MAX-PLANCK-GESELLSCHAFT

Sharing Scientific/Research Data

Peter Wittenburg
CLARIN Research Infrastructure
DASISH SSH Cluster Project
EUDAT Common Data Infrastructure
iCORDI-RDA Global Data Sharing & Interoperability

The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

CLARIN
Common Language Resources and Technology Infrastructure

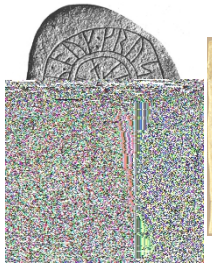
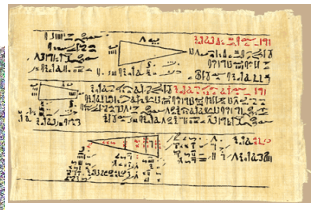
DASISH

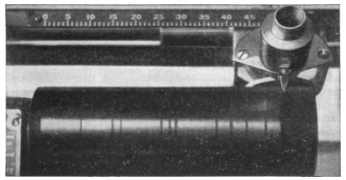
EUDAT

iCORDI

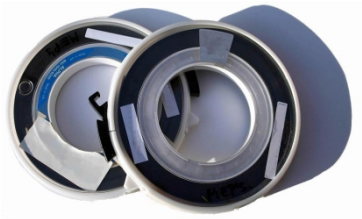
Sharing Data not new

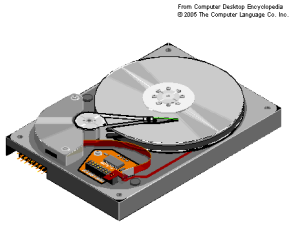
- of course nothing new
- in the good old times people used various carriers to exchange information - mostly a personal exchange in the analog era




- people started using and exchanging new carriers – the digital era



- something fundamentally changed

Sharing Data changes

- something fundamentally changed:
  - digital data can be copied exactly – can separate carrier and info
  - principle change: don't touch → touch frequently
  - independence of carriers allows using Internet
  - exchange can become anonymous (unknown producers and users)

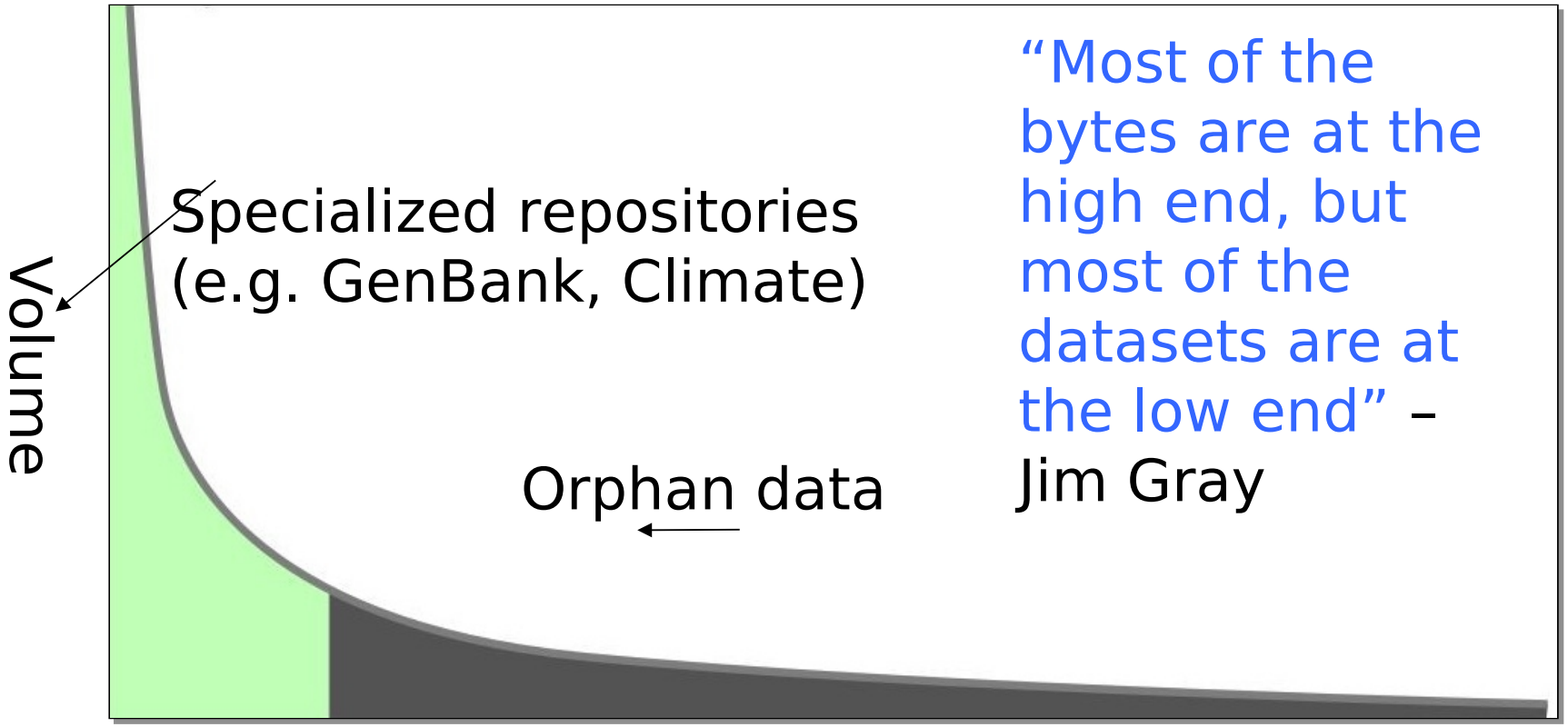
  - need to cope with changes
    - can we trust data – could be manipulated
    - can we trust creation/transformation process
    - can producers trust in seriousness of users
    - can we trust repositories of taking care (preservation, curation)
    - can we trust usage across borders
      (different legal and ethical systems)

Sharing Data - a bit more

- technological innovation has consequences
  - sensors becoming smaller and smarter -> huge amounts of data
  - sensors spread across the world -> even more data
  - computer simulation generates data -> even more data
  - mobiles allow massive crowd sourcing -> much & complex data
  - bit-streams are formatted/structured according to needs -> an increasing variety -> thus more complexity
  - experiments create regular data of huge sizes and with various conditions -> thus increasing complexity
  - analysis/transformations etc create huge amount of derivatives -> there are relations between files (better objects) and fragments of objects -> also increasing complexity
  - variety of transformations modify content -> thus adding history
  - etc.

our capability of creating data outperforms our capability of managing data

Regular-big and Long-Tail Data



Data in the Press

The Economist

ksgreen | Subscribe | Digital

World politics | Business & finance | Economics | Science & technology | Culture

Current issue | Previous issues | Special reports | Politics this week | Business this week

Technology

The data deluge

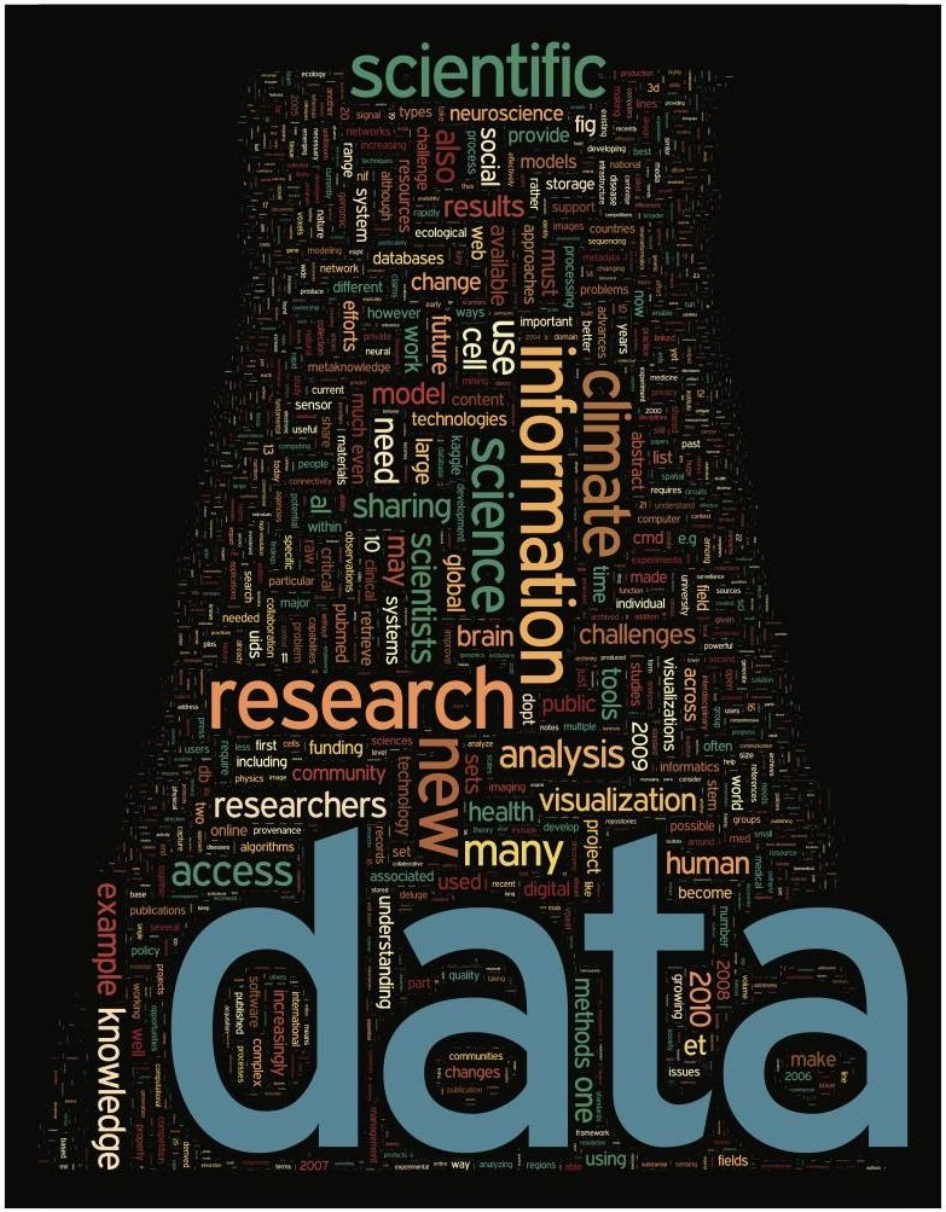
Businesses, governments and society are only starting to tap its vast potential

Feb 25th 2010 | from the print edition

Like 62 Tweet

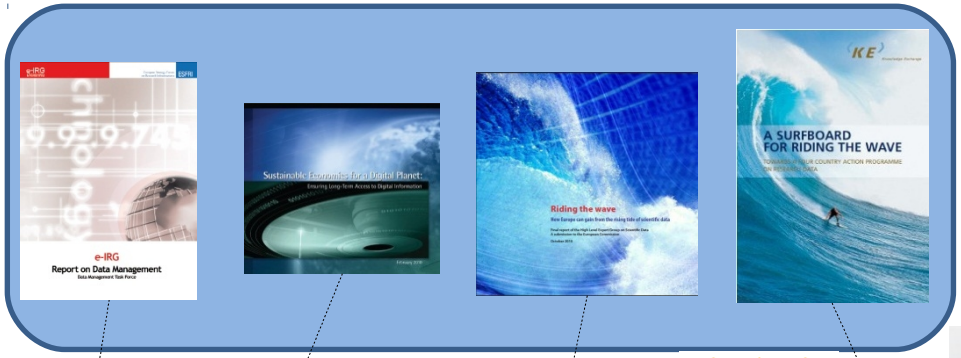
"Data is the new raw material of business: an economic input almost on par with capital and labor."

—Craig Mundie, Chief Research and Strategy Officer, Microsoft



Big Data presents considerable

Activities in Europe



OpenAIRE
Open Access Infrastructure for Research in Europe

2006/8
ESFRI

DataNet1

ands

ASIST Workshop

EUDAT

DataNet2

ALLEA Meeting Rome

iCORDI EC Flagship

UIPIU Data2012 Workshop

DAITF Prepar. Workshop

DWF Concept

2008	2009	2010	2011	2012

tackling first data topics

brainstorming on data issues, need for global action & first focussed actions

global interaction in place

Value of Data

elie Kroes (VP of European Commission)

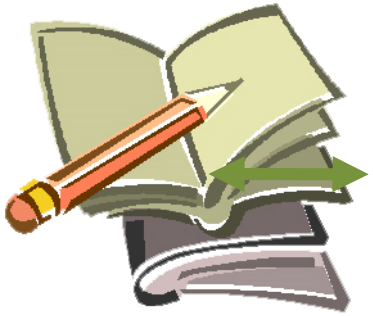
ata is the currency of modern research."

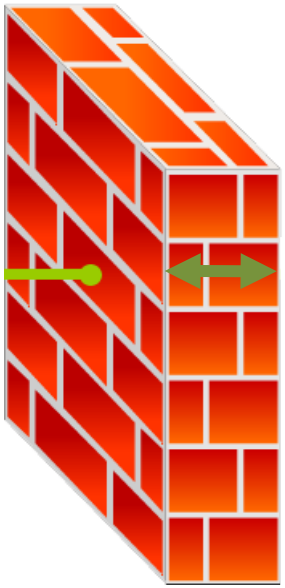
Implications

- we need to change our behavior in relation with data
- individual researchers and projects can't manage data anymore and take care of accessibility, preservation, curation etc.
- thus they need to hand over data to trusted repositories
- thus sharing in our era means accessing data from a repository and not from a researcher personally (will become an exception)

- just accessing newly created data? – what about sharing old data
  - C. Huc: 40% of data access is to old data
  - humanities: even more data to be accessed is old data
- we need mechanisms
  - to ensure that we get the data object we want
  - to get context and provenance information to interpret and re-use

Change of Culture





Wall of Silence

Some well-known problems:

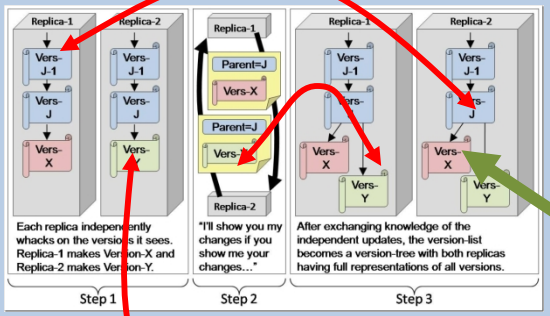
no-persistency, hardly any sharing, no correctness proof, etc.

