MAX-PLANCK-GESELLSCHAFT

Key Elements of Global Data Infrastructures

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

EUDAT

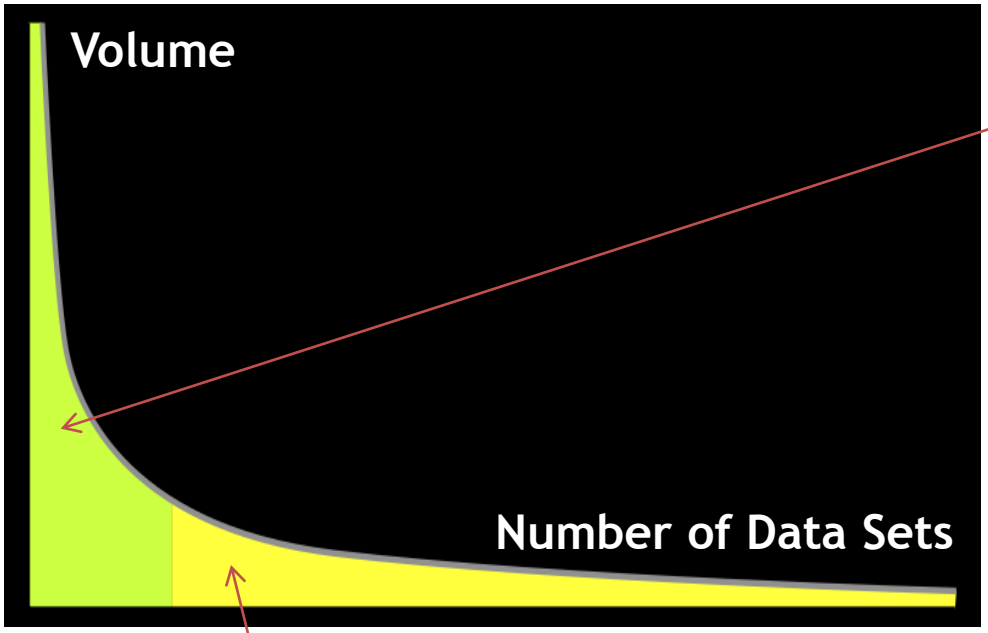
CLARIN
Common Language Resources and Technology Infrastructure

Big Questions in Complex Data Domain

- How to guarantee easy and persistent access to globally available data objects and collections?
- How to remove interoperability barriers to easily analyse large distributed data sets?

- many aspects and answers
  - focus on three different key elements for Data Infrastructures (DI)

Remind the Long Data Tail



Focus on Big Data
- in general raw data
- in general regular structure

Data Intensive Science
domain of numbers
- find patterns across globally spread collections
- fit parameters using big data

Focus also on Small Data
- often covers domain knowledge
- much more heterogeneous
- in general special structures and difficult semantics

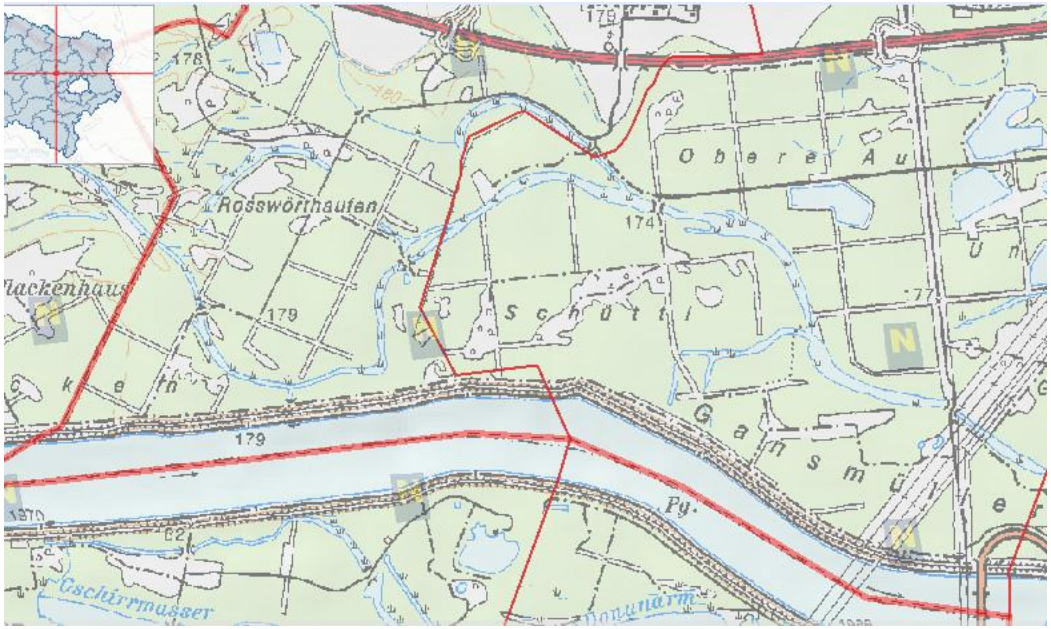
Smart Information Science
domain of symbols just one example
- semantically join kindred collections
- exploit using semantic knowledge

