CLARIN

Common Language Resources and Technology Infrastructure

A European Research Infrastructure

CLARIN - a European Research Infrastructure

Peter Wittenburg

Max-Planck Institut für
Psycholinguistik, Nijmegen

eResearch - Infrastructures

J. Taylor

“eScience is about global collaboration in key areas of science and the next generation of infrastructures that will enable it”

Requires new persistent platforms

- to enable researchers to combine resources and tools to solve the big challenges of today (global migration, crisis of cultures and minds)
- to increase the efficiency of researchers in the many small tasks
    - 40 % of the time of "knowledge workers" is spent, to find useful material (Forrester Research)

CLARIN Goal

What:

- Offer a distributed Research Infrastructure of integrated and interoperable Language Resources and Tools that serves researchers and students in the SSH

How:

- allow the combination of existing and web-accessible digital centers hosting resources in a common federation
- offer language tools and services as distributed services with a common web interface

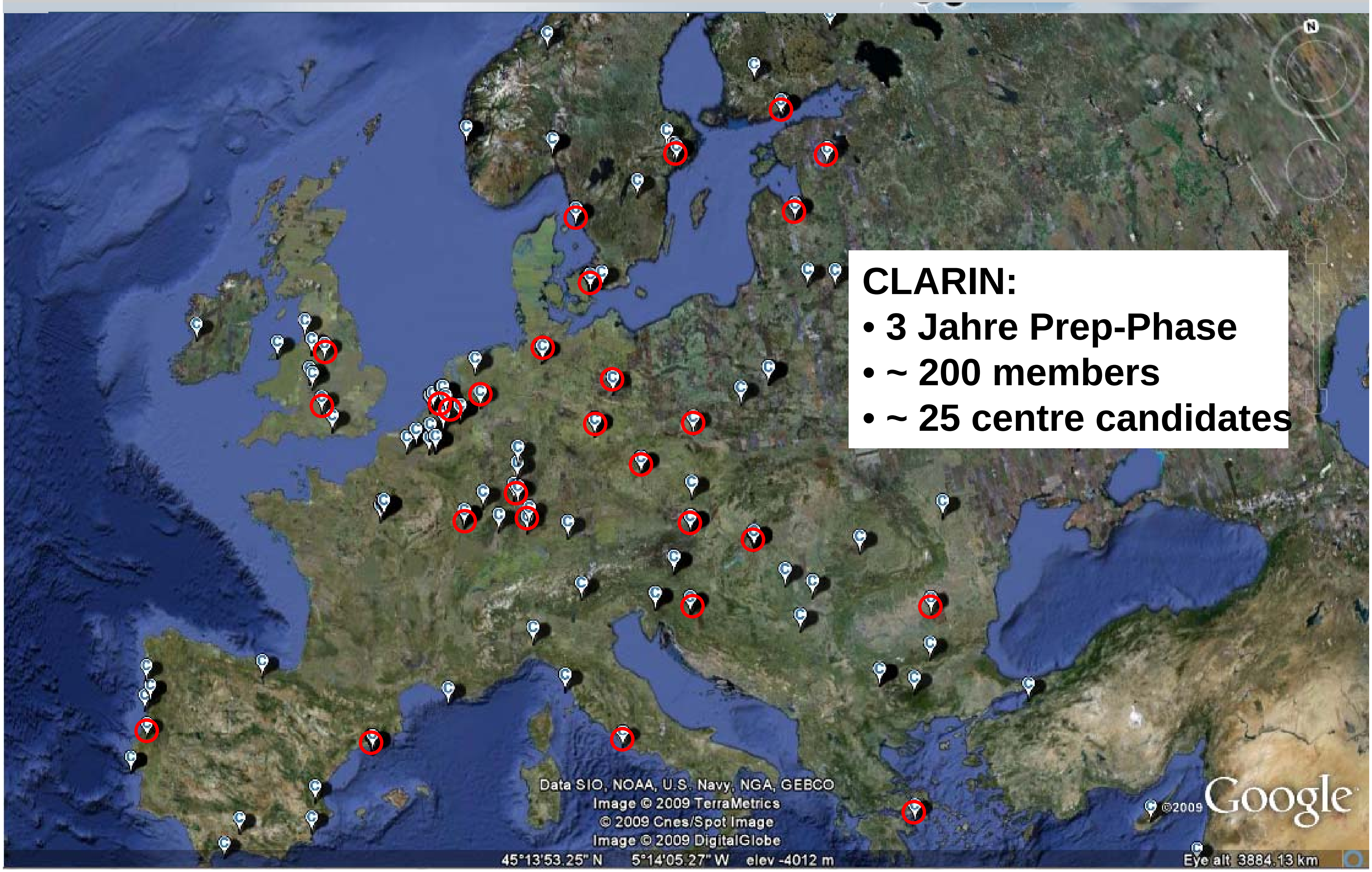
Key Application/Mission

A researcher authenticates at his own organization and creates a *virtual collection* of resources from different repositories and executing a *virtual pipeline of processes* on them.

King Arthur failed by the way
will CLARIN fail as well?

CLARIN is pan-European

The Language Archive

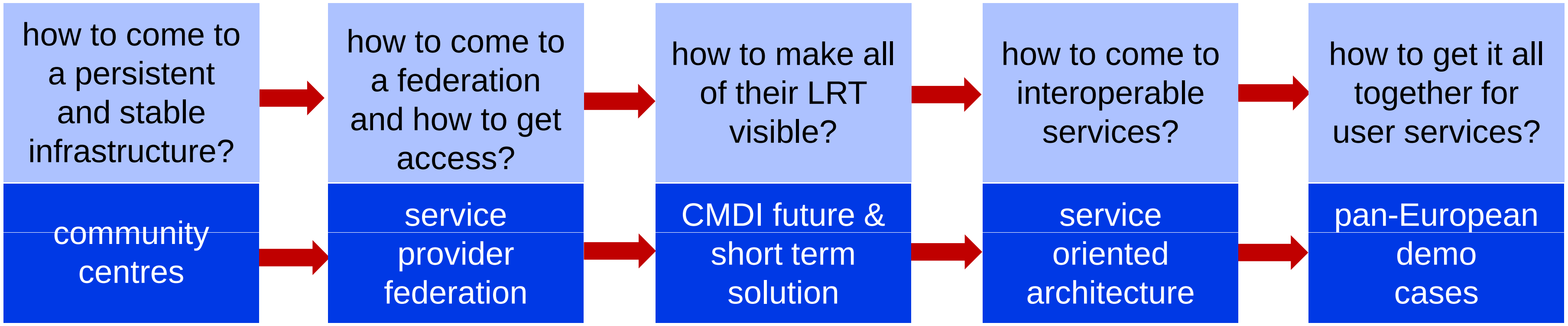

Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image © 2009 TerraMetrics
© 2009 Cnes/Spot Image
Image © 2009 DigitalGlobe
45°13'53.25" N 5°14'05.27" W elev -4012 m
©2009 Google
Eye alt 3884.13 km

CLARIN:

- **3 Jahre Prep-Phase**
- **~ 200 members**
- **~ 25 centre candidates**

CLARIN Work Dimensions

... at least IT oriented aspects



how to come to a persistent and stable infrastructure?

community centres

how to come to a federation and how to get access?

service provider federation

how to make all of their LRT visible?

