinary access and enrichment operations?
- How to simplify access conditions for researchers?
- How to establish trust of SSH researchers in the infrastructure services?

- can we build on solutions or buy services from others?

DASISH as part of an eco-System …

- 48+ RI to solve the same basic tasks again and again?

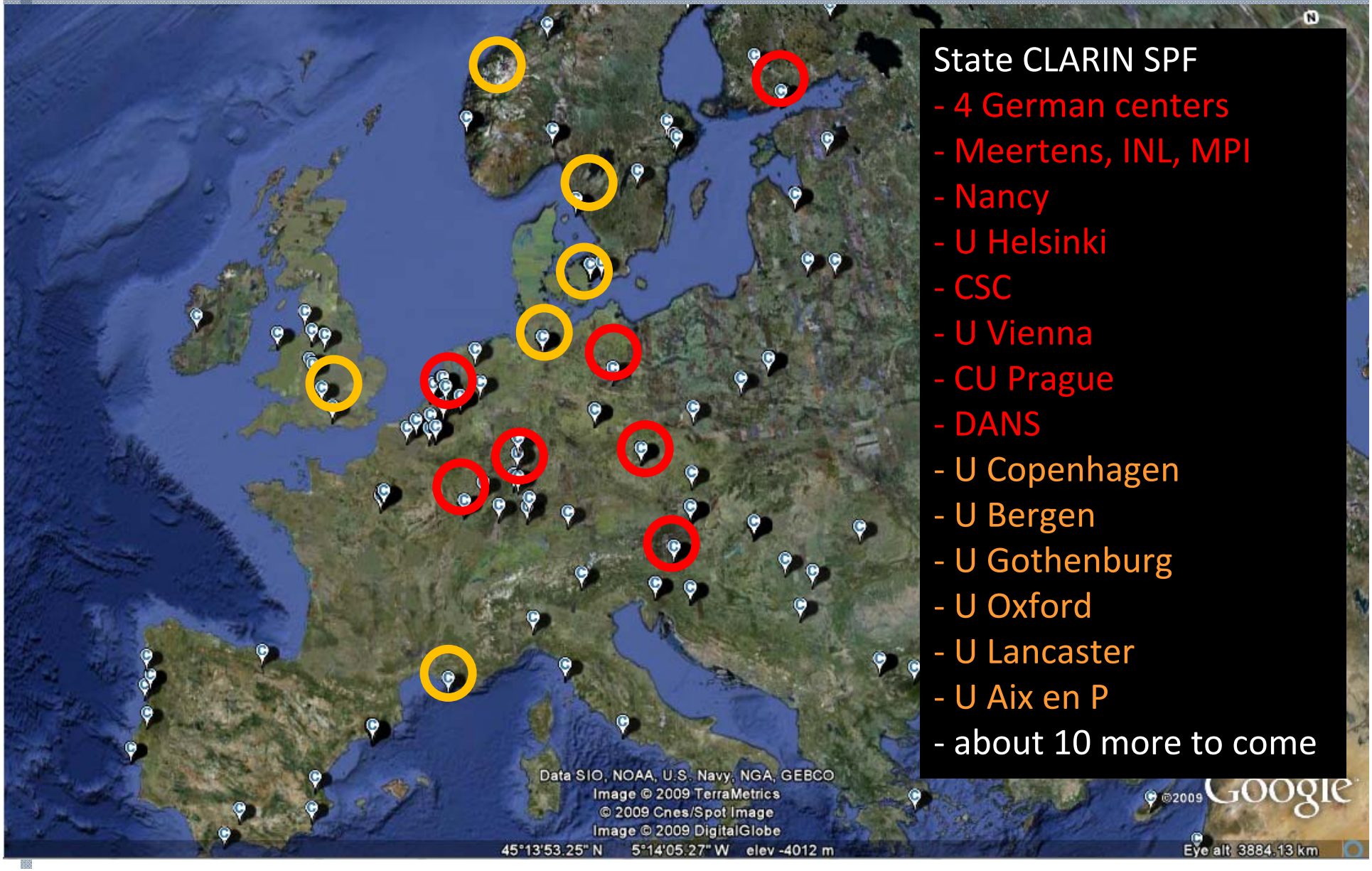


- doing it all yourself would not be efficient
  - but CERN, ESA, etc. will look different at this scenario
- need to build on common services where possible
  - but finding a proper mutual understanding is not simple