Data Infrastructures need Registries

"modern" societies have (cadastral) land registries

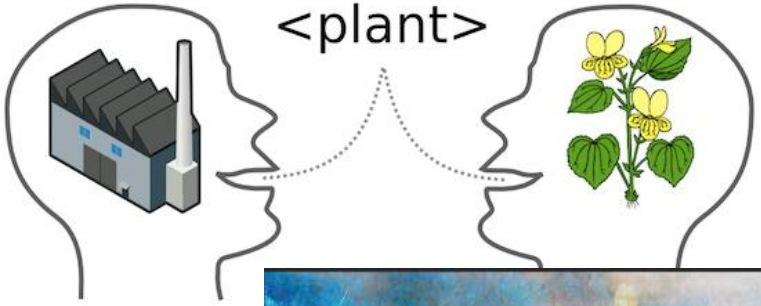
- dimensions, owner, claims, etc.
- also roads, electricity lines, etc.

- hierarchy of authorities
- ALL know how to find/read them



- functioning DI requires agreed registries of many types
  - centers/repositories, objectIDs, personIDs, etc.
  - need agreements on formats, content, APIs to support automatic access
  - need global agreements - need a big approach
  - does not make sense to support small islands
  - how can we make registry approach scale and who takes care about persistency
- we need to bundle forces => DAITF

Interoperability is essential, but ...



- Interoperability is relevant at many levels
  - integration of metadata
  - tools to automatically find information
  - execute one operation on data sets created by different researchers
  - it's all about interpreting syntax and semantics and bridging

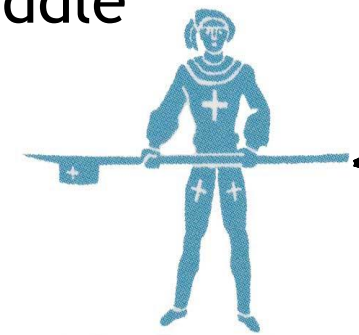
- interoperable DI requires adhering to basic IT principles
  - make your syntax/formats explicit and register them in known registries
  - make your semantics (elements, vocabularies) explicit and register them
  - use persistent identifiers for ALL references
  - does not make sense to support small islands but need to accept different models

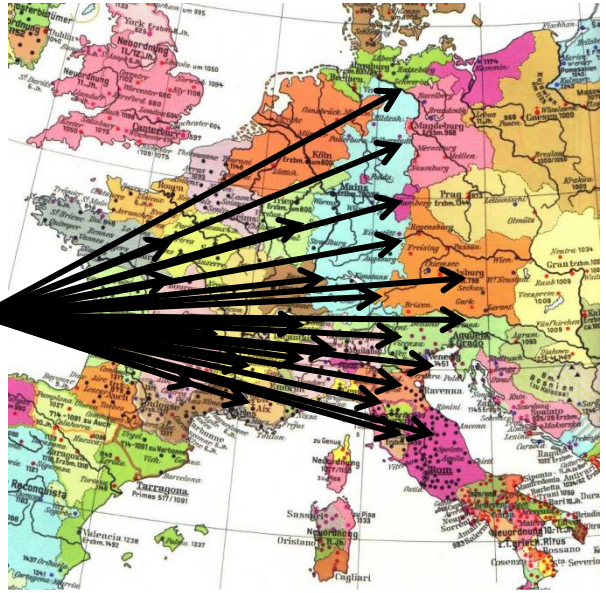
- we need to bundle forces => DAITF

Smooth Distributed Authentication a Must

are currently in the dark middle ages

- have many identities ☹
- IdPs are small kingdoms
- no trust in/for academia
- “inefficient” practices come into place