CMDI future & short term solution

how to come to interoperable services?

service oriented architecture

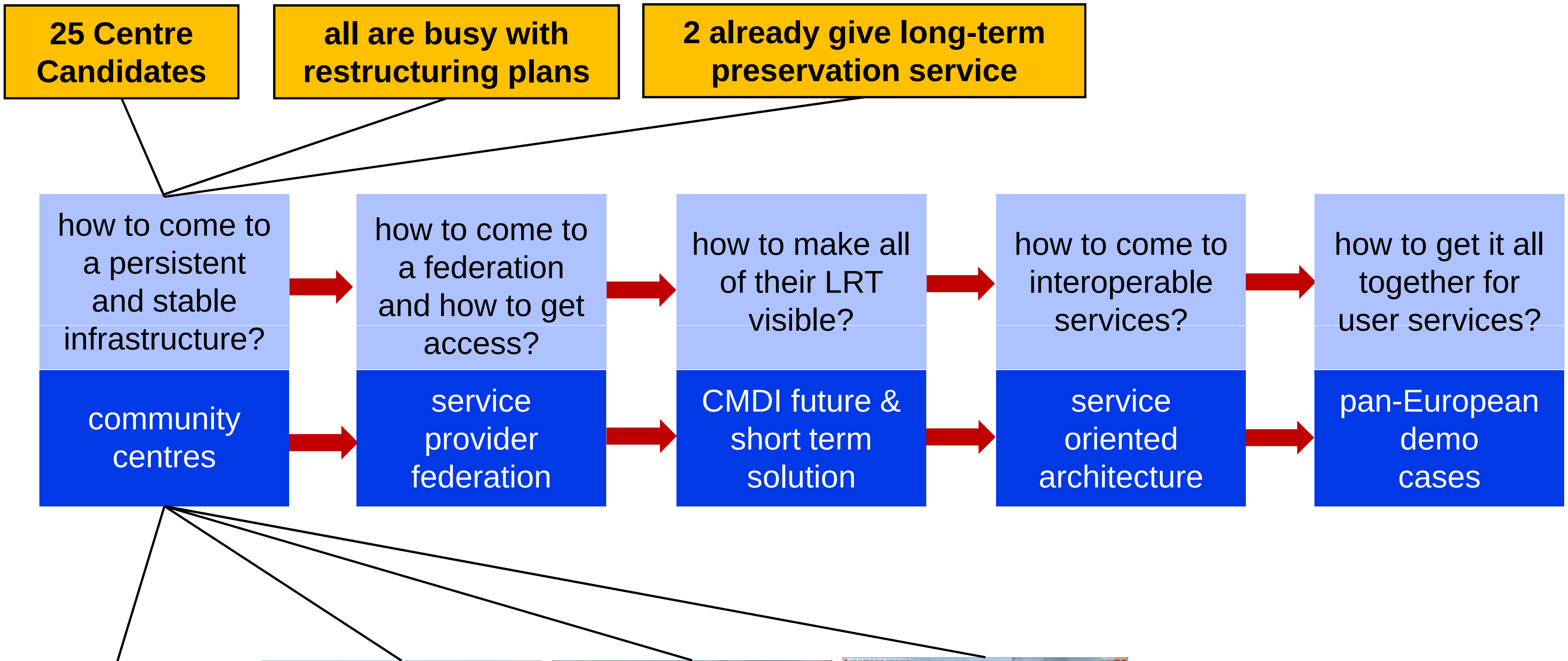
how to get it all together for user services?

pan-European demo cases

CLARIN has other very important aspects:

- Relation with SSH disciplines - mainly driven by national funds
- Education/Training, Help/Support/Advice, Dissemination
- Harmonization of licencing and Code of Conducts
- Specification of the ERIC legal framework to ensure persistency

Community Centres



25 Centre Candidates

all are busy with restructuring plans

2 already give long-term preservation service

how to come to a persistent and stable infrastructure?

how to come to a federation and how to get access?

how to make all of their LRT visible?

how to come to interoperable services?

how to get it all together for user services?