Change of Culture

why should I change?

should I really look into this data mess?

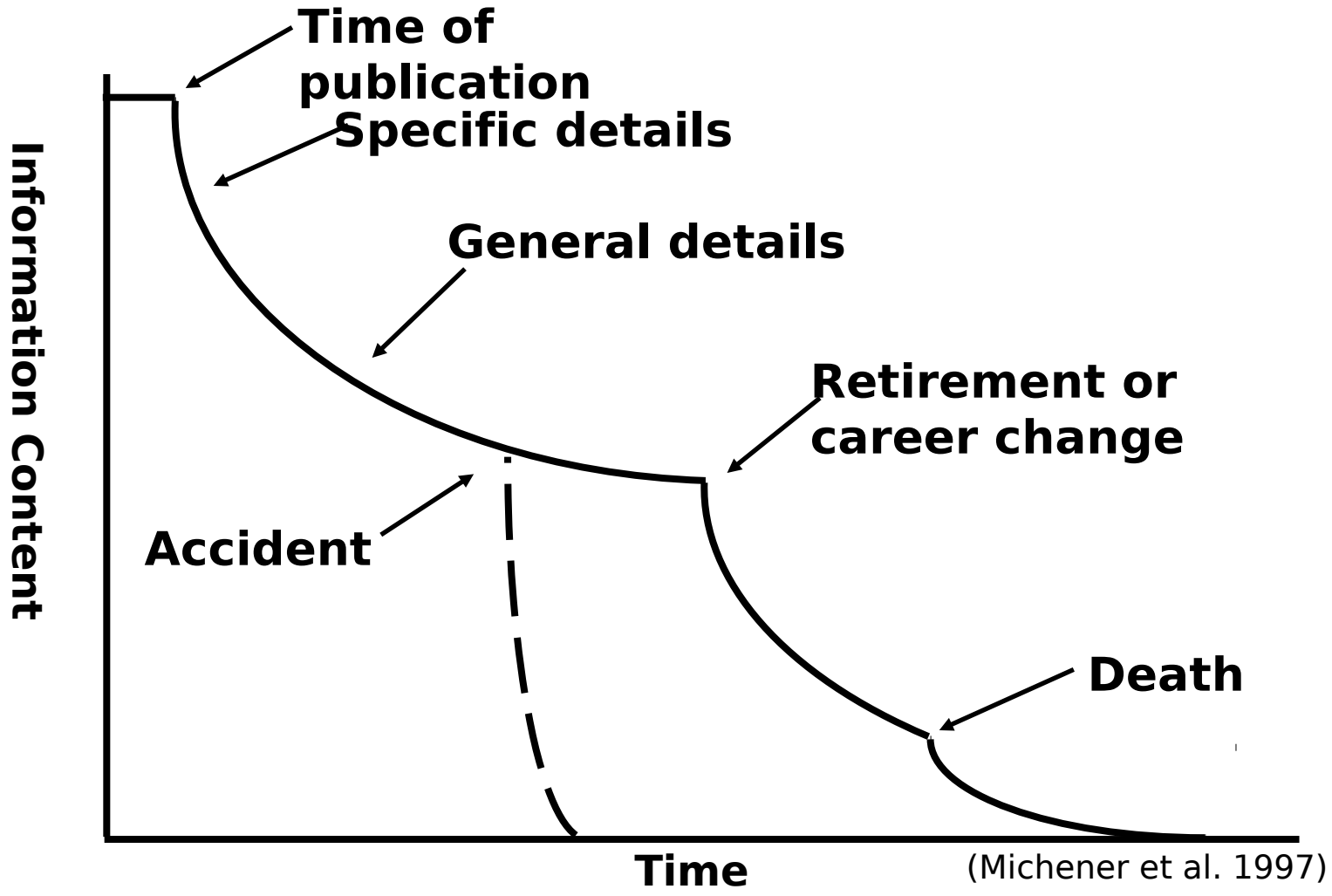
Linked Data Universe



Change in culture required - will not be that easy:

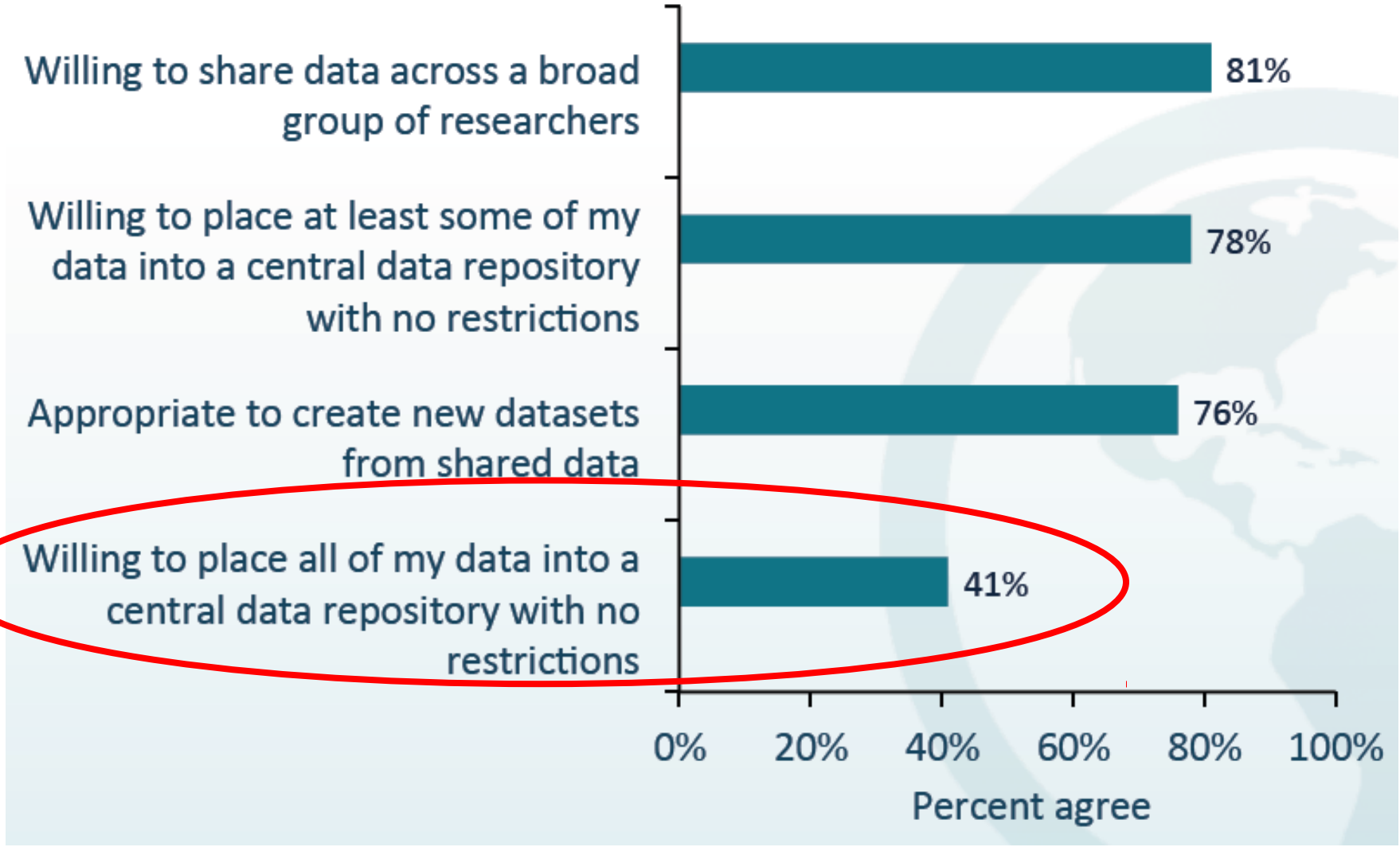
- more work (management, curation), costs?, career?, quality?, etc.
- benefits for small and grand research challenges?

Suffer from Data Entropy



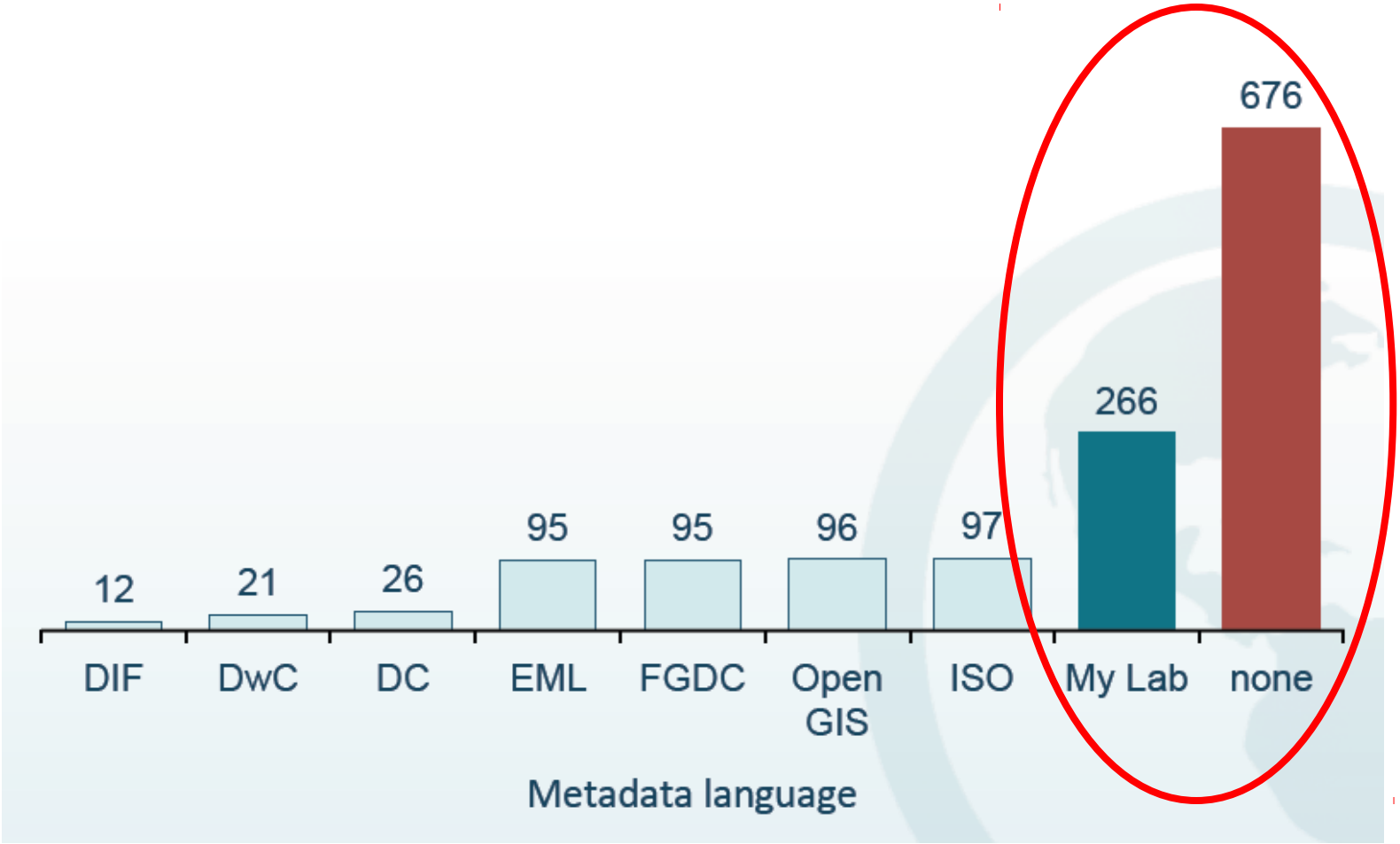
(Michener et al. 1997)

Willingness of sharing



Rebecca Koskela: DataONE

Reality of finding



Rebecca Koskela: DataONE

Need a Data Infrastructure

ing The Wave (EC's HLEG on Scientific Data)

e emerging infrastructure for scientific data must be flexible
nble, secure yet open, local and global, affordable yet
-performance. Obviously, this is a tall order – and here is n
technology that we know to[illegible] *orrow to*
ieve it a[illegible] *al*
nework[illegible] *rsities,*
ernment[illegible] *em –*
t types[illegible] *erformance*
rics shou[illegible] *this framework would ensure the*
tworthiness of data."

need a flexible, open, global data infrastructure

no one technology solution – not a monolithic design

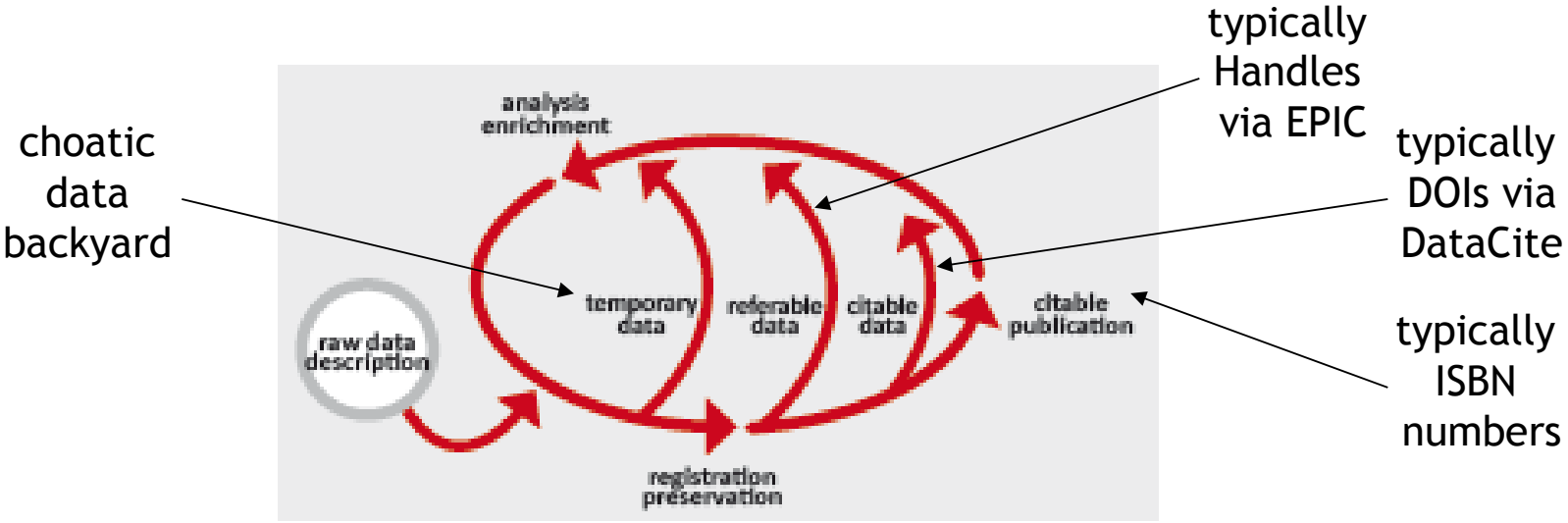
trust is a core concept to be taken serious