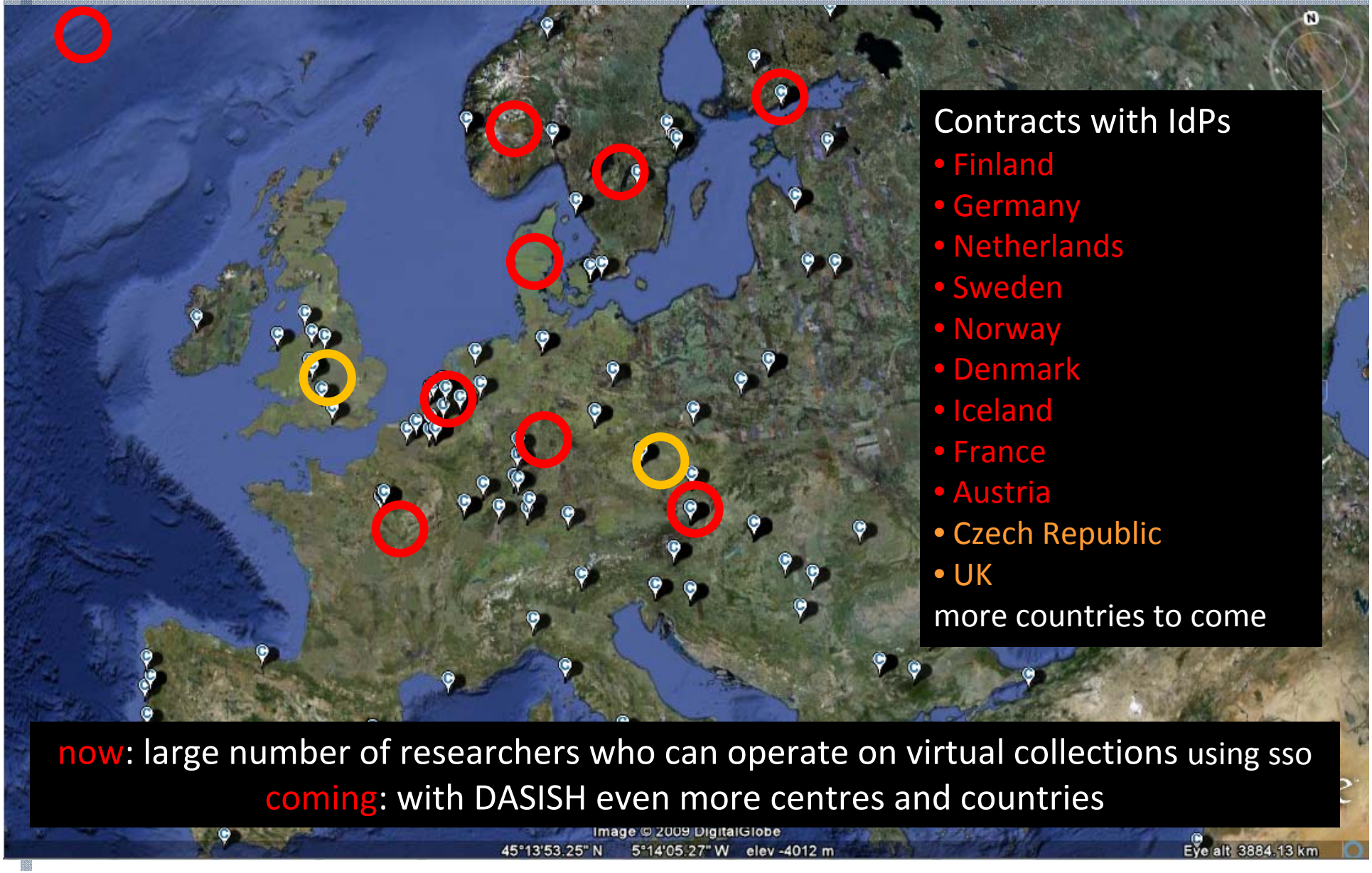
eco-System examples…

1. federation building
2. data services
3. web-based usage scenario
4. knowledge dissemination