community centres

service provider federation

CMDI future & short term solution

service oriented architecture

pan-European demo cases

CLARIN Centres

Common Language Resources and Technology Infrastructure

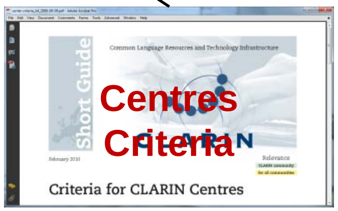
February 2009

Relevance

CLARIN community

for all communities

CLARIN Centres



Centres Criteria

Common Language Resources and Technology Infrastructure

February 2010

Relevance

CLARIN community

for all communities

Criteria for CLARIN Centres

Long-term Preservation

Common Language Resources and Technology Infrastructure

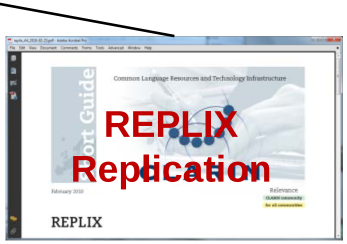
February 2009

Relevance

CLARIN community

for all communities

Long-term Preservation and Access



REPLIX Replication

Common Language Resources and Technology Infrastructure

February 2010

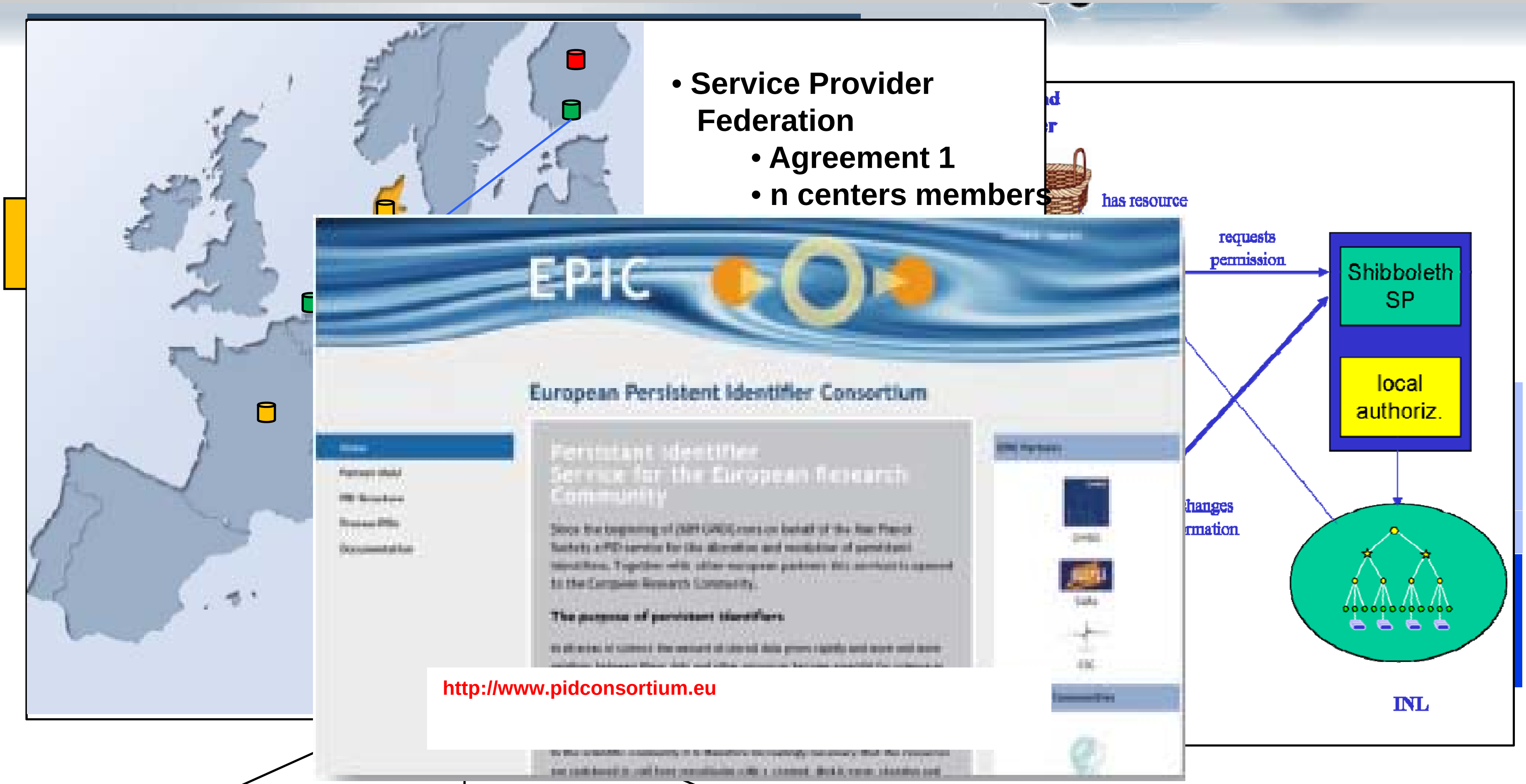
Relevance

CLARIN community

for all communities

REPLIX

Service Provider Federation



- Service Provider Federation
  - Agreement 1
  - n centers members

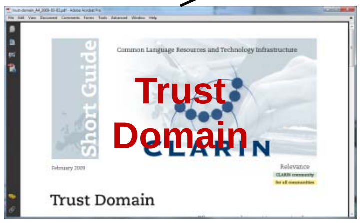
EPIC

European Persistent Identifier Consortium

Persistent Identifier Service for the European Research Community

The purpose of persistent identifiers

http://www.pidconsortium.eu



Trust Domain

Short Guide

Common Language Resources and Technology Infrastructure

CLARIN

Trust Domain

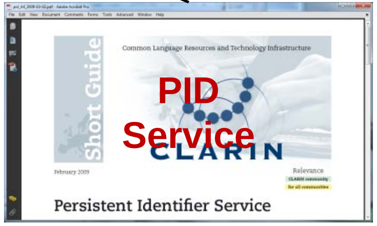
Initial Federation

Short Guide

Common Language Resources and Technology Infrastructure

CLARIN

Initial CLARIN Service Provider Federation



PID Service

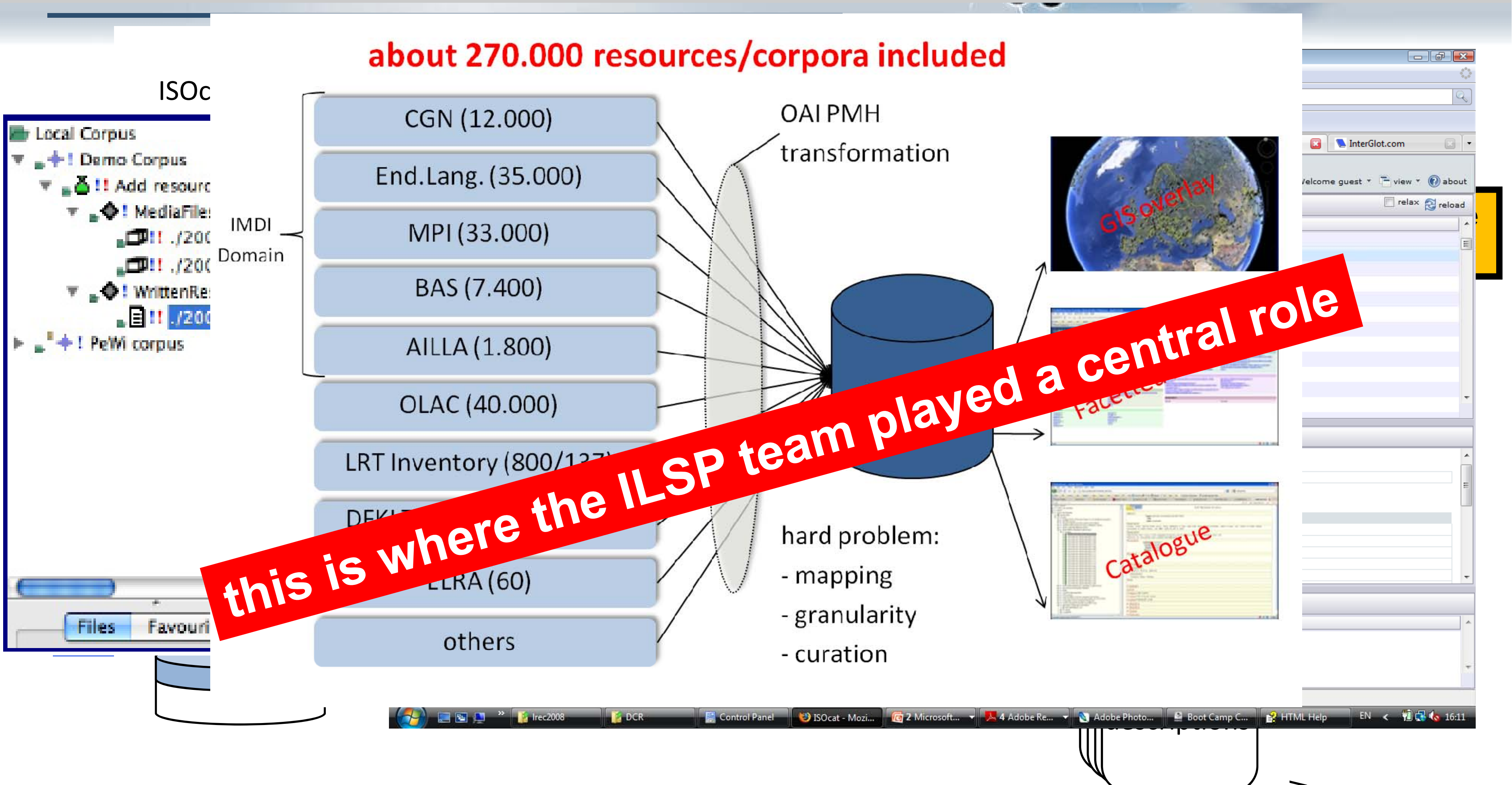
Short Guide

Common Language Resources and Technology Infrastructure

CLARIN

Persistent Identifier Service

Metadata Domain



about 270.000 resources/corpora included

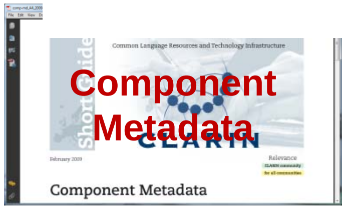
IMDI Domain

- CGN (12.000)
- End.Lang. (35.000)
- MPI (33.000)
- BAS (7.400)
- AILLA (1.800)
- OLAC (40.000)
- LRT Inventory (800/137)
- DEKL...
- ELRA (60)
- others