Rethinking Data

e are we talking about?
to rethink the way we are dealing with data!
something we did as a side job to a real task!
something where we used directories/files to someth

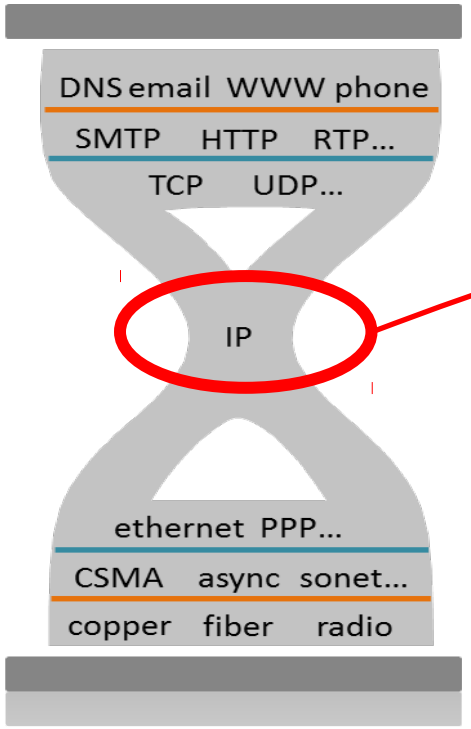
Data is in Scientific Cycle

- huge amounts of data objects are created automatically as part of workflows and by manual activities (think of massive crowd sourcing)
- new data objects will be used immediately and people/workflows will refer (use, citation) to them



- thus we need unique and persistence references to refer to data objects and need contextual/provenance information to allow re-usage
- AND: create them immediately otherwise costs increase by factors

Learn from Internet



Value Added Services

Internet Protocol Suite

Network Technology

Analysis Persistent Reference Citation
Apps Custom Clients Plug-Ins
Resolution System Typing

PID

Digital Objects
Data Sets RDBMS Files
Local Storage Cloud Computed

Value Added Services

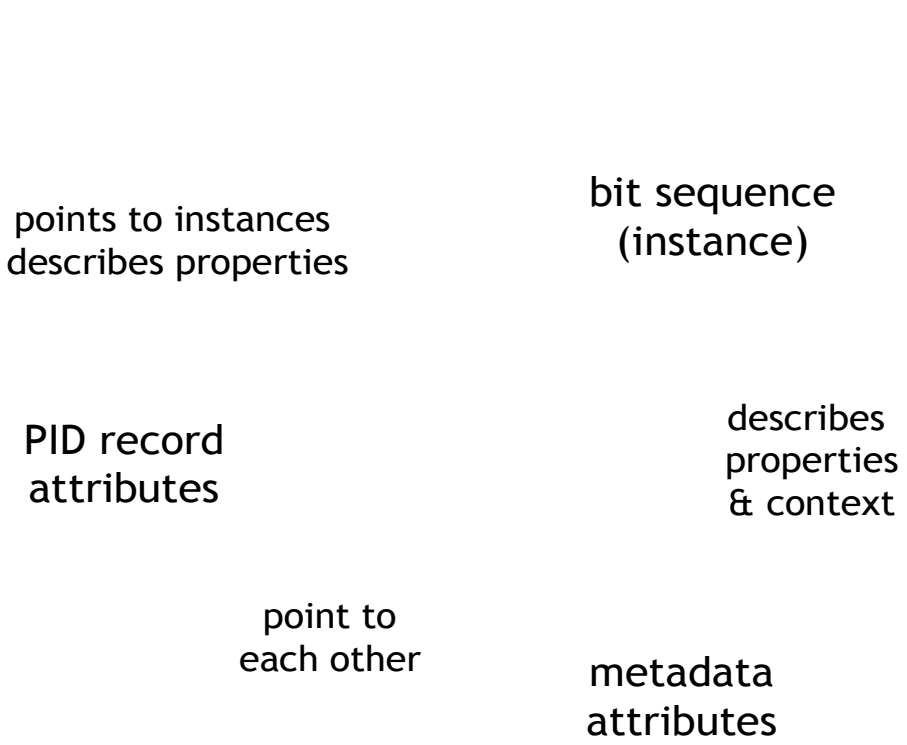
Persistent Identifiers

Data Sources

Internet Domain
nodes with IP numbers
packages being exchanged
standardized protocols

Data Domain
Objects with PID numbers
objects being exchanged
standardized protocols

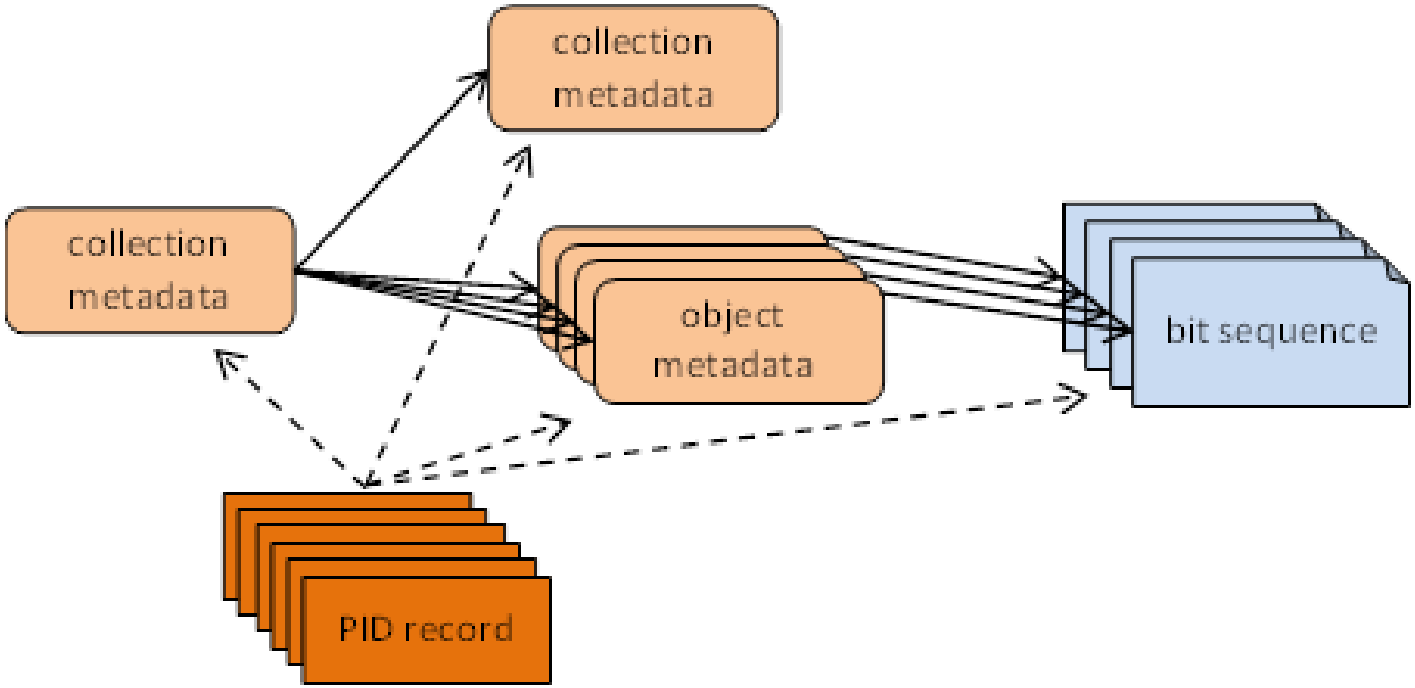
Domain of Registered Data Objects



•need to take care of domain of DO existing of

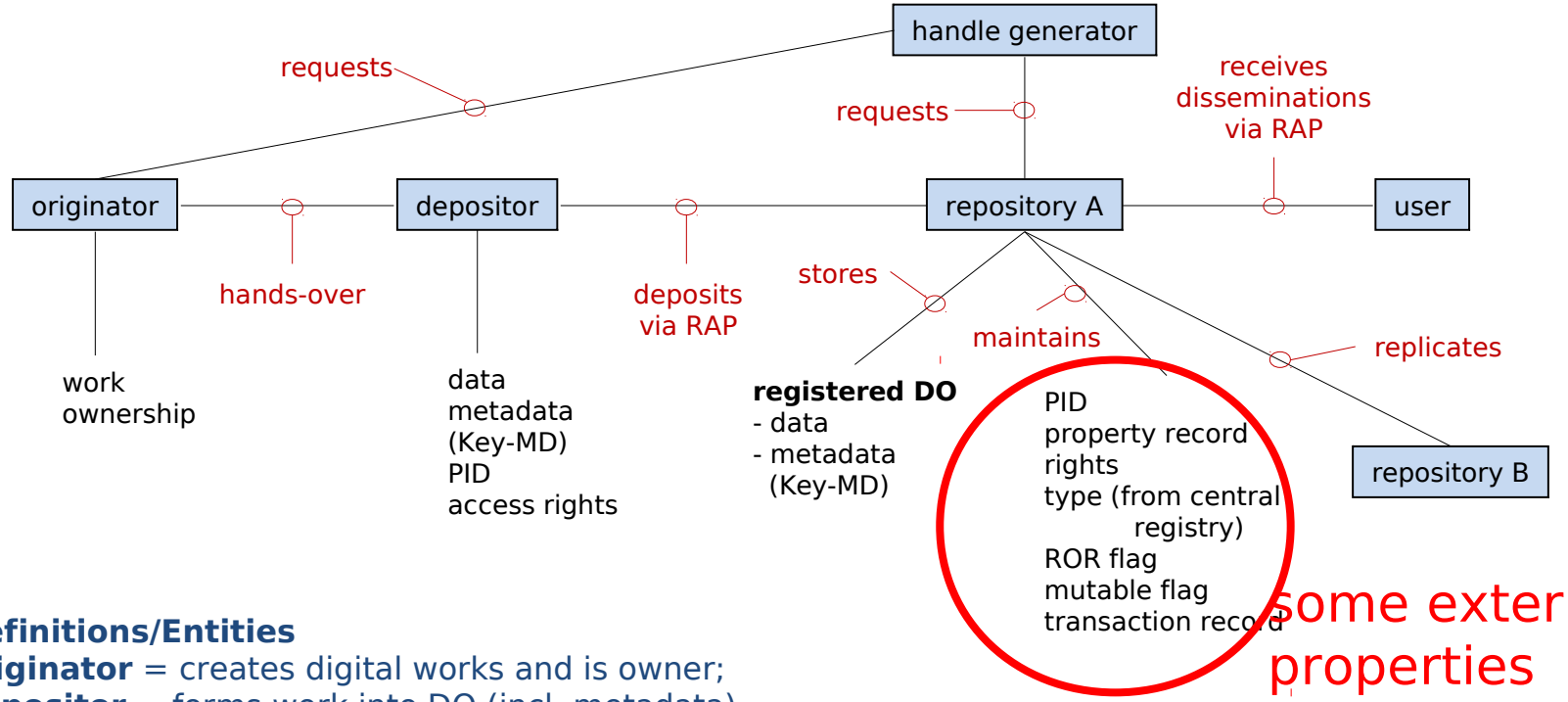
- instances of bit sequences stored at different repositories
- a PID that points to all instances
- a metadata object storing contextual and provenance information
- PID and MD store „external“ properties of data objects
- utterly important for PID is checksum to prove integrity
- utterly important for PID and MD is information indicating authenticity

Domain of Registered Collections



- **a collection is an aggregation of DOs**
  - collections designed at creation time under specific contexts (experiment, etc.)
  - users can create arbitrary (virtual) collections aggregating objects and collections for specific purposes such as a dissertation (they need to be referable and citable as well)

Data Object World of Bob Kahn



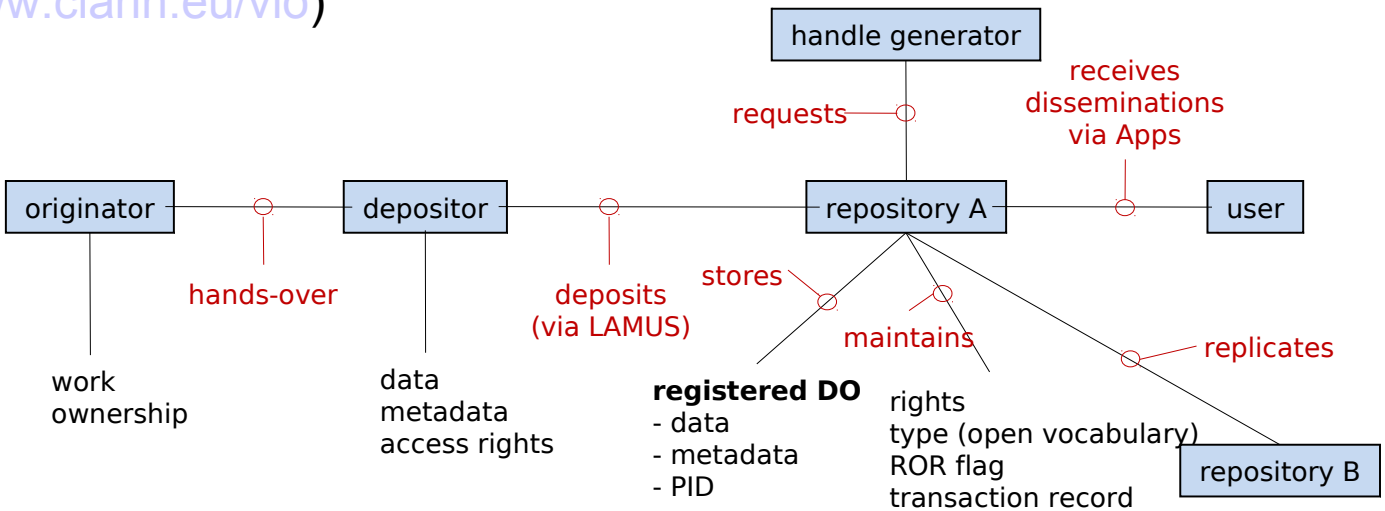
some external properties

Definitions/Entities
originator = creates digital works and is owner;
depositor = forms work into DO (incl. metadata),
digital object (DO) = instance of an abstract data type;
registered DOs are such DOs with a Handle;
repository (Rep) = network accessible storage to store DOs;
RAP (Rep access protocol) = simple access protocol
Dissemination = is the data stream a user receives
ROR (repository of record) = the repository where data was stored first;
Meta-Objects (MO) = are objects with properties
mutable DOs = some DOs can be modified
property record = contains various info about DO
type = data of DOs have a type