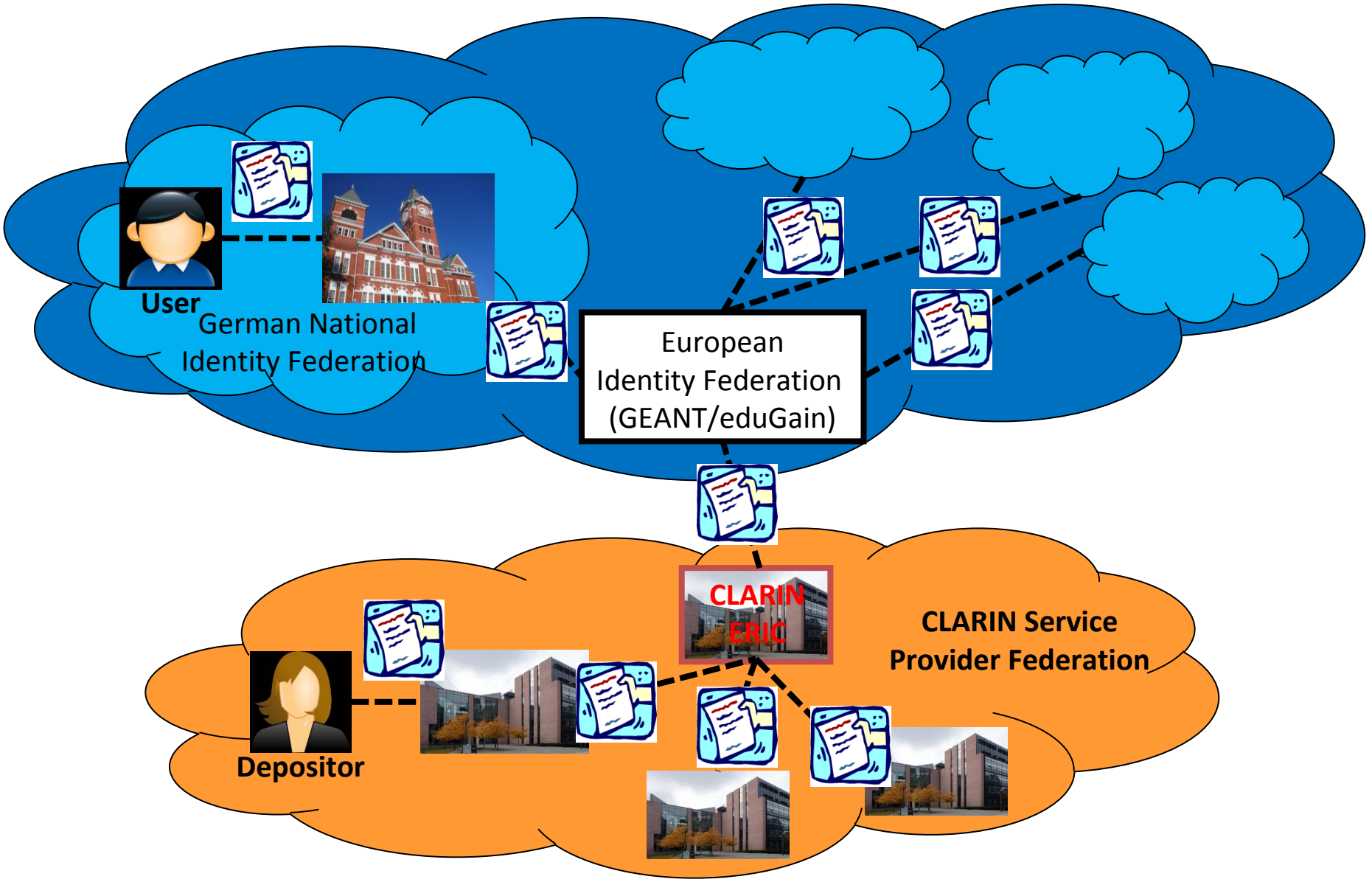
Example 1: trust federation



Example 1: trust federation

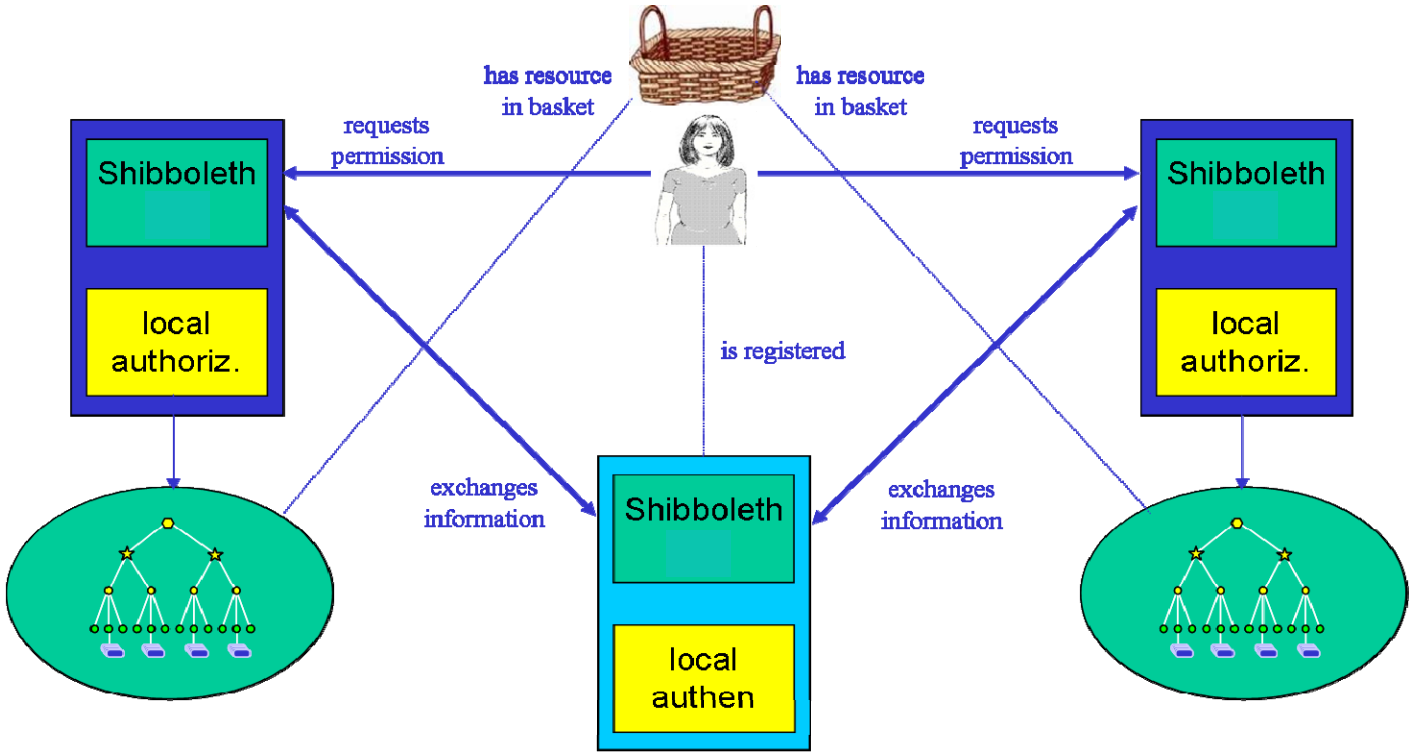


Example 1: trust federation



Federation essentials for SSH

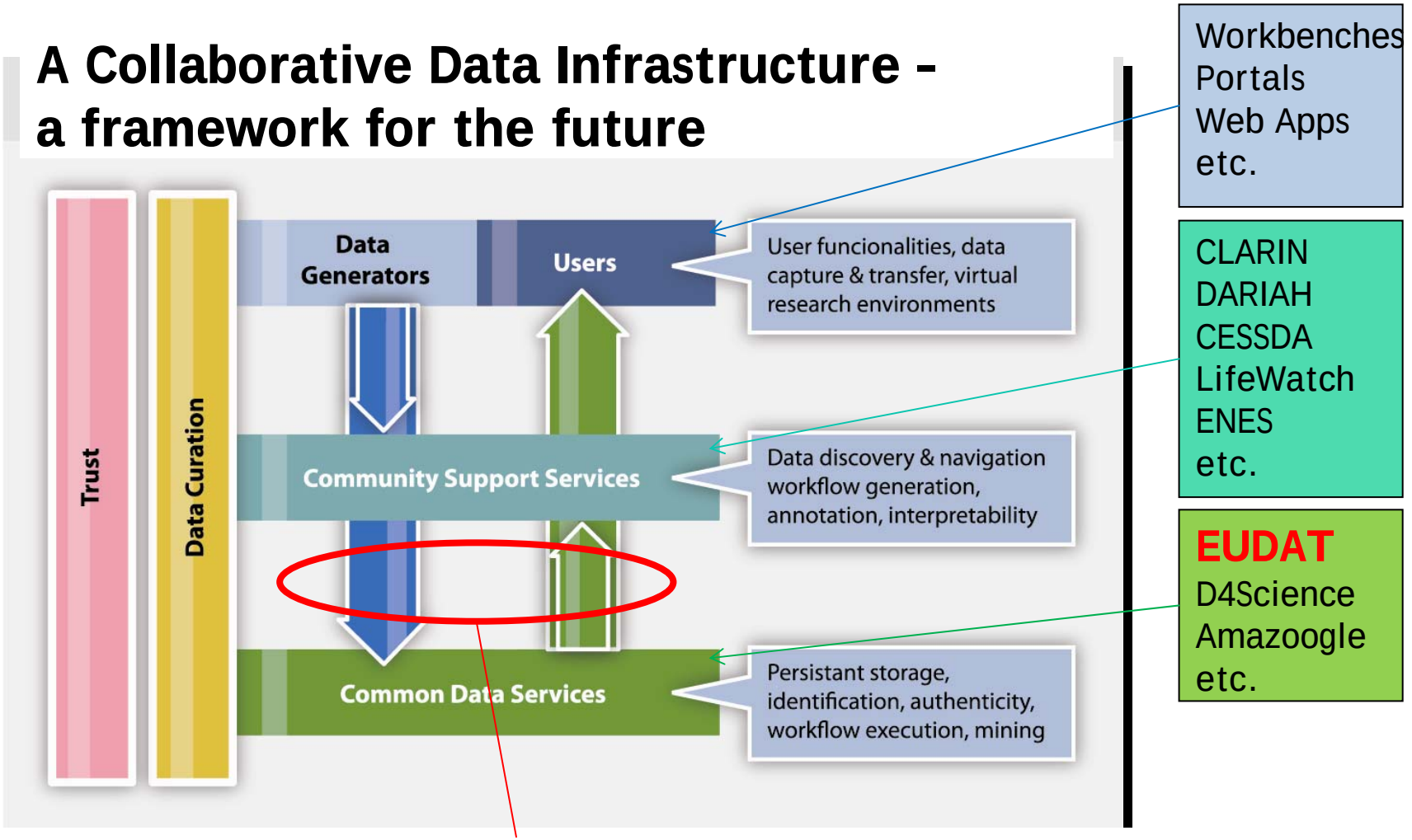
- it must be transparent since otherwise SSH researchers will not use it!
- it's not a matter of use cases - it's getting the whole community on board
- single identity granted by home institution as basis
- single sign-on to allow virtual collection building etc.