OAI PMH transformation

hard problem:
- mapping
- granularity
- curation

this is where the ILSP team played a central role



Component Metadata

Component Metadata

Metadata now

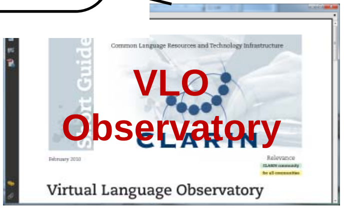
Create CLARIN Metadata Now

Virtual Collection

Virtual Collections

ISOcat Registry

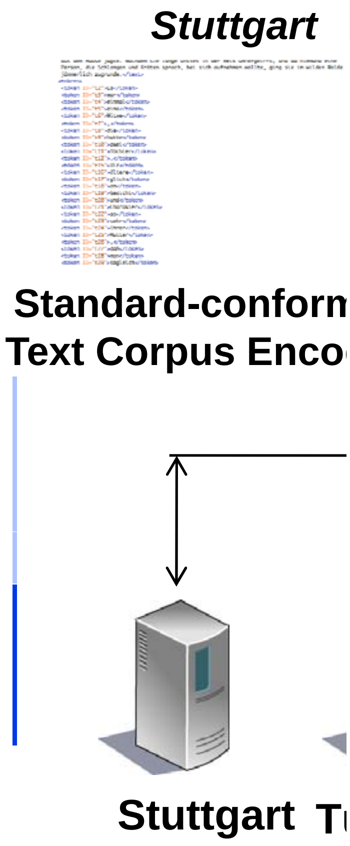
Concept Registry Service

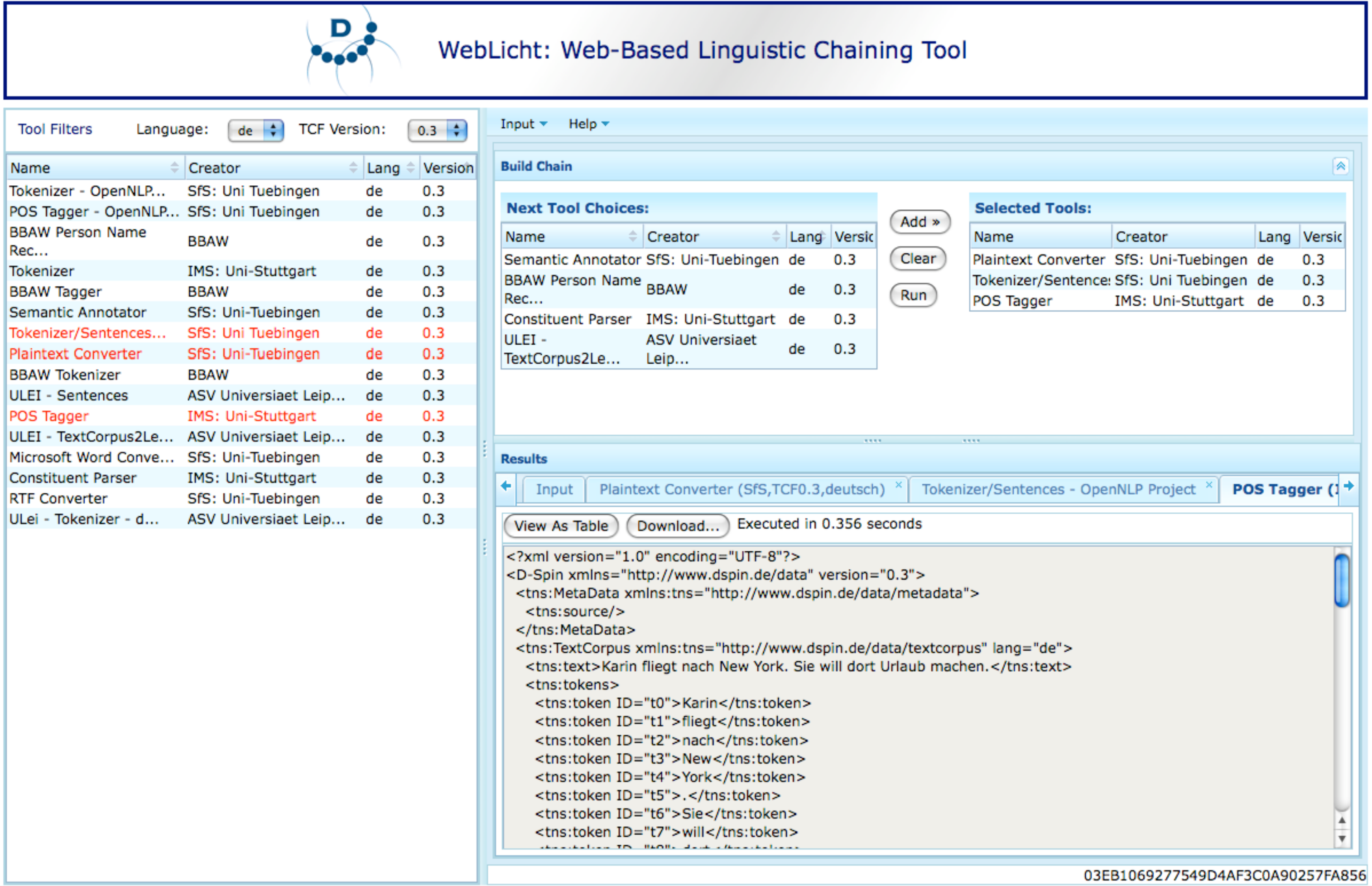


VLO Observatory

Virtual Language Observatory

Service Oriented Architecture





Language: de
TCF Version: 0.3

Name	Creator	Lang	Version
Tokenizer - OpenNLP...	SfS: Uni Tuebingen	de	0.3
POS Tagger - OpenNLP...	SfS: Uni Tuebingen	de	0.3
BBAW Person Name Rec...	BBAW	de	0.3
Tokenizer	IMS: Uni-Stuttgart	de	0.3
BBAW Tagger	BBAW	de	0.3
Semantic Annotator	SfS: Uni-Tuebingen	de	0.3
Tokenizer/Sentences...	SfS: Uni Tuebingen	de	0.3
Plaintext Converter	SfS: Uni-Tuebingen	de	0.3
BBAW Tokenizer	BBAW	de	0.3
ULEI - Sentences	ASV Universiaet Leip...	de	0.3
POS Tagger	IMS: Uni-Stuttgart	de	0.3
ULEI - TextCorpus2Le...	ASV Universiaet Leip...	de	0.3
Microsoft Word Conve...	SfS: Uni-Tuebingen	de	0.3
Constituent Parser	IMS: Uni-Stuttgart	de	0.3
RTF Converter	SfS: Uni-Tuebingen	de	0.3
ULei - Tokenizer - d...	ASV Universiaet Leip...	de	0.3

Next Tool Choices:

Name	Creator	Lang	Versic
Semantic Annotator	SfS: Uni-Tuebingen	de	0.3