- from Kahn & Wilensky paper on Digital Objects from 2006 as basis for interactions
- worked extremely well

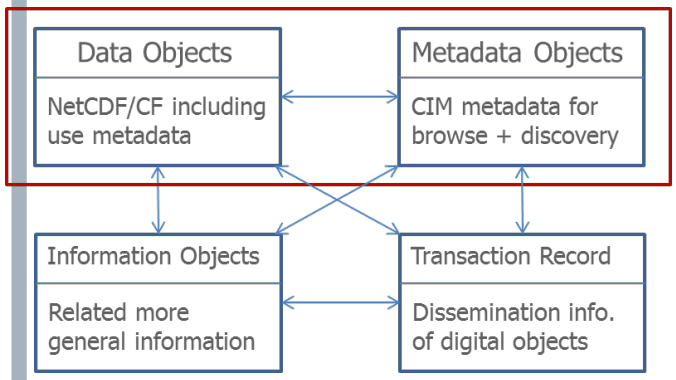
what happens in Language Community

- **CLARIN (Language Resource and Technology Community)**
  - about 200 centers in Europe with about 30 „community center“ candidates
  - have 4 types of centers (DataONE: tiers) from strong to weak requirements
  - 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)

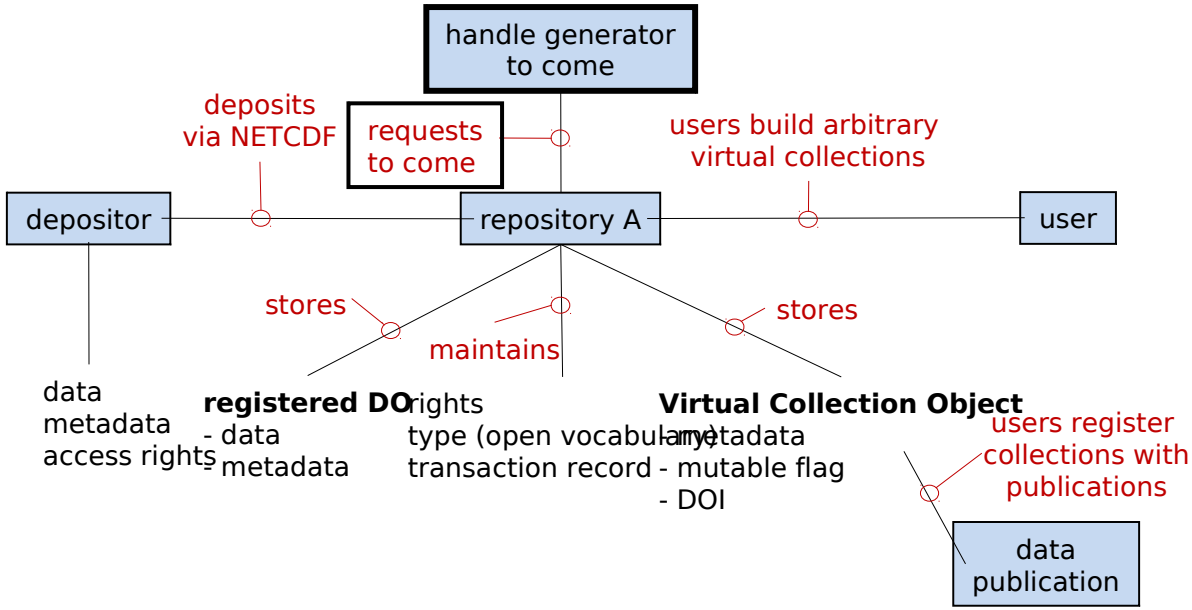


what happens in Climate Community

- **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

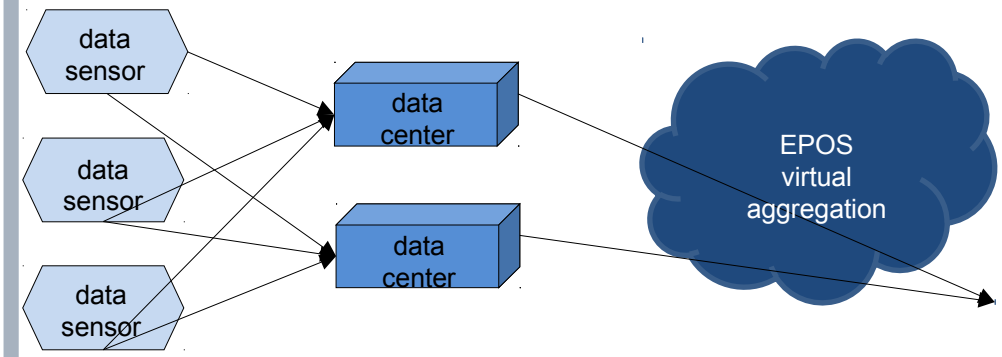


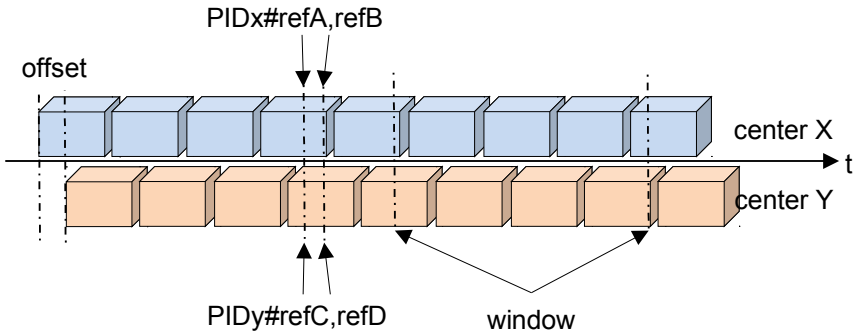
Identification of distinct data objects and P2P infrastructure



What happens in earth observation c.

- **EPOS (Seismologists, Vulcanologists, 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 „vulcano 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
  - need to synchronize with US and other colleagues





Quality and Certification

eed to check quality of data when we want to share them.

- producers
- repositories (as new player in the chain)

Data Management Plan

- be aware: for NSF, Dutch, etc. grants you need to present a data management plan
  - yet no specific requirements but they will come
- what's that – let's take DMP example from UK Digital Curation Center
  1. Introduction and Context

     (name, funder, budget, duration, aim, policies, dates, etc.)
  1. Data Types, Formats, Metadata, Standards, Capture Methods
  2. Ethics and Intellectual Property (CoC, ownership, copyright, etc.)
  3. Access, Data Sharing and Re-use (who else interested, why not sharing, costs, restrictions, embargo, etc.)
  4. Short-Term Storage and Data Management (volume, storage media, responsibilities, backup, security, etc.)
  5. Deposit and Long-Term Preservation (strategy, duration, MD, repository/archive, appraisal/retention, curation, policies, etc.)
  6. Resourcing/Review (staff, roles, costs, checks, etc.)

Certification of Repositories

- it's all about trust building – you should get what you want
- we now have three major quality assessment procedures:
  - Data Seal of Approval (DSA): light procedure (2 pw)
    assessment of claims a repository is making!!!
  - NESTOR guidelines (DE – DIN)
  - Repository Audit & Certification (RAC): heavy procedure (3 pm)

- DSA criteria for repositories in more detail:
  - data must be found on Internet
  - data must be accessible (accepting ethical & legal restrictions)
  - data is available in usable formats
  - data is reliable
  - data can be referred to
  - separation in producer, consumer and repository roles

Certification of Repositories

- repository has explicit mission, ensures compliance with legal&ethical norms, applies documented processes and procedures for data management, has a long-term preservation strategy, carries out archiving according to explicit workflows, assumes responsibility wrt data access, enables users to use and refer to data, ensures integrity of data and metadata and ensures authenticity of data and metadata
- repository's technical infrastructure supports OAIS

- producers in DSA
  - producers deposit data in a repository with sufficient information for others to assess the scientific and scholarly quality of data
  - producers provide data in formats recommended by repository
  - producers provide data together with metadata as required by repository

Certification of Repositories

- consumers in DSA
  - consumers must comply with access regulations of repository
  - consumers conform to CoC guiding exchange and proper use of data, knowledge and information
  - consumers respect licenses with respect to use of data

- regulations, CoC, etc. at many different levels:
  - repository
  - institution
  - community
  - state
  - OECD, UNESCO-WIPO, Creative Commons, etc.

Canonical Workflows

Need insight in canonical Workflows to understand basic layers components for sharing & re-using.
Need to world-wide harmonize essential components!
Need to adhere to basic IT principles!

Access Workflow

Enabling Technologies

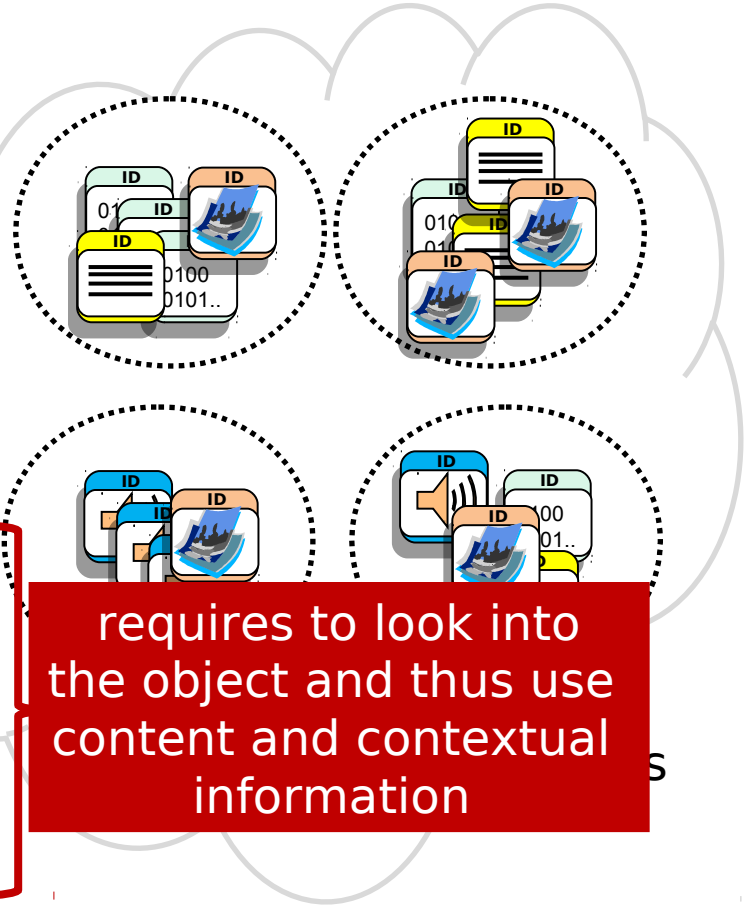
Scientists, Data Curator
End Users, Applications

Discovery

Access
(ref. resolution, protocols, AAI)

Interpretation

Reuse



requires to look into the object and thus use content and contextual information

Management Workflow

Data Managers
Data Scientists

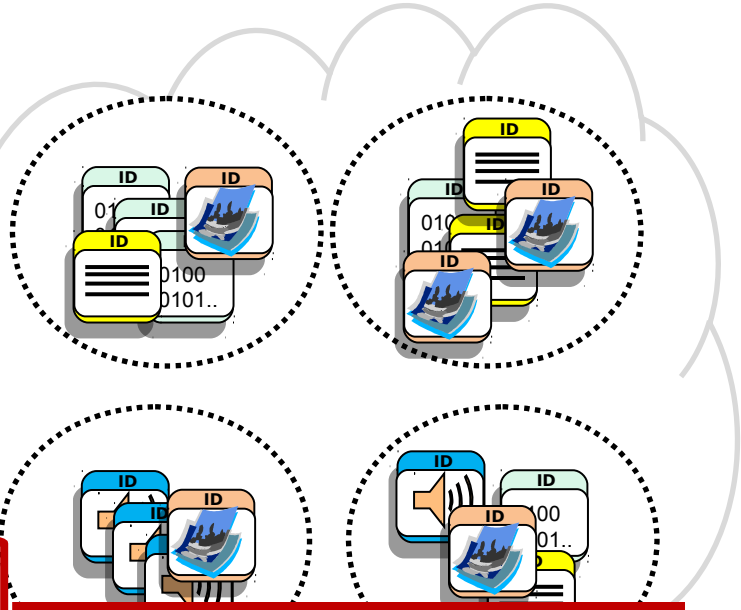
Enabling Technologies

Collections + Properties

Access
(ref. resolution, protocols, AAI)

formalized policies workflow engine

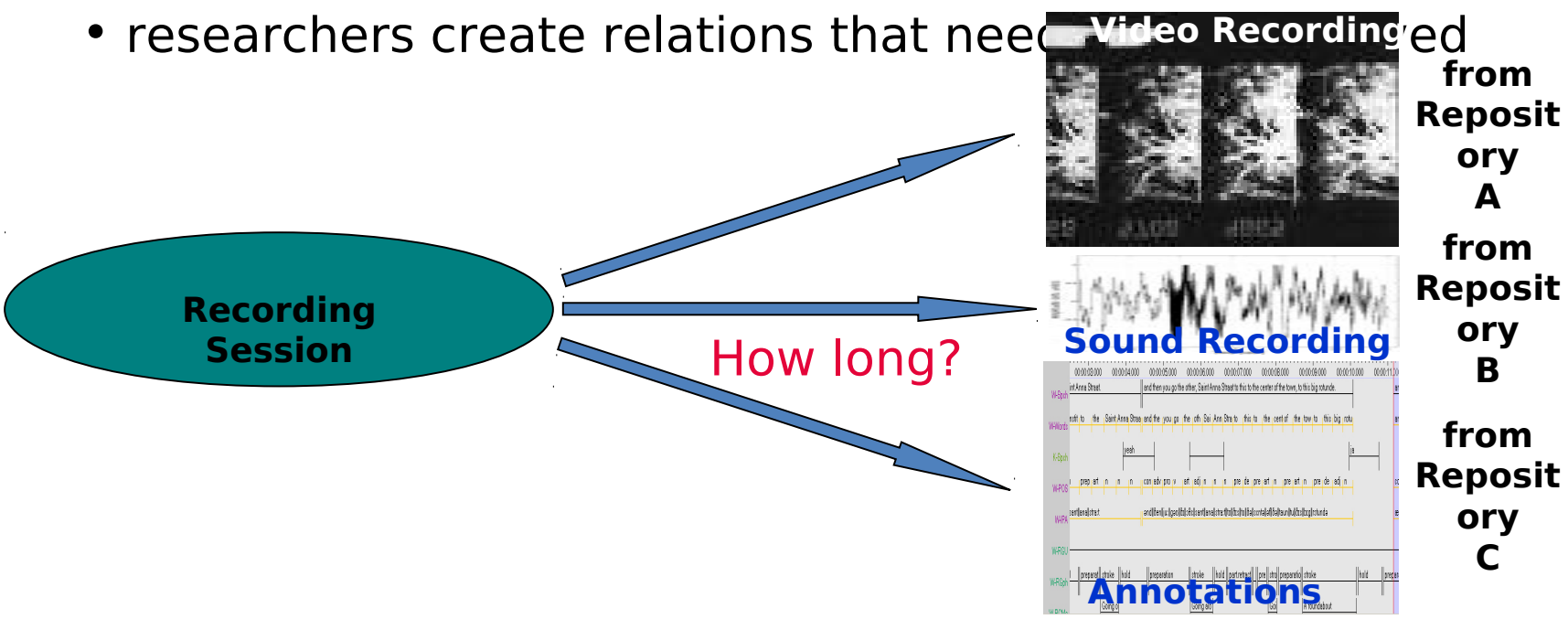
Assessment



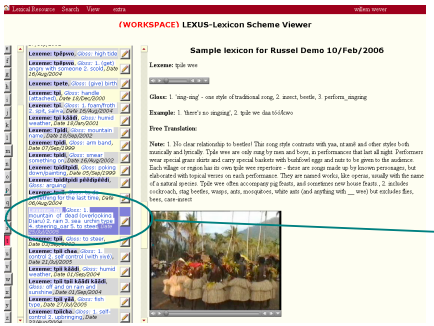
can all be done based on properties stored in PID/Metadata attributes (in general external prop.)