Example 2: Data Preservation

Report der High Level Expert Group on Scientific Data
from 6. October 2010

Collaborative Data Infrastructure

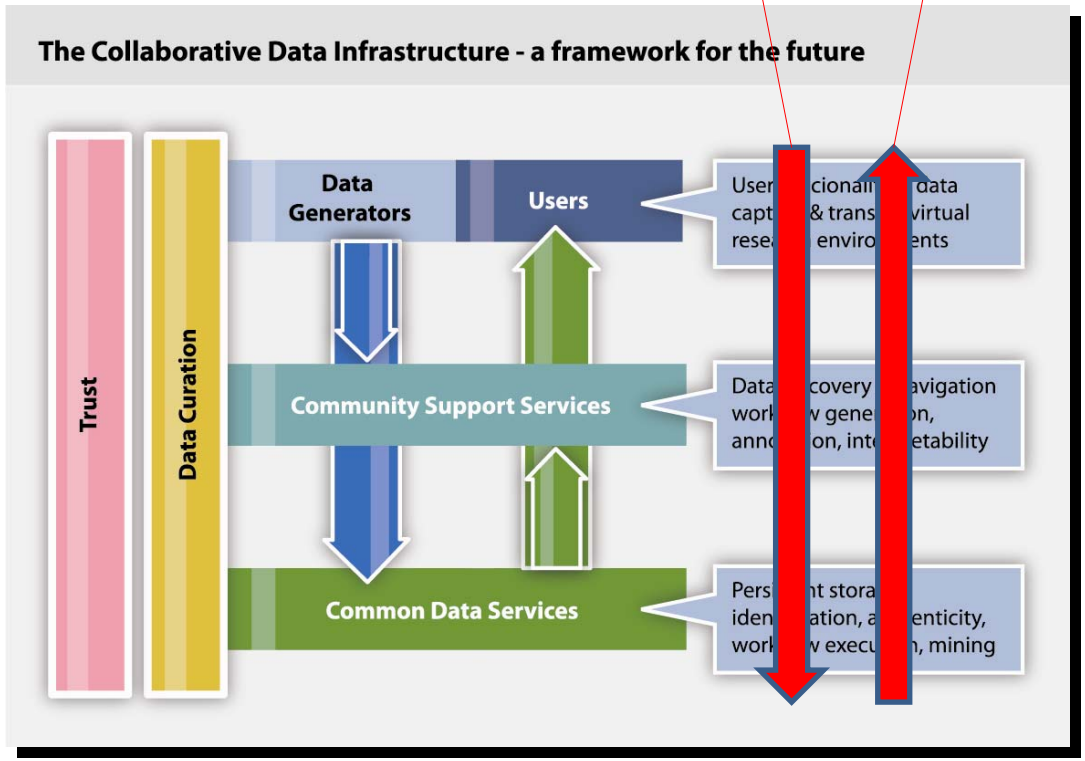


SSH communities have different data organization solutions
i.e. what is the right, abstract interface?