BBAW Person Name Rec...	BBAW	de	0.3
Constituent Parser	IMS: Uni-Stuttgart	de	0.3
ULEI - TextCorpus2Le...	ASV Universiaet Leip...	de	0.3

Selected Tools:

Name	Creator	Lang	Versic
Plaintext Converter	SfS: Uni-Tuebingen	de	0.3
Tokenizer/Sentence:	SfS: Uni Tuebingen	de	0.3
POS Tagger	IMS: Uni-Stuttgart	de	0.3

Input | Plaintext Converter (SfS,TCF0.3,deutsch) × | Tokenizer/Sentences - OpenNLP Project × | POS Tagger (1

Executed in 0.356 seconds

```
<?xml version="1.0" encoding="UTF-8"?>
<D-Spin xmlns="http://www.dspin.de/data" version="0.3">
  <tns:MetaData xmlns:tns="http://www.dspin.de/data/metadata">
    <tns:source/>
  </tns:MetaData>
  <tns:TextCorpus xmlns:tns="http://www.dspin.de/data/textcorpus" lang="de">
    <tns:text>Karin fliegt nach New York. Sie will dort Urlaub machen.</tns:text>
    <tns:tokens>
      <tns:token ID="t0">Karin</tns:token>
      <tns:token ID="t1">fliegt</tns:token>
      <tns:token ID="t2">nach</tns:token>
      <tns:token ID="t3">New</tns:token>
      <tns:token ID="t4">York</tns:token>
      <tns:token ID="t5">.</tns:token>
      <tns:token ID="t6">Sie</tns:token>
      <tns:token ID="t7">will</tns:token>
```

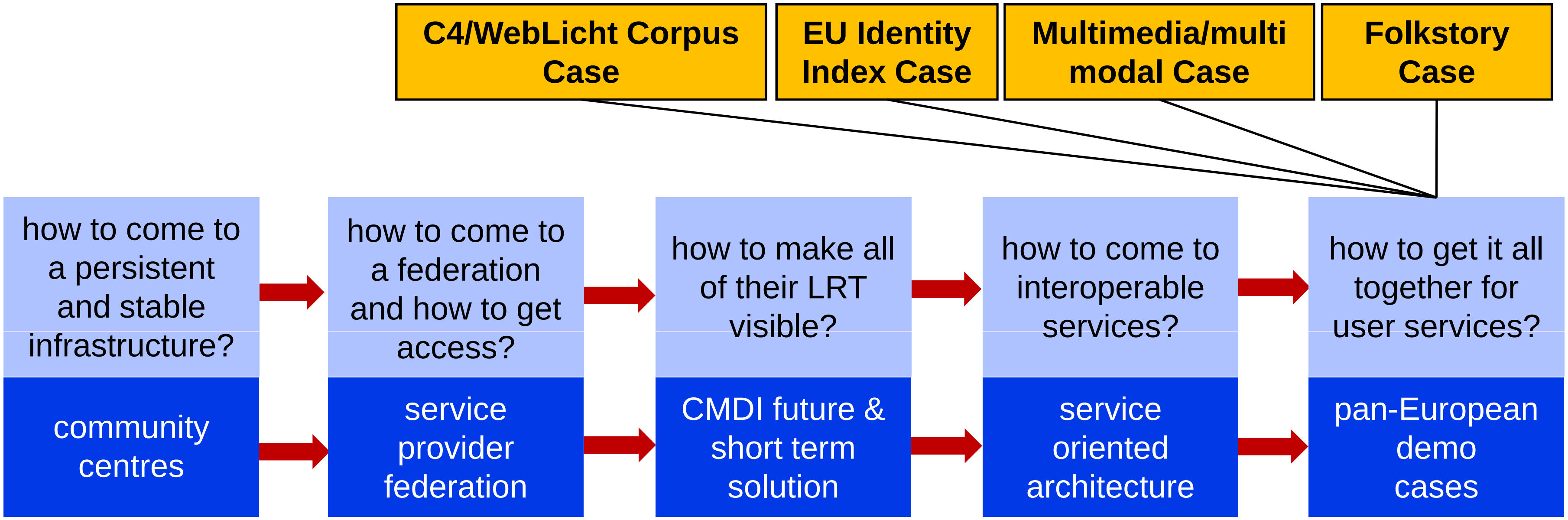
03EB1069277549D4AF3C0A90257FA856

Service Oriented Infrastructure

Web Services Interoperability

Standards & Best Practices

Demo Cases (just started)