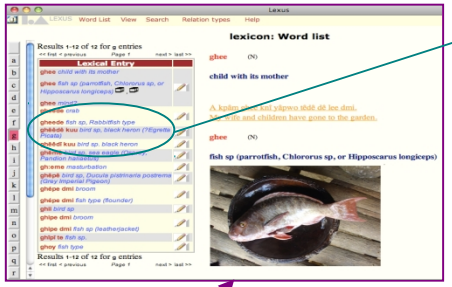
Why PIDs

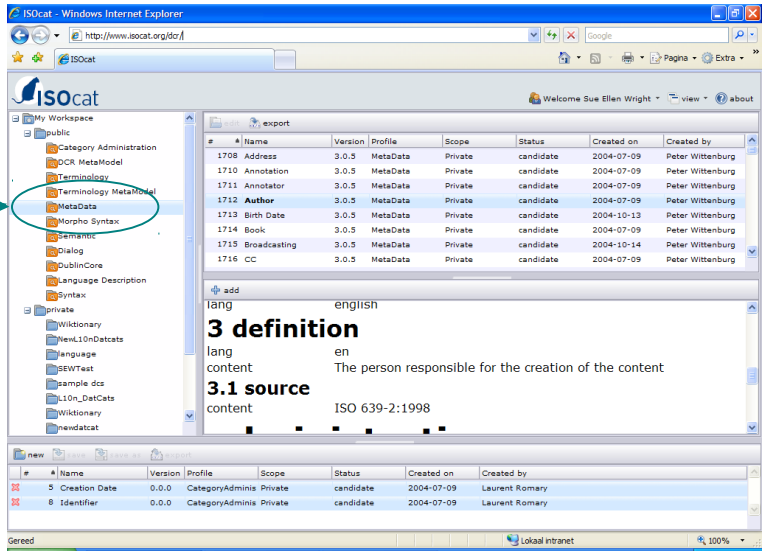
- assume that we have a recording of an extinct language and some
  annotations that tell us what someone said about medicine etc
- researchers create relations that nee ed



Why PIDs







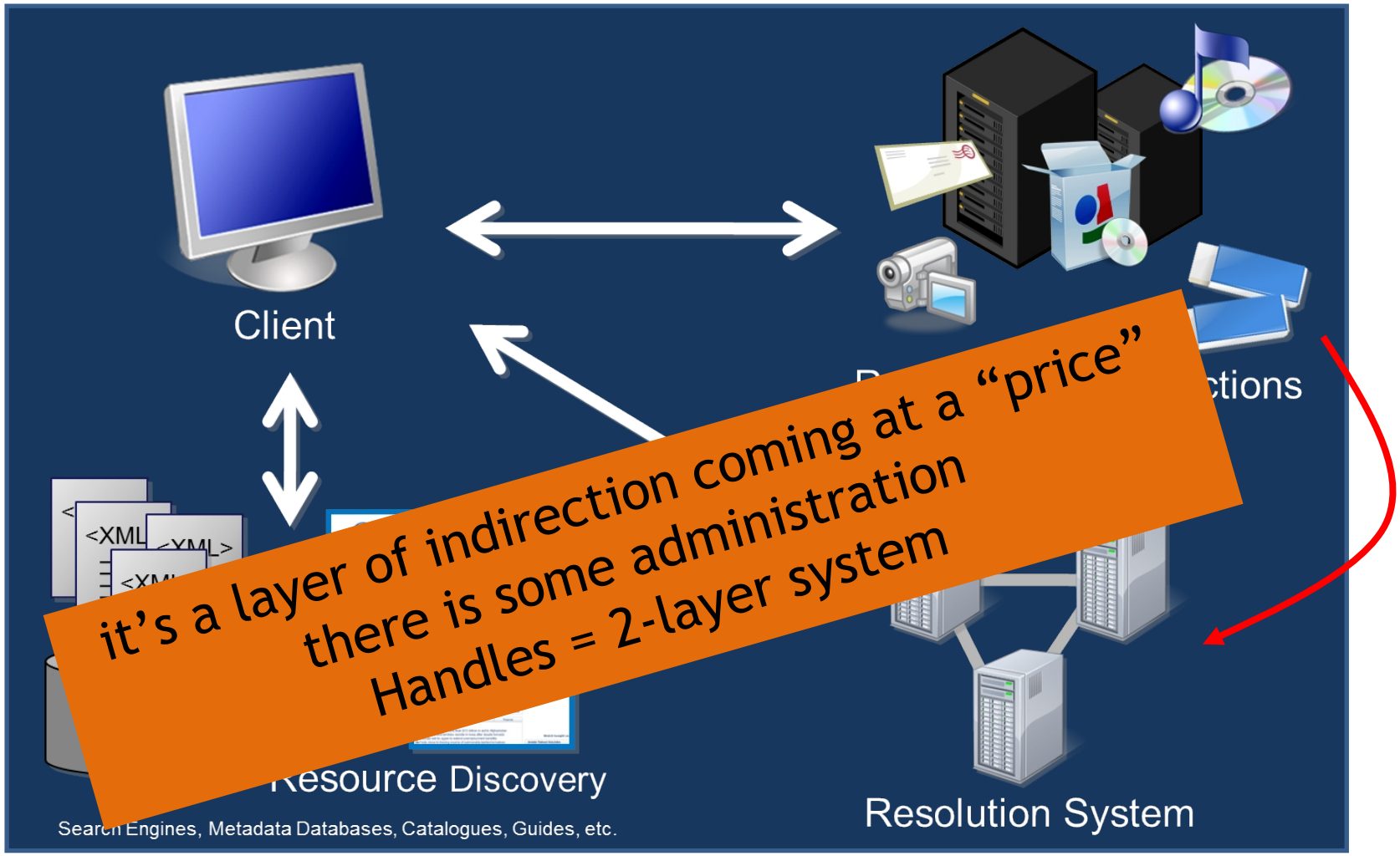
eResource1
Repository 1

eResource2
Repository 2

How long?

Ontology
open registry

PIDs (Handles/DOIs) not for free



DOI = Handle + business model; for science: EPIC = Handle + sc-bm

PIDs embedded in Metadata

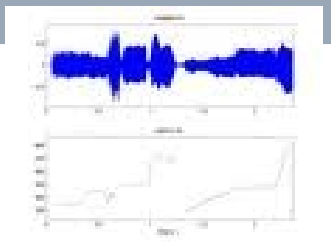

```
<?xml version="1.0" encoding="UTF-8"?>
<METATRANSCRIPT ArchiveHandle="hdl:1839/00-0000-0000-0005-82B0-2"
    Date="2006-07-18" FormatId="IMDI 3.0"
    Originator="Editor - Profile:SESSION.Profile.xml" Type="SESSION"
    Version="1" xmlns="http://www.mpi.nl/IMDI/Schema/IMDI"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI ./IMDI_3.0.xsd">
 <Session>
      <Name>DBD_RIF_14_12_01_064</Name>
      <Title>Dutch Bilingualism Database, Ethnic Dutch, Session 64</Title>
    ..........

<MediaFile>
<ResourceLink ArchiveHandle="hdl:1839/00-0000-0000-0004-DC6B-0">
    http://corpus1.mpi.nl/qfs1/media-archive/dbd_data/boumans/T-
    Cult/Metadata/../Media/dbd_rif_14_12_01_064.wav</ResourceLink>
    ..........
```

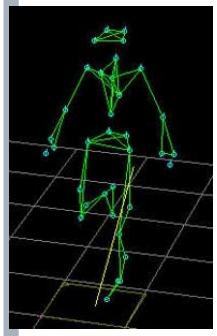
Data & repository in MPI

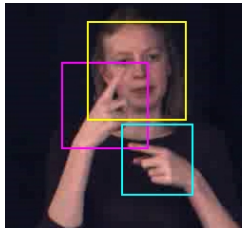
What happens in my institute to
increase possibilities of sharing and re-using?

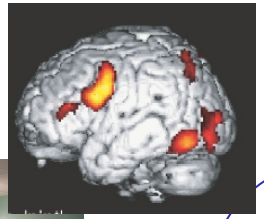
What's in the big pot?

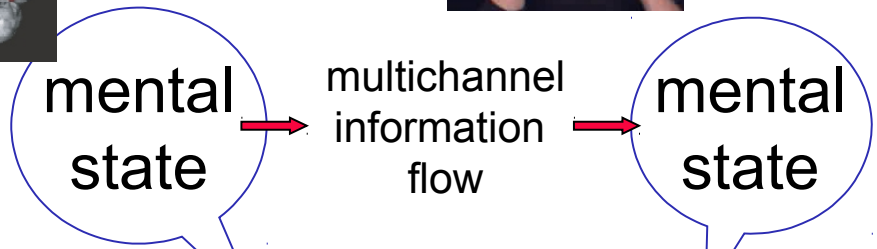


- using all channels in interaction
  - speech sounds
  - suprasegmental information (pitch, intensity, etc)
  - eye movements
  - head movements
  - hand/arm movements (gestures)
  - body movements
  - Virtual Reality
  - EEG/fMRI
  - etc.









- task: understand each other
- NOT: produce grammatic speech

DOBES Project

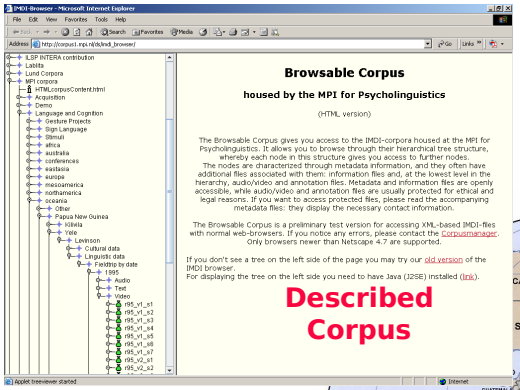
DOBES = Documentation of Endangered Languages

some facts

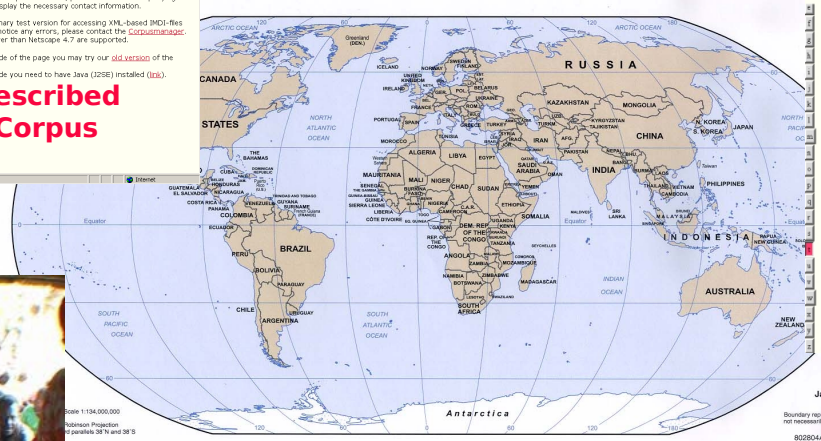
- started 2000 with 7 international teams and 1 archive team
- 2012: now 68 documentation teams working almost every where

- cross-disciplinary approach: linguists, ethnologists, musicologists, biologists, ship builders, etc.
- every year

Online Archive as a Result



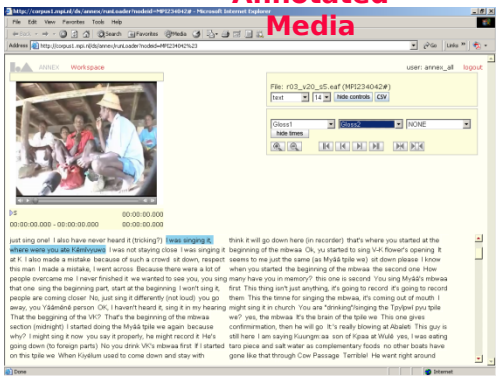
Described Corpus



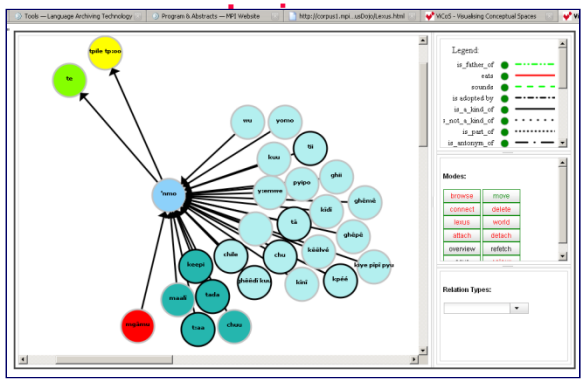
Multimedia Lexicon

Video clips, sound files & images

Annotated Media



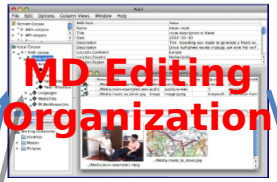
Typed Relations within the

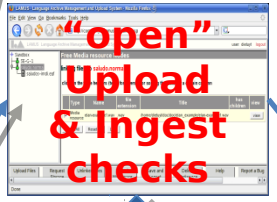


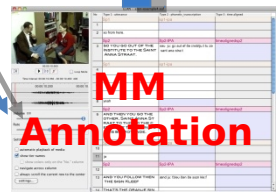
machinery at MPI – standards based

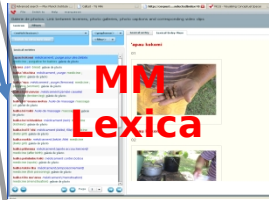
MAX-PLANCK-GESELLSCHAFT

MPI & external researchers





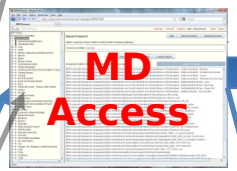


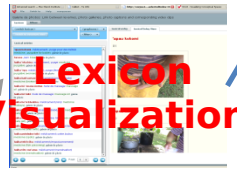


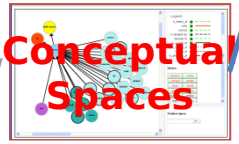
Organizational
- ~ 22 experts
- DSA Quality Seal
- 50 y data guarantee
- 5 y TLA guarantee