How to organize CDI

“bottom up” from Communities

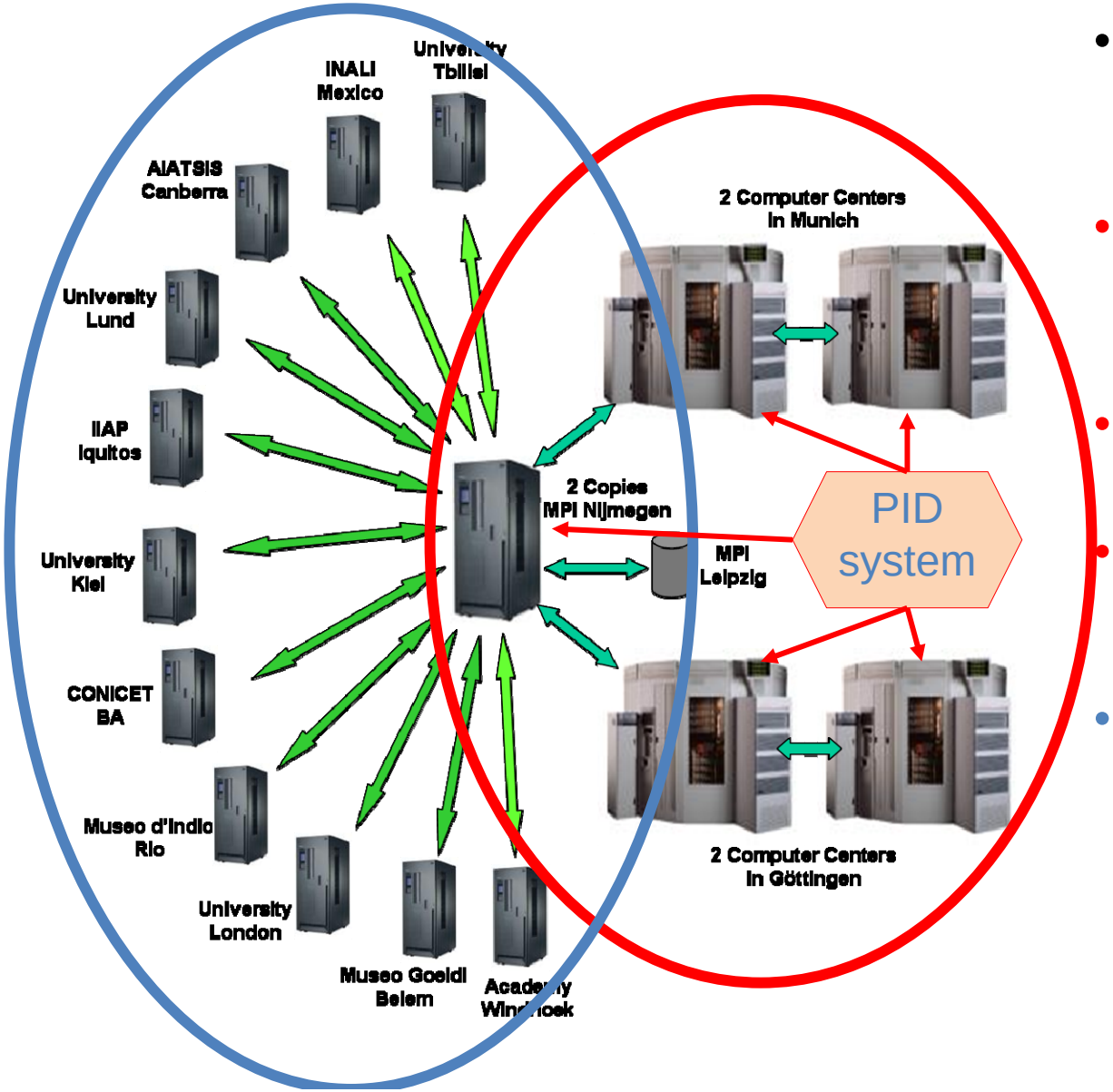
“top down” from IT



- **need to start with bottom-up approach since communities will/can not change**
- **there will be phases where IT experts need to generalize**

- **in recent years much e-Infra work driven by IT**
- **currently we see a move towards bottom up i.e. different languages, different solutions, etc.**

existing island solutions but …



- since 2004 a LTP strategy in Max Planck Society
- yet no systematic European solution !!

  80% data endangered
- yet no safe and rule-based replication !!
- using EPIC services and iRODs
- in addition 13 regional archives worldwide to help human heritage to survive

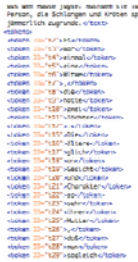
  (10 requests)

Data preservation issue …

- we need a YouTube for SSH data
  - with preservation option (cannot expect curation service)
  - with metadata support and proper IPR solution
- we need a common preservation layer
  - most community centers good for short/mid-term offers
- EUDAT is going to provide both services

Example 3: Distributed Workflows

Stuttgart



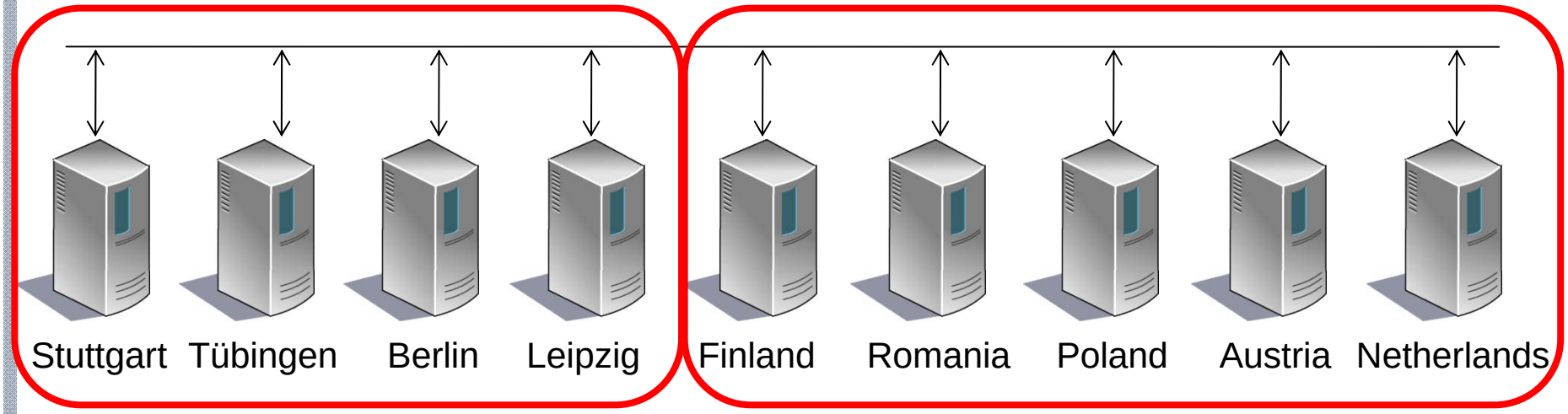
Standard-conformant
Text Corpus Encoding