C4/WebLicht Corpus Case

EU Identity Index Case

Multimedia/multi modal Case

Folkstory Case

how to come to a persistent and stable infrastructure?

how to come to a federation and how to get access?

how to make all of their LRT visible?

how to come to interoperable services?

how to get it all together for user services?

community centres

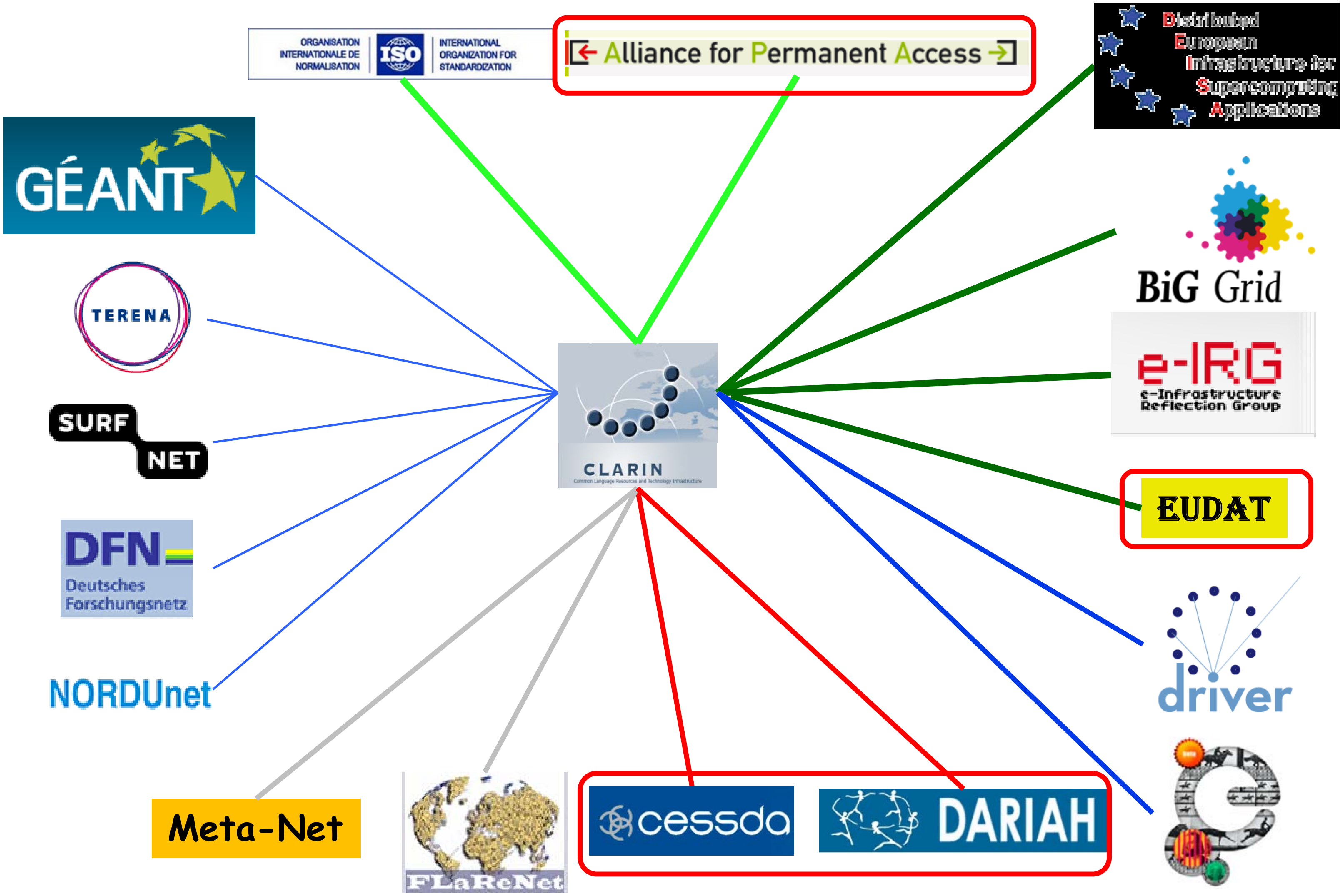
service provider federation

CMDI future & short term solution

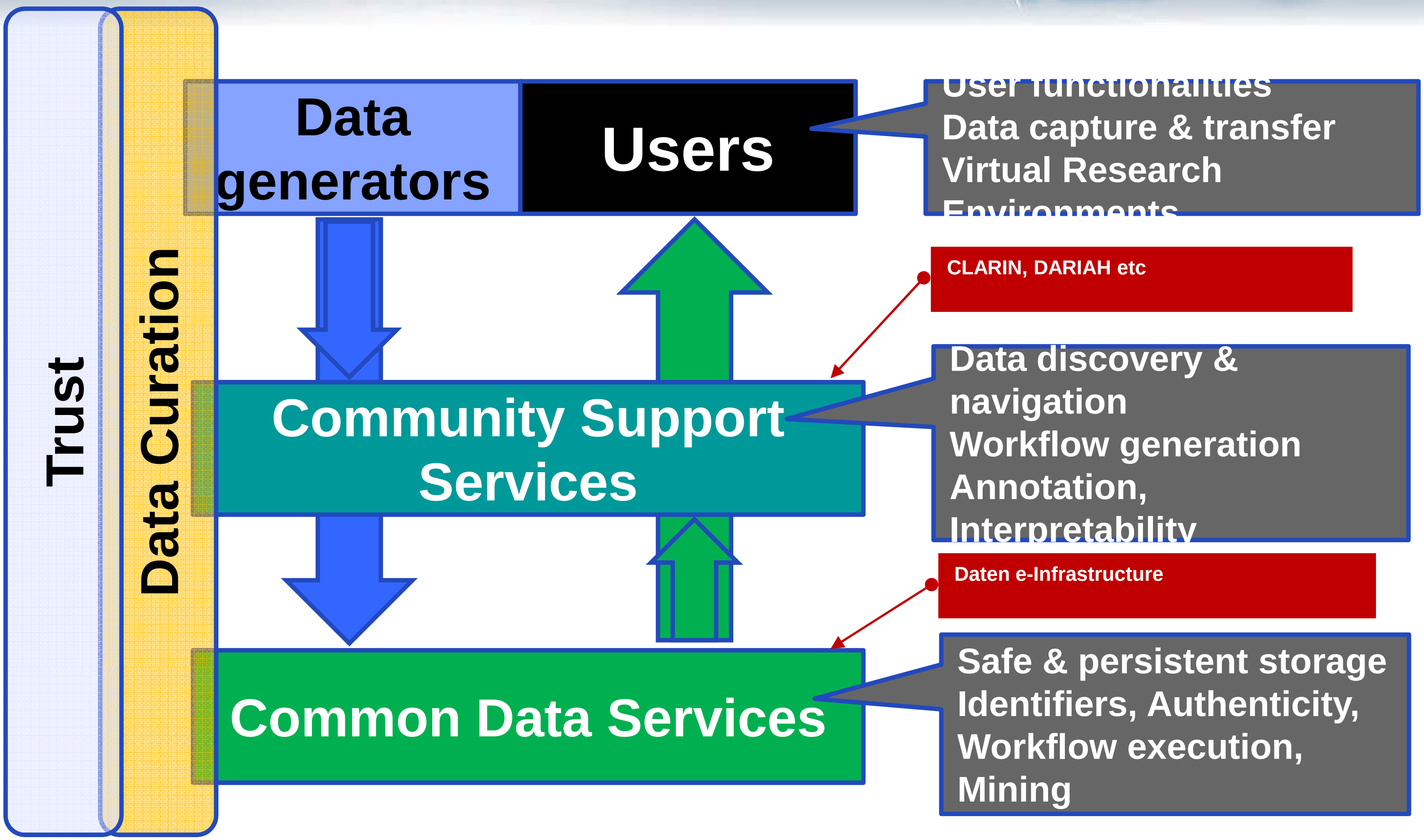
service oriented architecture

pan-European demo cases

not alone ...


ORGANISATION INTERNATIONALE DE NORMALISATION
ISO
INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
Alliance for Permanent Access
Distributed European Infrastructure for Supercomputing Applications
GÉANT
TERENA
SURF NET
DFN Deutsches Forschungsnetz
NORDUnet
CLARIN
Common Language Resources and Technology Infrastructure
BiG Grid
e-IRG
e-Infrastructure Reflection Group
EUDAT
driver
Meta-Net
FLaReNet
cessda
DARIAH

need to take care of data ...



Trust

Data Curation

Data generators

Users

User functionalities
Data capture & transfer
Virtual Research Environments

CLARIN, DARIAH etc