The Language Archive
80/200 TB
annual audit

External Services
- PID Service
- AAI Service
- 4 Replications
- Service Hosting
- etc



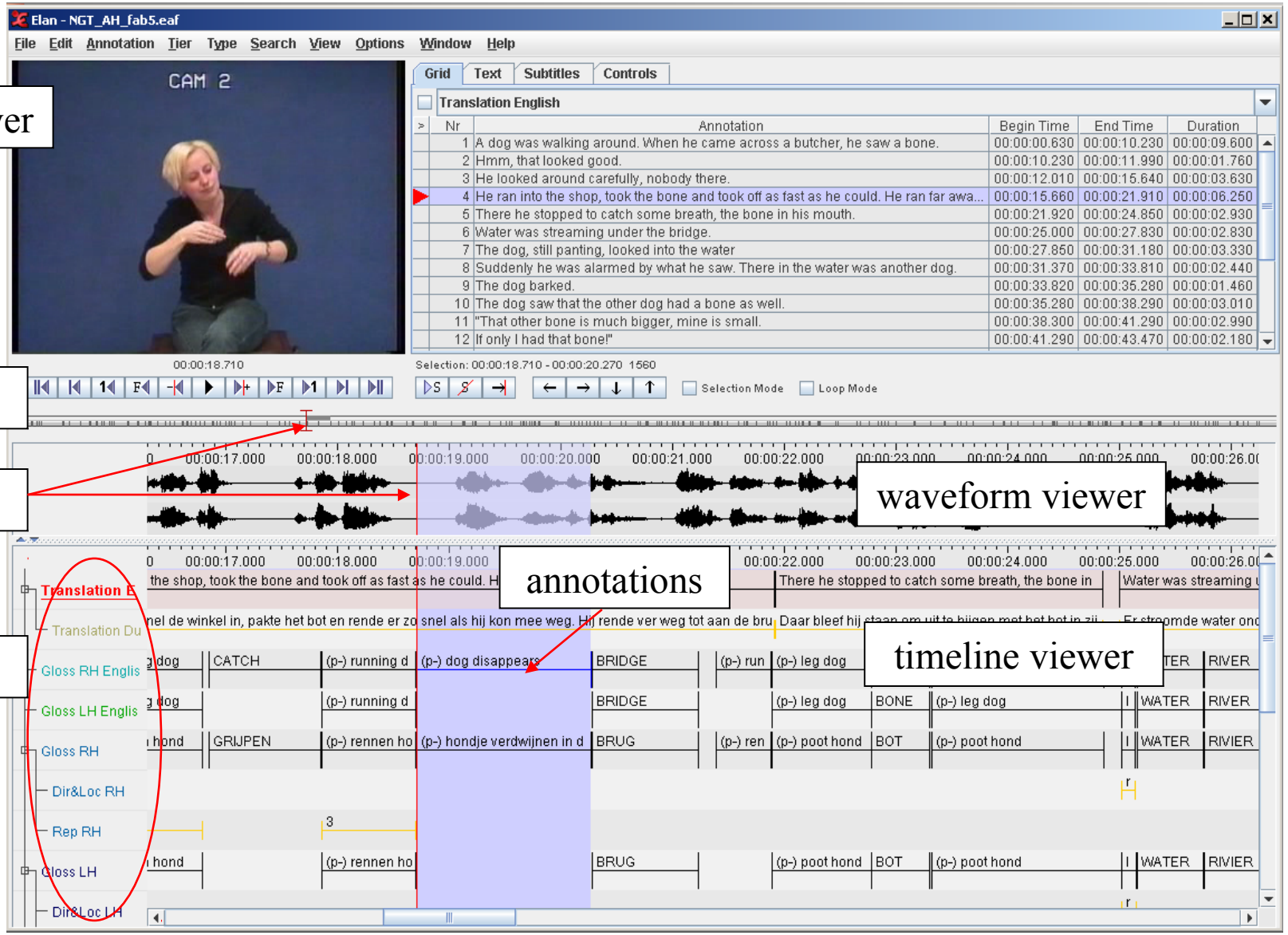




MPI & external researchers

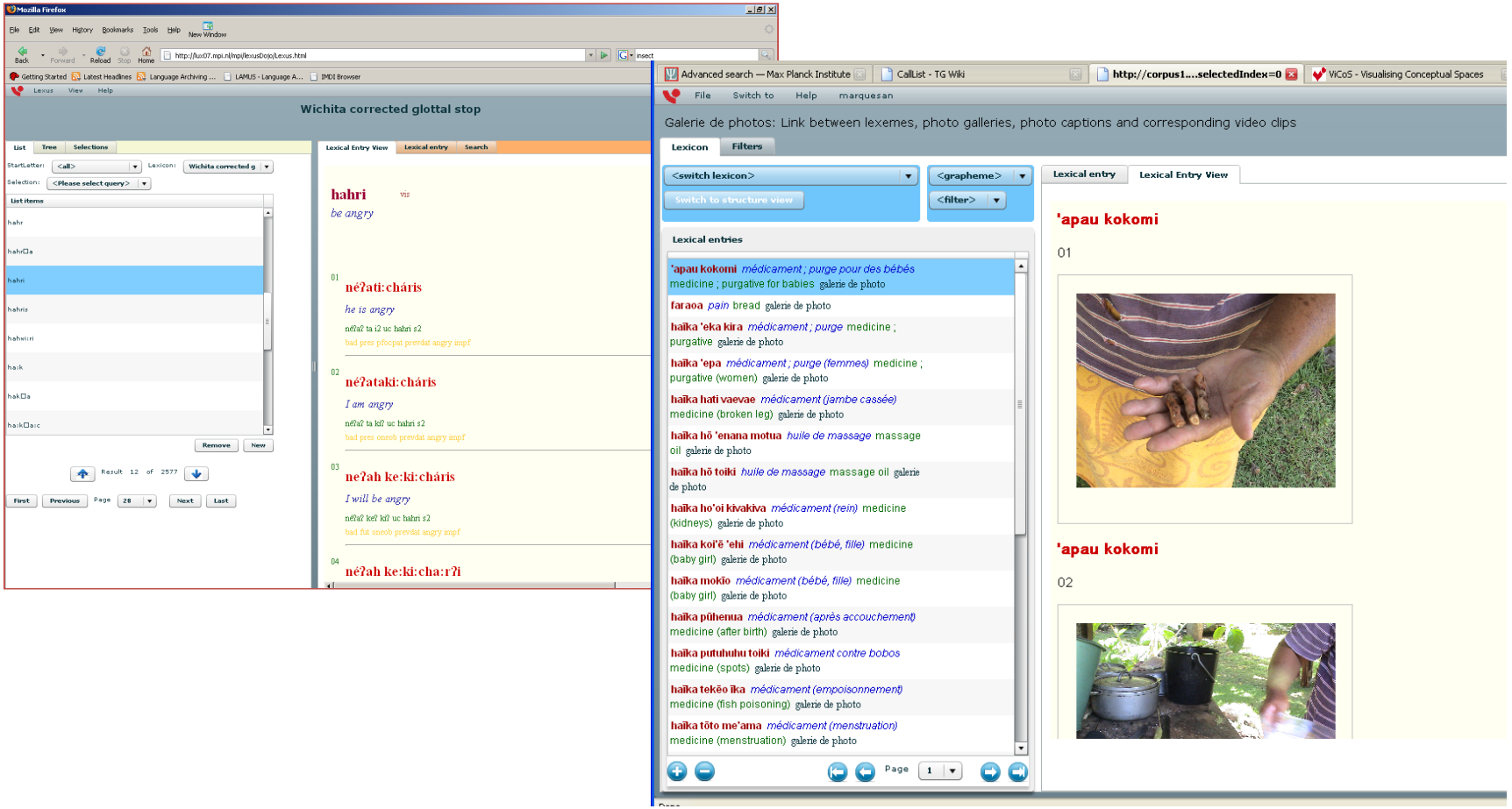
CLARIN WS

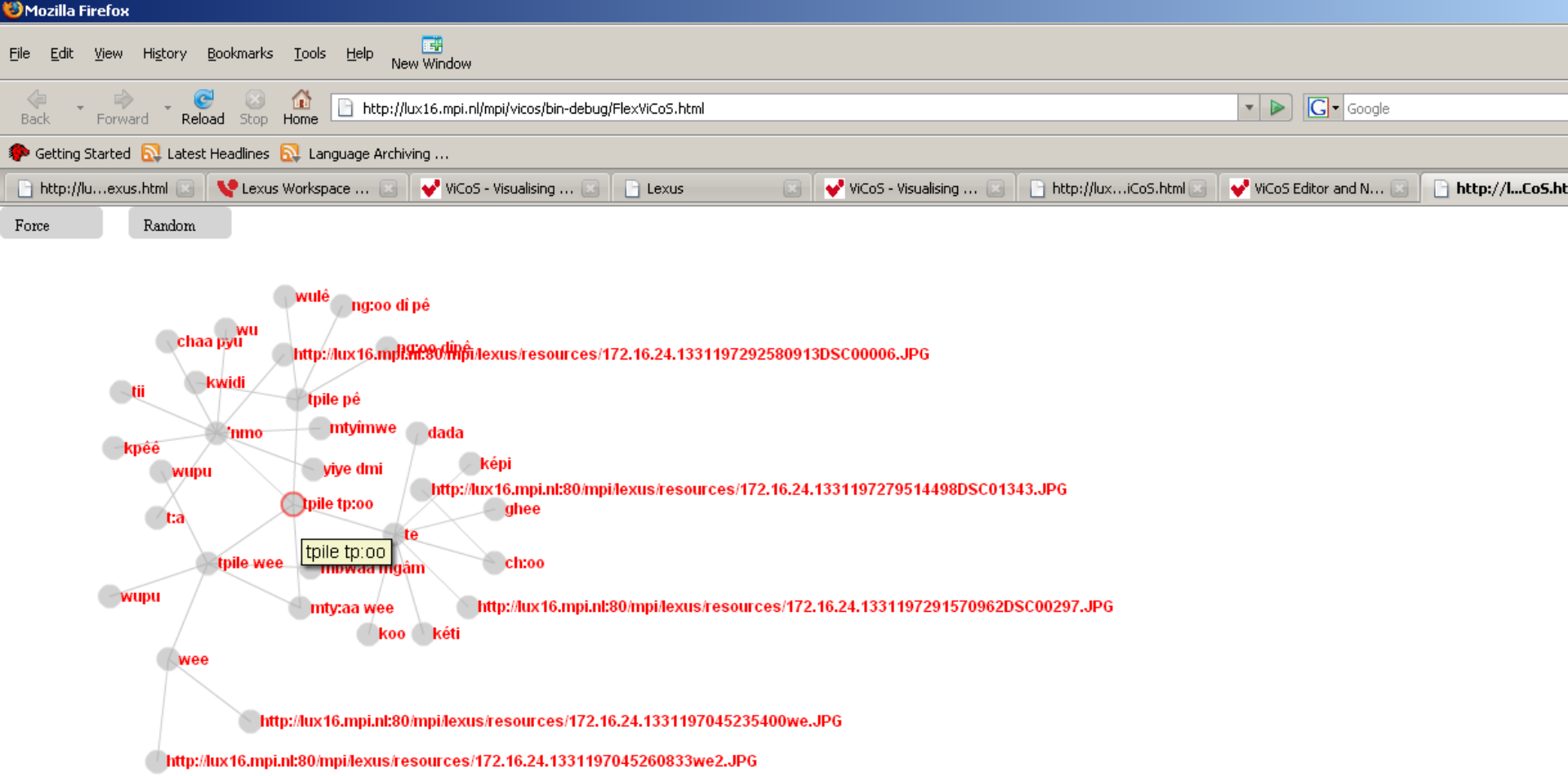
ELAN Annotation Tool (ANNEX)