Tübingen

Web 2.0 Application for
Tool Chaining
and Execution

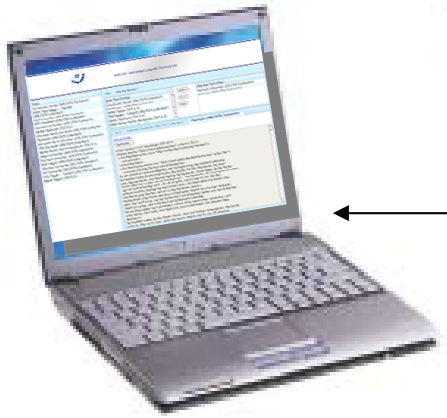
Leipzig

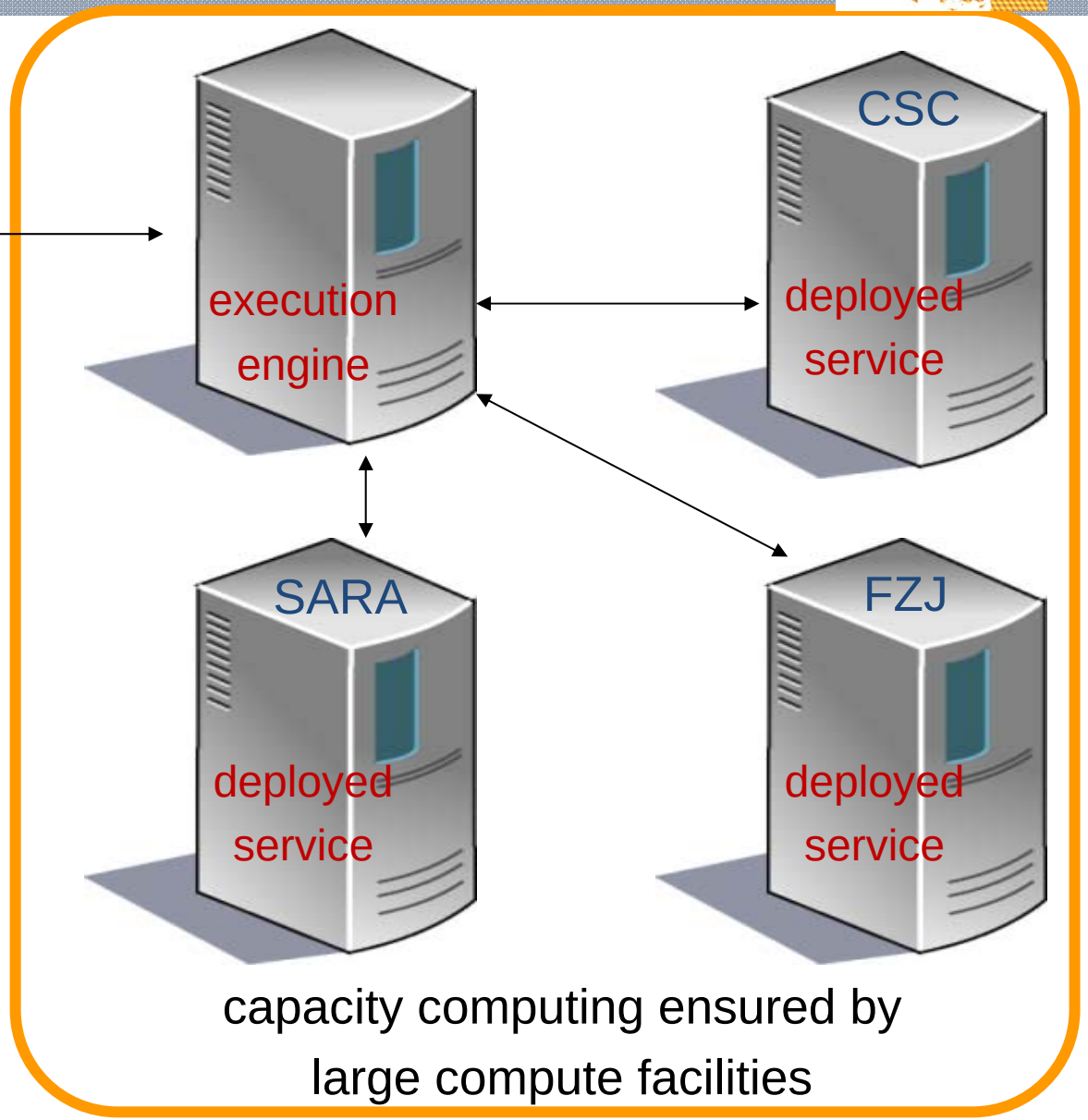
Repository



- complete chaining system in operation - running on departmental servers
- needs to be available for all interested researchers in Europe (+ beyond)