Community Support Services

Data discovery & navigation
Workflow generation
Annotation,
Interpretability

Daten e-Infrastructure

Common Data Services

Safe & persistent storage
Identifiers, Authenticity,
Workflow execution,
Mining

Architecture created by EC High Level Expert Group will be a guideline for coming decades

why European?

- live in a multilingual Europe with a joint historical tradition and need to exploit this strength
- many research questions are cross-national
- required standards cannot be national
- sharing costs in all respects is more efficient
- finally it's about global competition also in SSH

Why now?

- there is the ESFRI process and all countries are synchronized which is a unique chance to build infrastructures
- in total 44 initiatives on the ESFRI roadmap and there is the potential of gain by an eco system of RI
- we need to organize our resource domain due to huge increase of data (MPI: 200 TB)
- we need to take care to not loose our cultural and scientific memory
- there is a huge uptake of RI and there will be many funding streams!!!

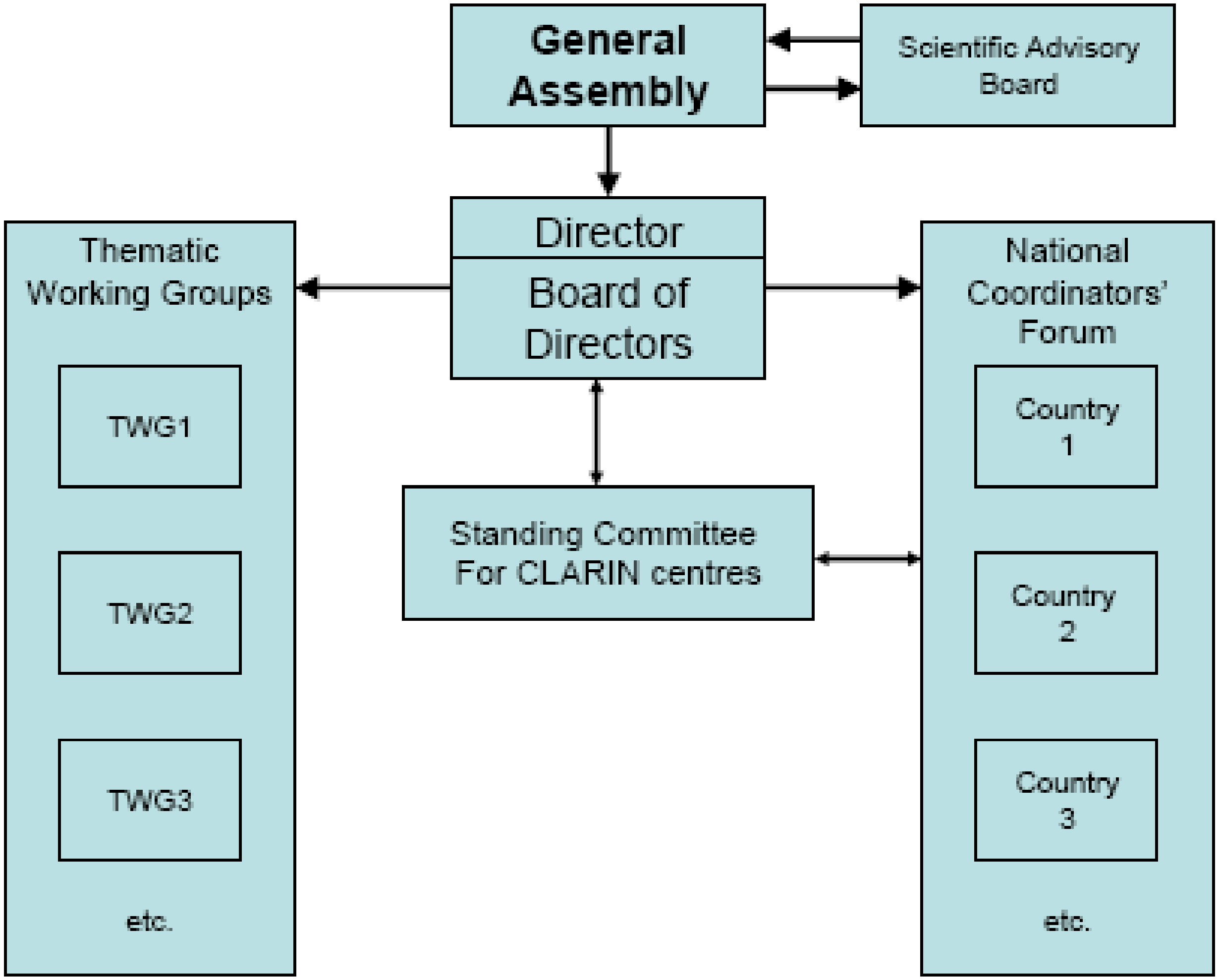
who and when?