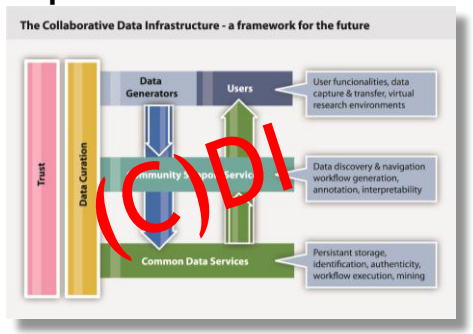
- efficient DI requires smooth distributed authentication
  - need simple mechanisms supporting single identity and single sign-on
  - need trust in academia to exchange attributes (Code of Conduct is promising)
  - must be worldwide since data is worldwide
  - can’t be true that all set up own user databases
- we need to bundle forces => DAITF

Bundle forces => DAITF

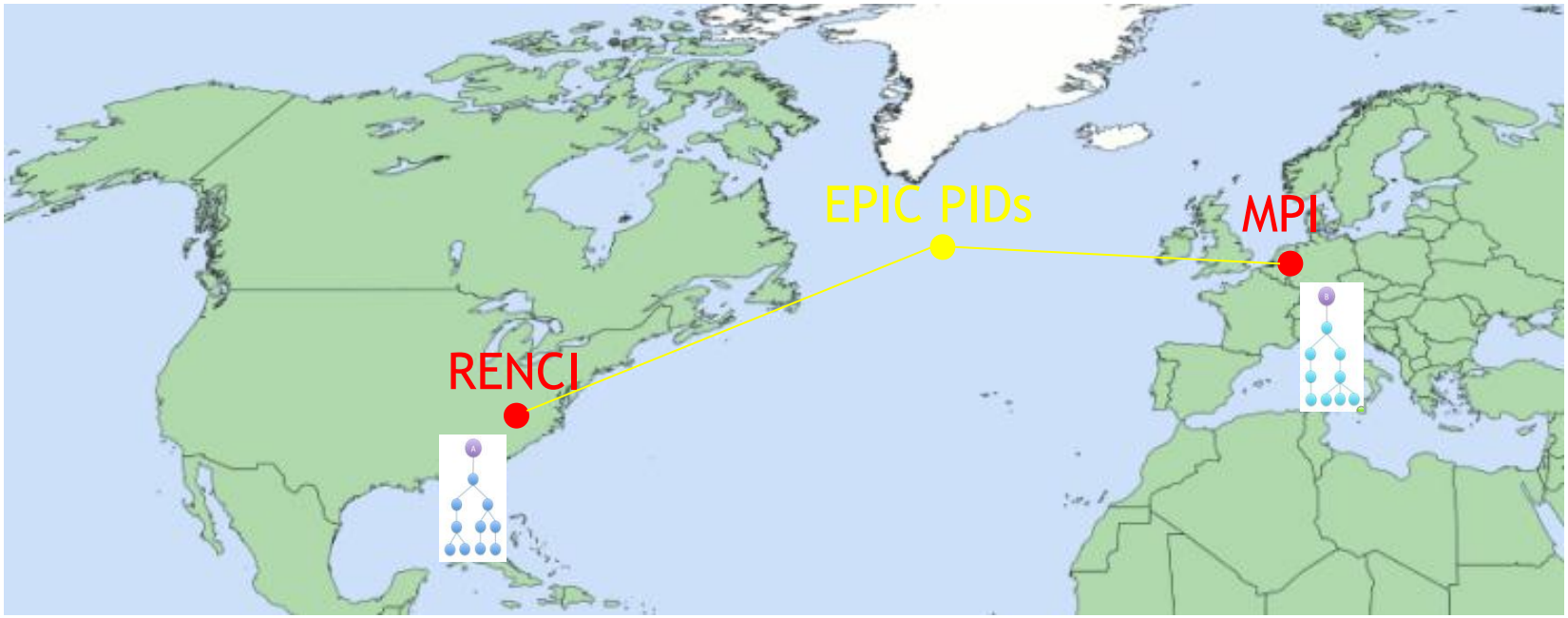
- Data Access and Interoperability Task Force
  - sounds like IETF which is not per accident
  - most agree it needs to be a grass-roots based process
  - how to organize global interaction and harmonization process
  - it is embedded in many existing activities
  - do we need something separate - will we join?

- EC is ready to fund this activity
  actually EUDAT and OpenAIRE reserved some funds already
  iCORDI is committed to push this ahead in coming years
- US seems to be ready to support this activity
- various other countries sent people to a pre ICRI workshop

- improving integration and interoperability - establishing DI - will be a stepwise process
- we are already working on that path (policy-rule based replication across Atlantic)



Policy-Rule based Replication across Atlantic



What's worth mentioning?

- all data/metadata organizations and attributes are maintained
- explicit PIDs are used to check validity
- users could immediately start working with copies
- ready for auditing