LEXUS Multimedia Lexicon Tool

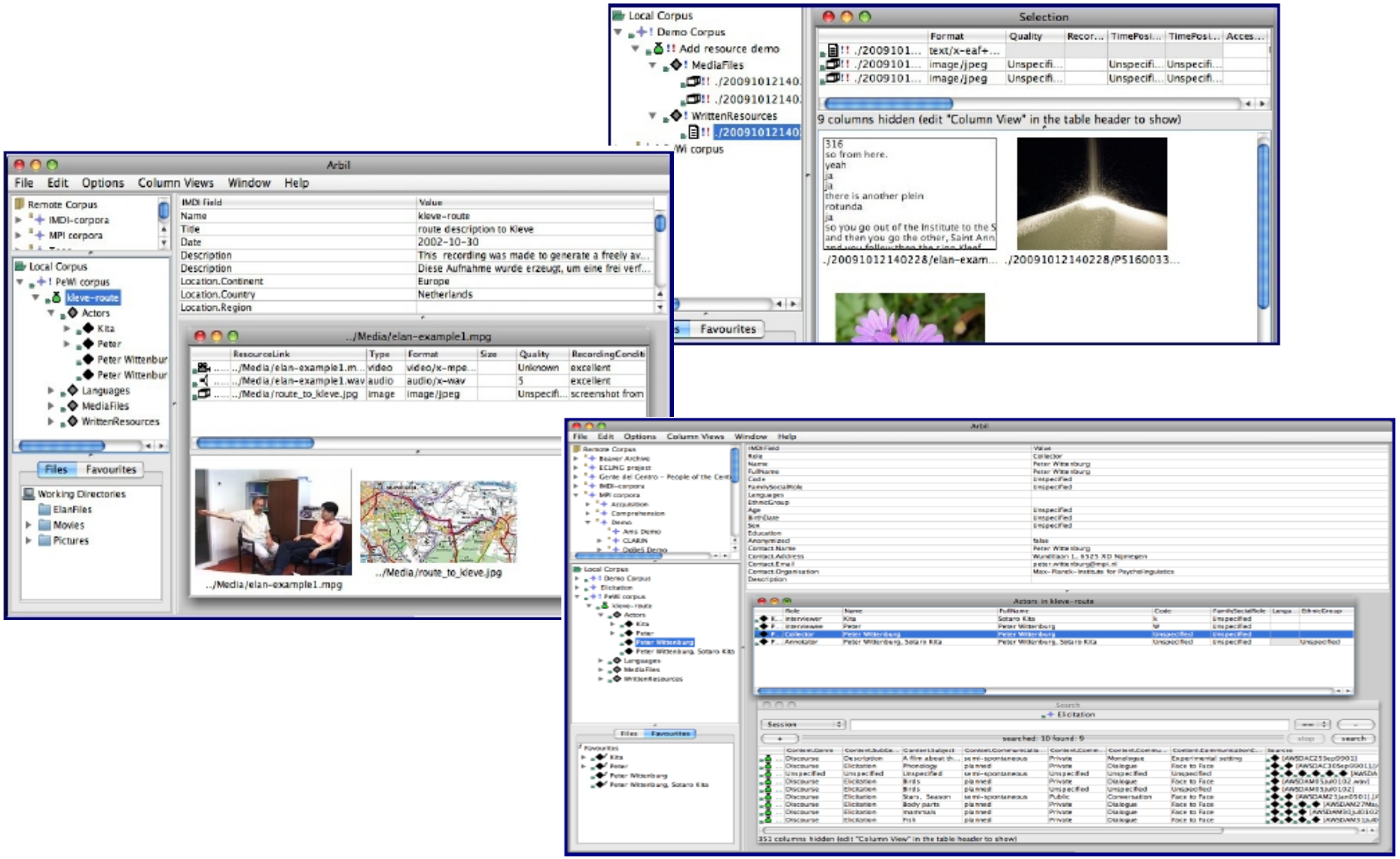
- Creation of lexica from scratch, import lexica from other formats (Toolbox, XML, Chat)
- User definable views of word list and lexical entries



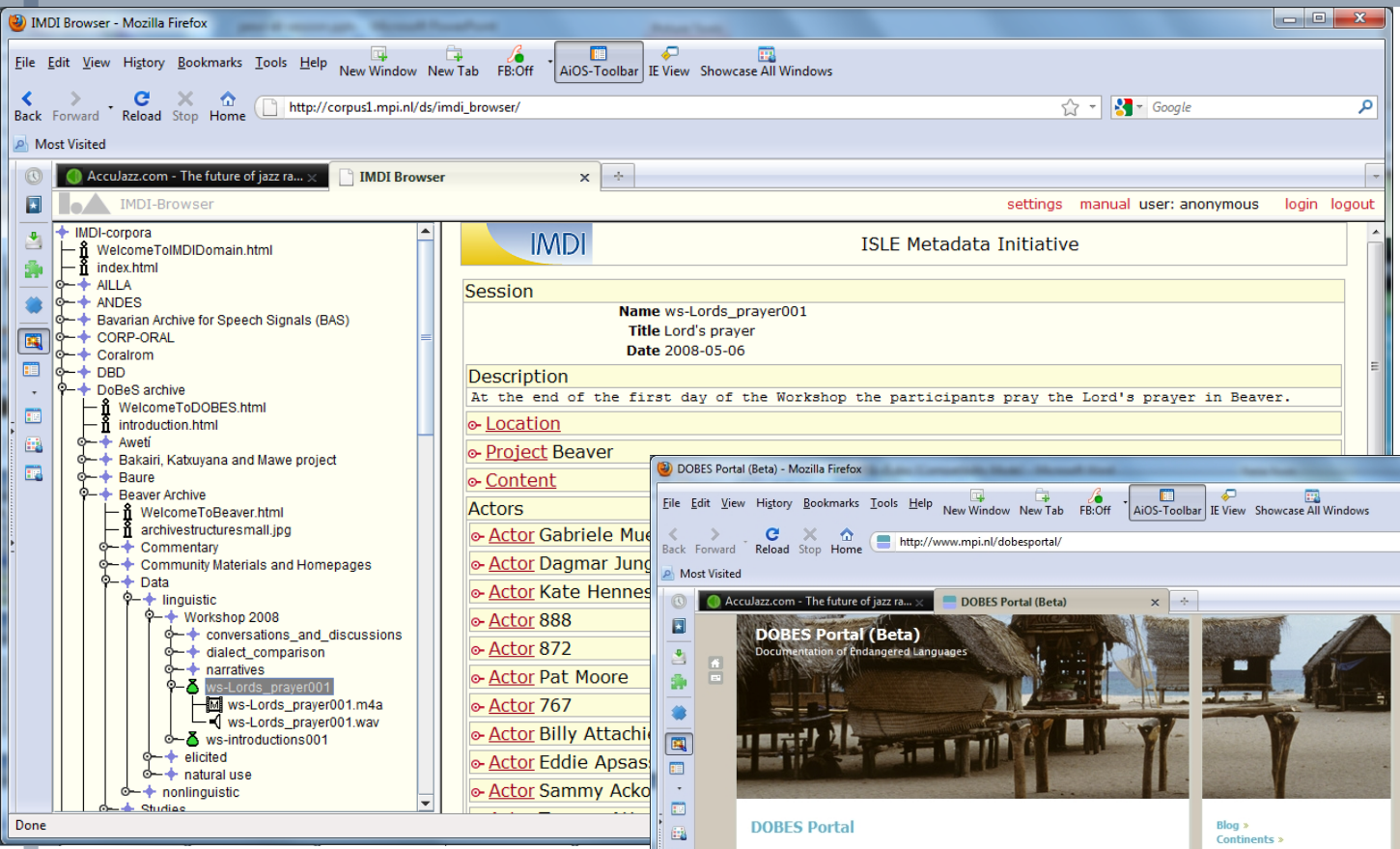


VICOS
Conceptual Spaces

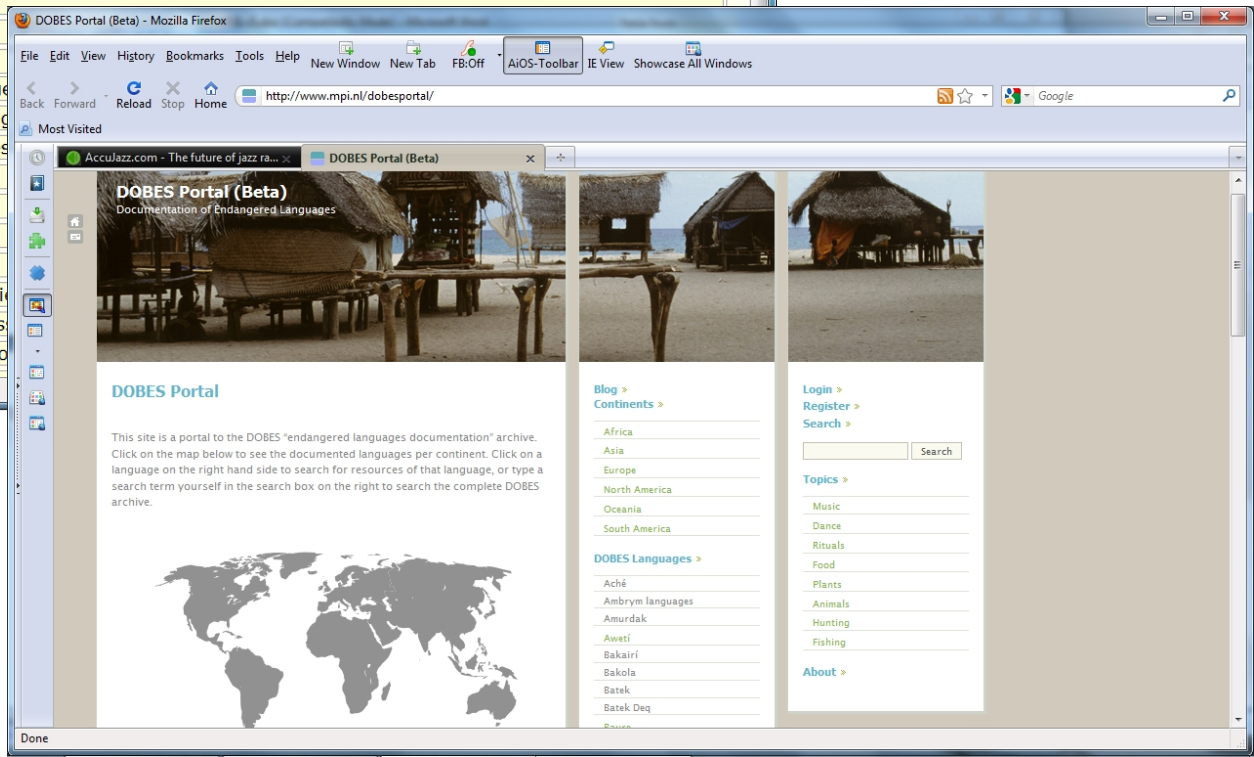
ARBIL: Metadata Editor & Organizer



Access to Archive

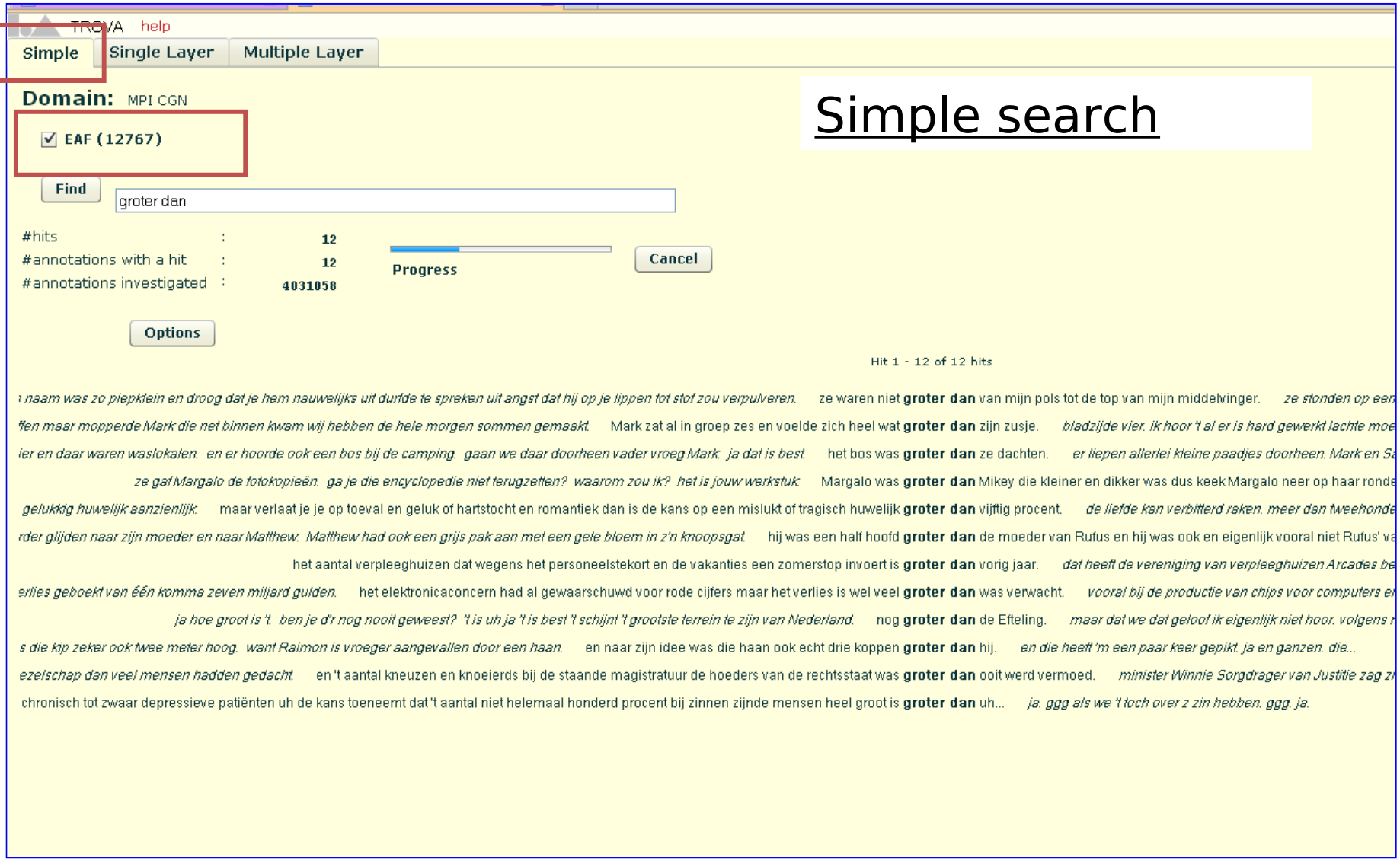


good old browser and search vs.