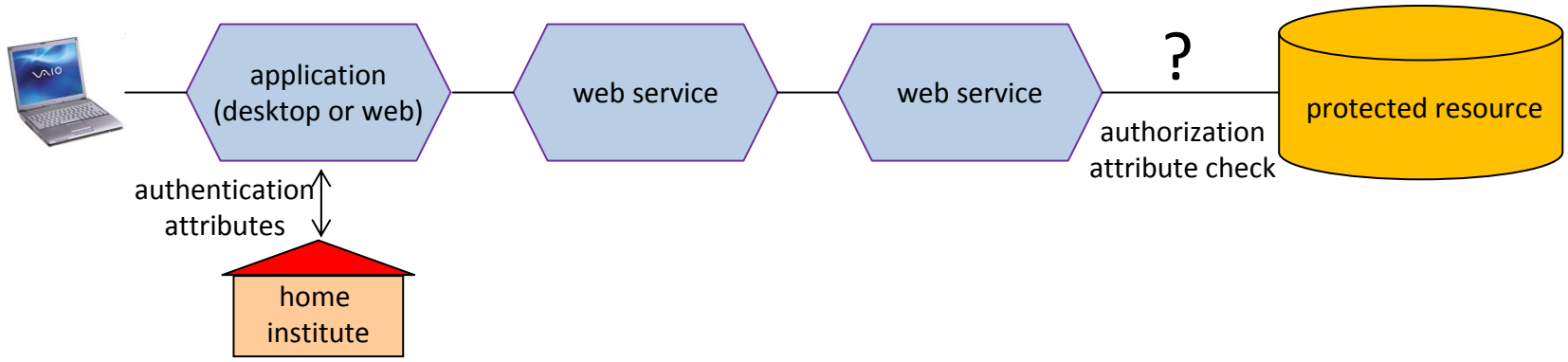
Example 3: Distributed Workflows





Example 3: problem to solve …

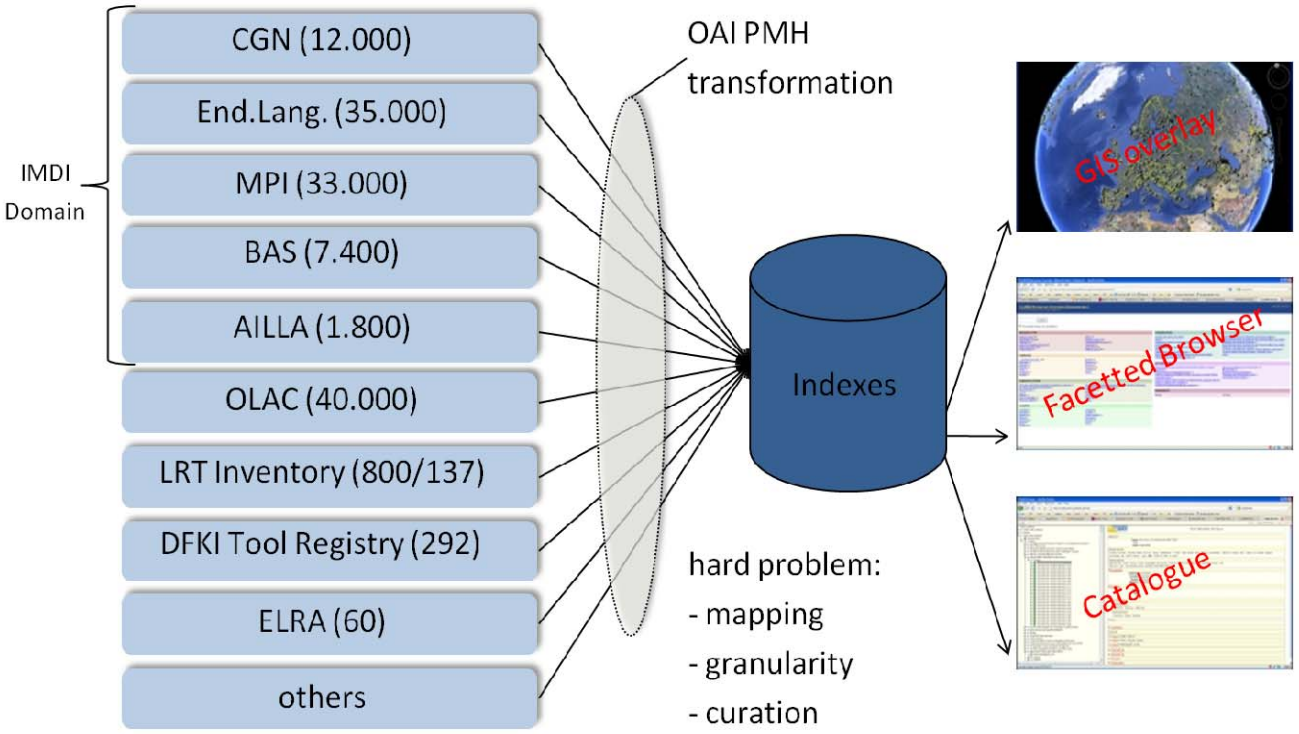
- yet no robust solution for attribute delegation for web services



- have joint projects with Dutch Grid colleagues, but must be a service for everyone in Europe (and beyond)

Example 4: service dissemination …

Virtual Language Observatory



> 100.000 resources / > 500 tools/services in “offer” - all open metadata

can EGI have a role here (MoU)?

- all metadata is open
- thus resources and tools are “published”
- want to move towards automatic profile matching
- center need to give advice
- in DASISH cross-training etc.

interactions on the way …

- in collaboration with eduGain/GEANT - yet no common solution for AAI
- interacting with EGI - yet not a clear picture
- interacting with EUDAT with clear expectations

- discussing issues, but bottom up principle is essential
- working on an eco-system sounds convincing - but will take time to come to a seamless and cost efficient setup

- services must be lean to be affordable
- need competition built in
- what are the costs - who will pay what?

Thanks for
your attention.

Note to Cloud

- “Cloud” now used by everyone for everything?
  - Cloud is a technology - it does not solve all our problems
    - it does not solve long-term curation/preservation
  - it allows to store much data and protect access from outside
    - one domain of authority simplifies
    - but that’s not the only issue
    - issue for researchers is internal data access and flow control
  - it allows easy service deployment
  - it caters for scalable capacity computing
- as Community we don’t care too much which technology is used
  - robustness, persistence
  - decent level of security
  - not forcing us to change our data organization