- current EU CLARIN consortium in prep phase (08-10): 32 partners from 24 countries

- CLARIN construction phase from 2011; main funds by national programs - but additional funding streams by EC connected to RI

- legal issue: foundation of a European Research Infrastructure Consortiums (ERIC) as basis for future with automatic qualification to participate in programs

Organisation of the CLARIN ERIC


General Assembly

Scientific Advisory Board

Director

Board of Directors

Thematic Working Groups

TWG1

TWG2

TWG3

etc.

Standing Committee For CLARIN centres

National Coordinators' Forum

Country 1

Country 2

Country 3

etc.

CLARIN
Utrecht

who seems to be on board?

Belgium, Bulgaria, Germany, Denmark, Estonia,

Latvia, Finland, Croatia, Netherlands, Norwegen,

Austria, Portugal, Spain, Czech Republic, Hungary,

South Tirol, ?

Some are discussing: FR, SW, GR?, etc.

Advantage of membership

- privilaged access to CLARIN federation
- networked with CLARIN centres (direct technology transfer)
- a word when discussing priorities, agreements, best practices
- access to EC funding streams
- access to education and training programs to make our young generation competitive

Weitere Informationen

- CLARIN web site: http://www.clarin.eu
- CLARIN office: clarin@clarin.eu

- CLARIN Newsletter:

  http://www.clarin.eu/newsletter
- CLARIN members:

  http://www.clarin.eu/members

Thanks for your attention.

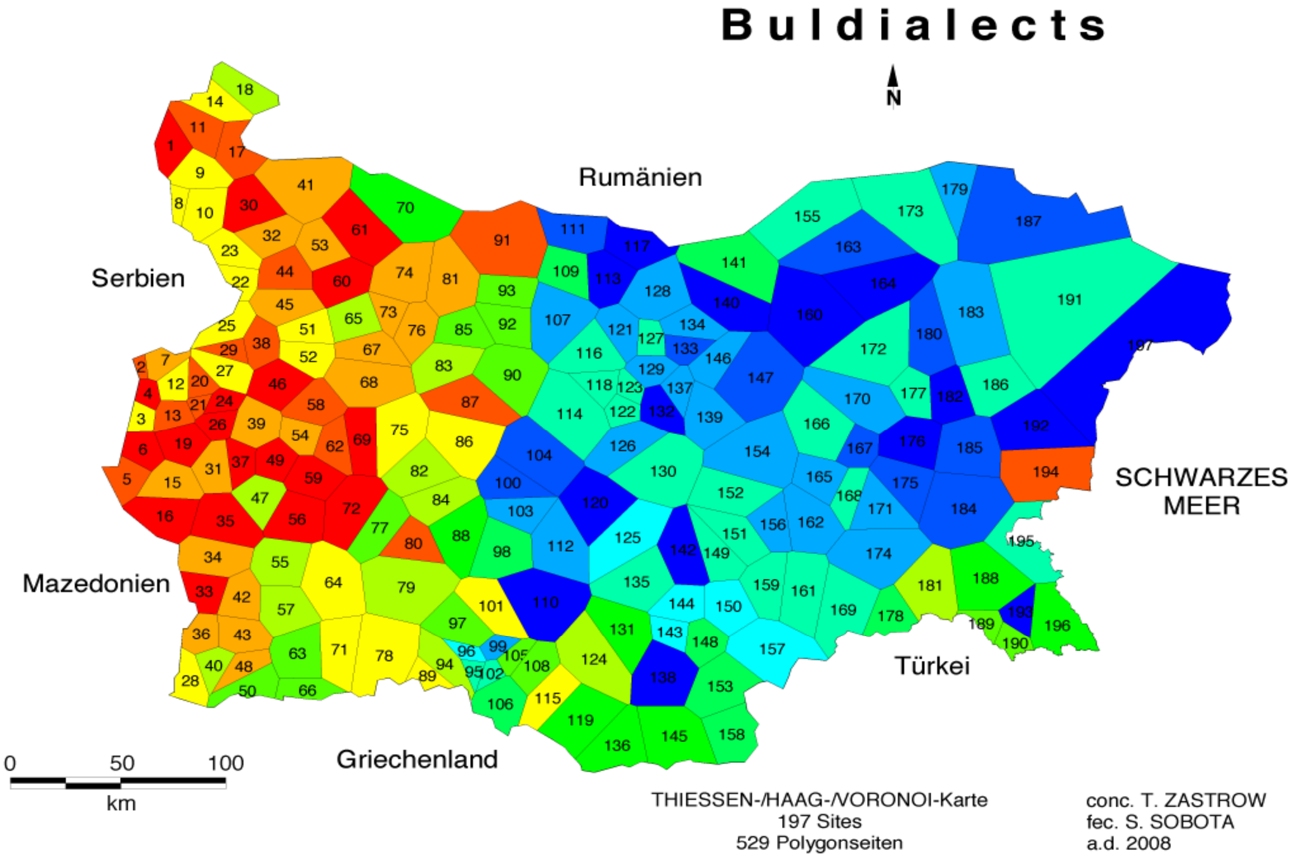
CLARIN Usage Scenario

- Scenario: A Serbian and a German PhD student want to study language variation in the Balkan area

- Resource: via VLO they find all relevant language variation data for that area

- Tools/Services: Modern clustering methods available via the web allow to quickly build dialect continua on top of a geographic map; visualization services allow to pipeline this to get a nice output

Visualization of Dialect Data: Clustering



THIESSEN-/HAAG-/VORONOI-Karte
197 Sites
529 Polygonseiten

conc. T. ZASTROW
fec. S. SOBOTA
a.d. 2008

CLARIN Usage Scenario

- Scenario: Linguists, sociologists and ethnologists want to study the cultural and linguistic differences of parliament debates in SE, DE and GR about the swine flue and compare how such global problems are dealt with
- Resource: building a virtual collections of all debates (Audio, Video, Transkription)
- Tools/Services: allowing researchers to analyse and annotate gestures, intonation, word choices, timing etc where partly powerful computers need being used
- Vision: in 2011/12 such computational services will be made available in CLARIN 2011