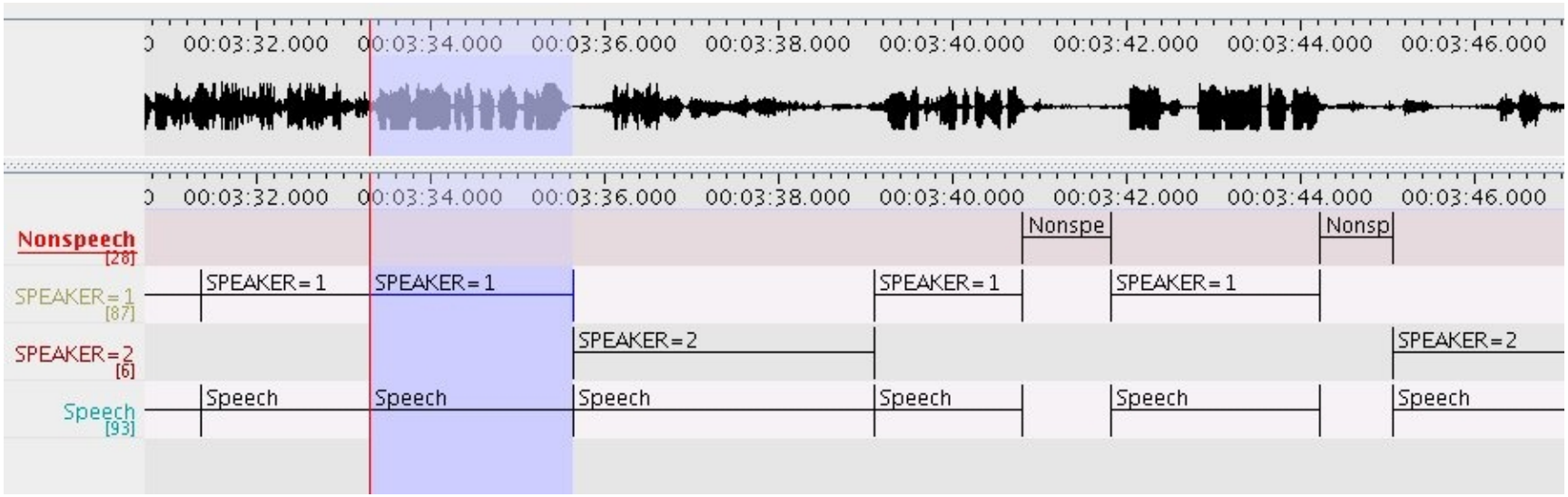
attractive community portals

TROVA Search Engine (ELAN/Archive)

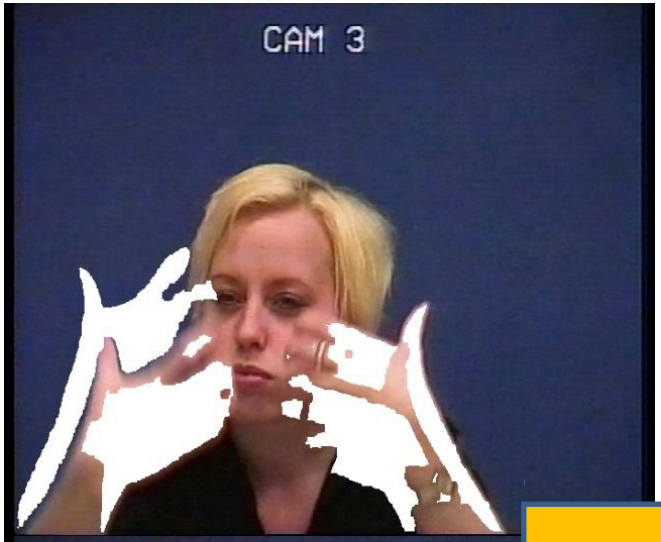


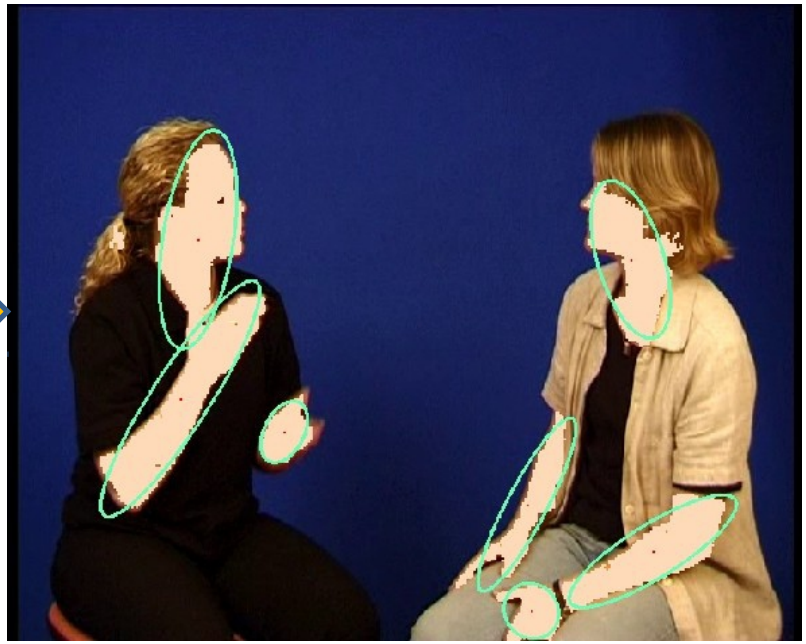
Simple search

Speaker Clustering



Skin-color Detection





Other stuff in Archive?

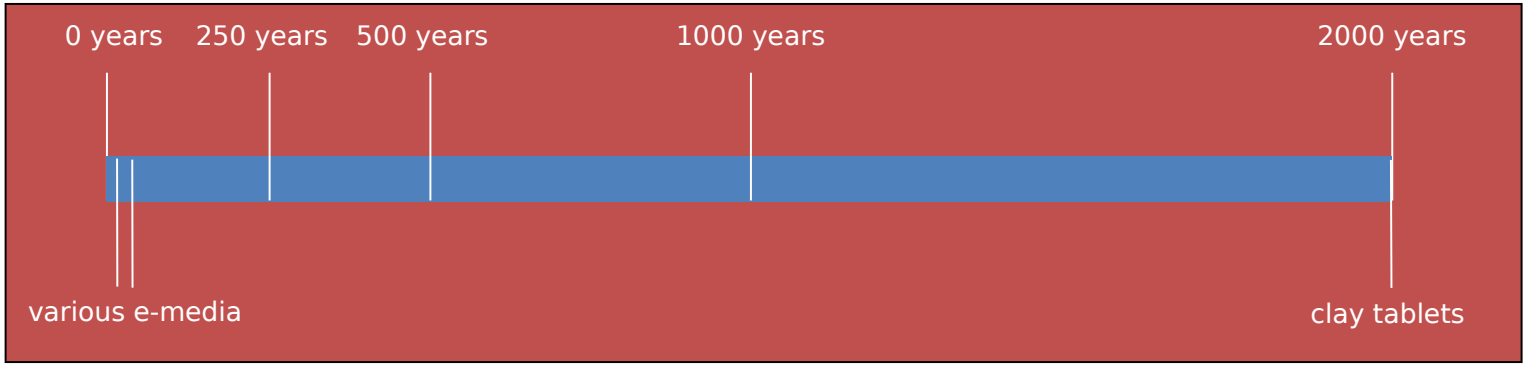
- most of the data in the archive is compliant with open standards and we do checks at ingest (JHOVE library and own checks)

- but science is dynamic – continuously new formats and proprietary stuff
  - Word, Excel, etc. (unconstrained)
  - encapsulation (databases – what is an object?)
  - Matlab files, etc. (no standards etc)

- need to have two separate branches in archive
  - guided, controlled, curated
  - unguided, uncontrolled, non-curated

Long Term Preservation

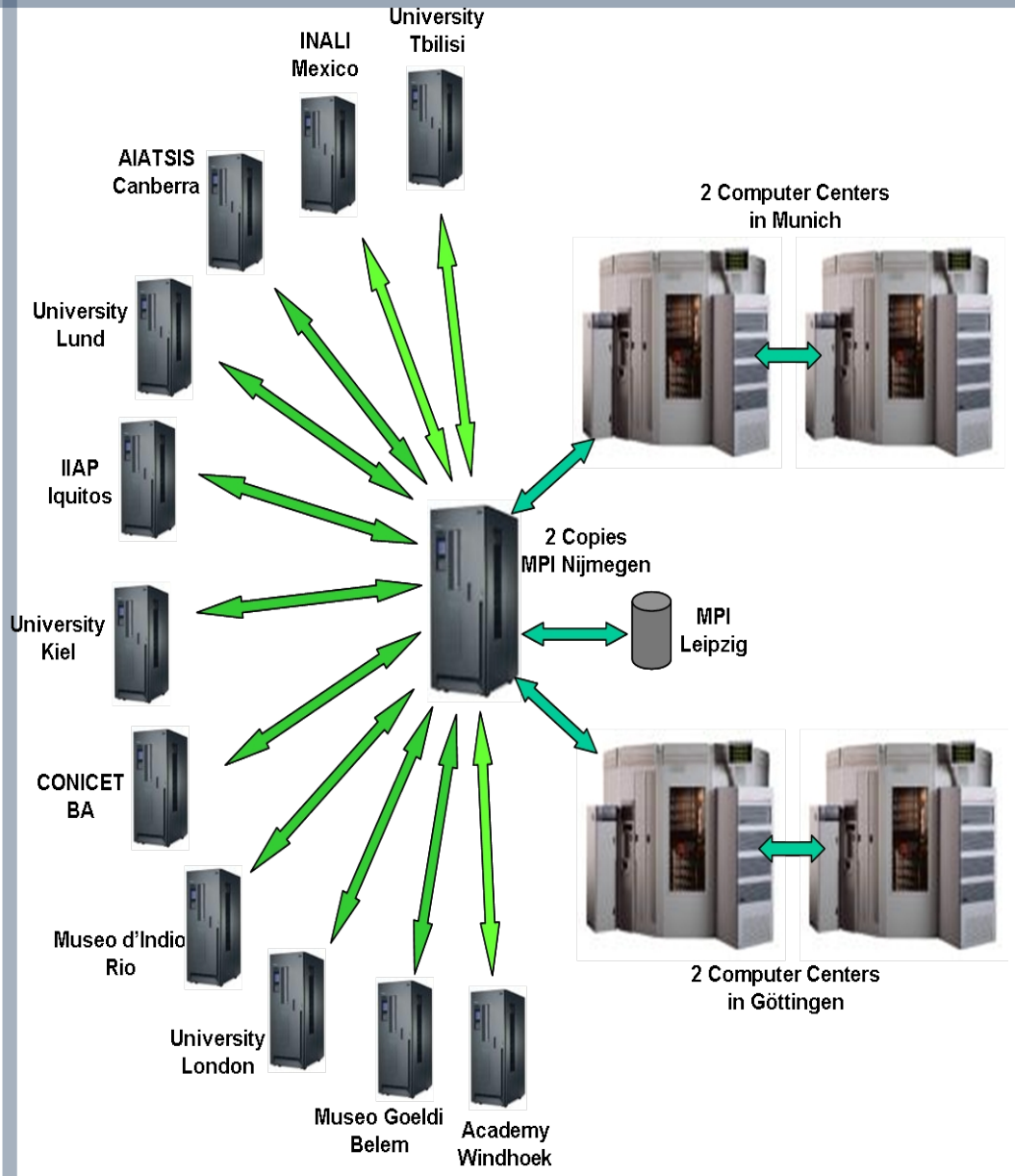
- bit stream preservation



- 80% of all language and culture recordings are endangered due to deterioration of carrier substrate
  - for logistic reasons much data will be lost for ever

- two strands: carrier migration and replication
- migration: every 4 years almost all hw (except TL mechanics)

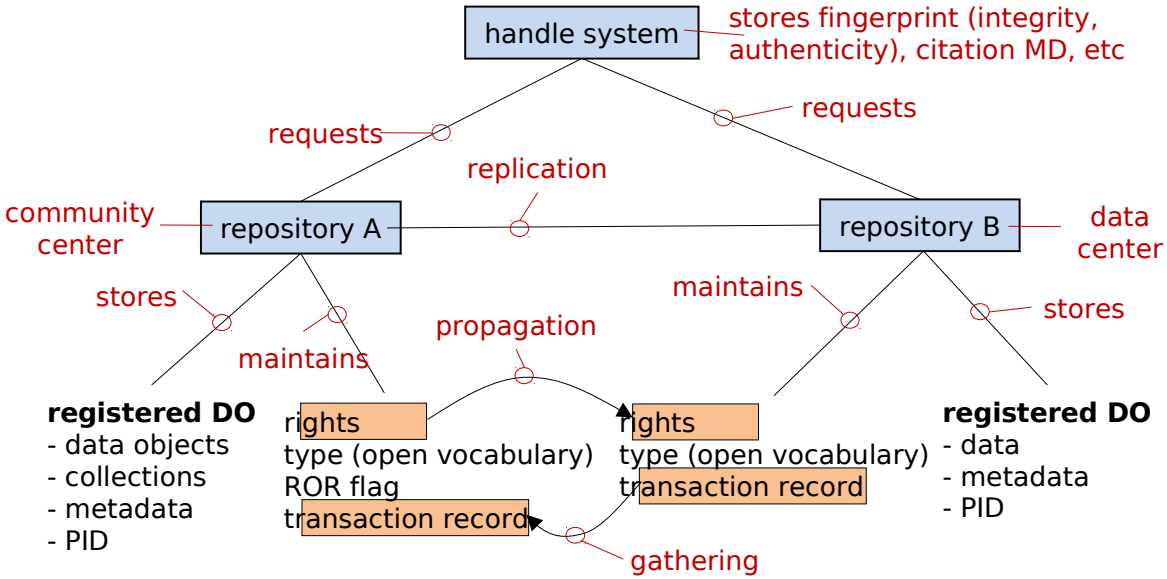
Data Migration



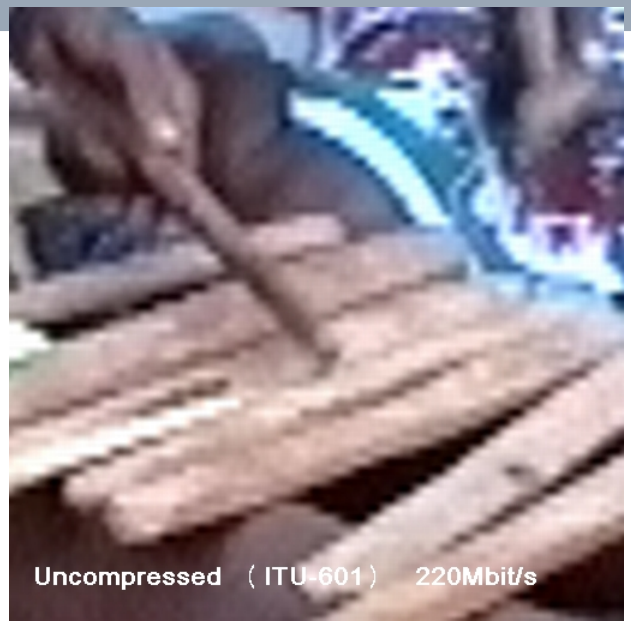
- stable, robust, organized and coherent online archive with 75 Terabyte of resources
- all metadata described and all associated with PIDs
- 4 full dynamic copies at remote CC with 50 years guarantee
- in addition 11 regional repositories with more to come
- open deposit service

Safe Replication

- safe replication between CLARIN center and RZG data center
- purpose: preservation, computation (AV Recognition) and access optimization
- total amount: 80 Terabytes
- requires policy rule based approach due to quality assessment (Data Seal)
- iRODS, Handles, CMDI Metadata
- deployment of Archive/Access software stack as well



Codecs and Curation Challenge

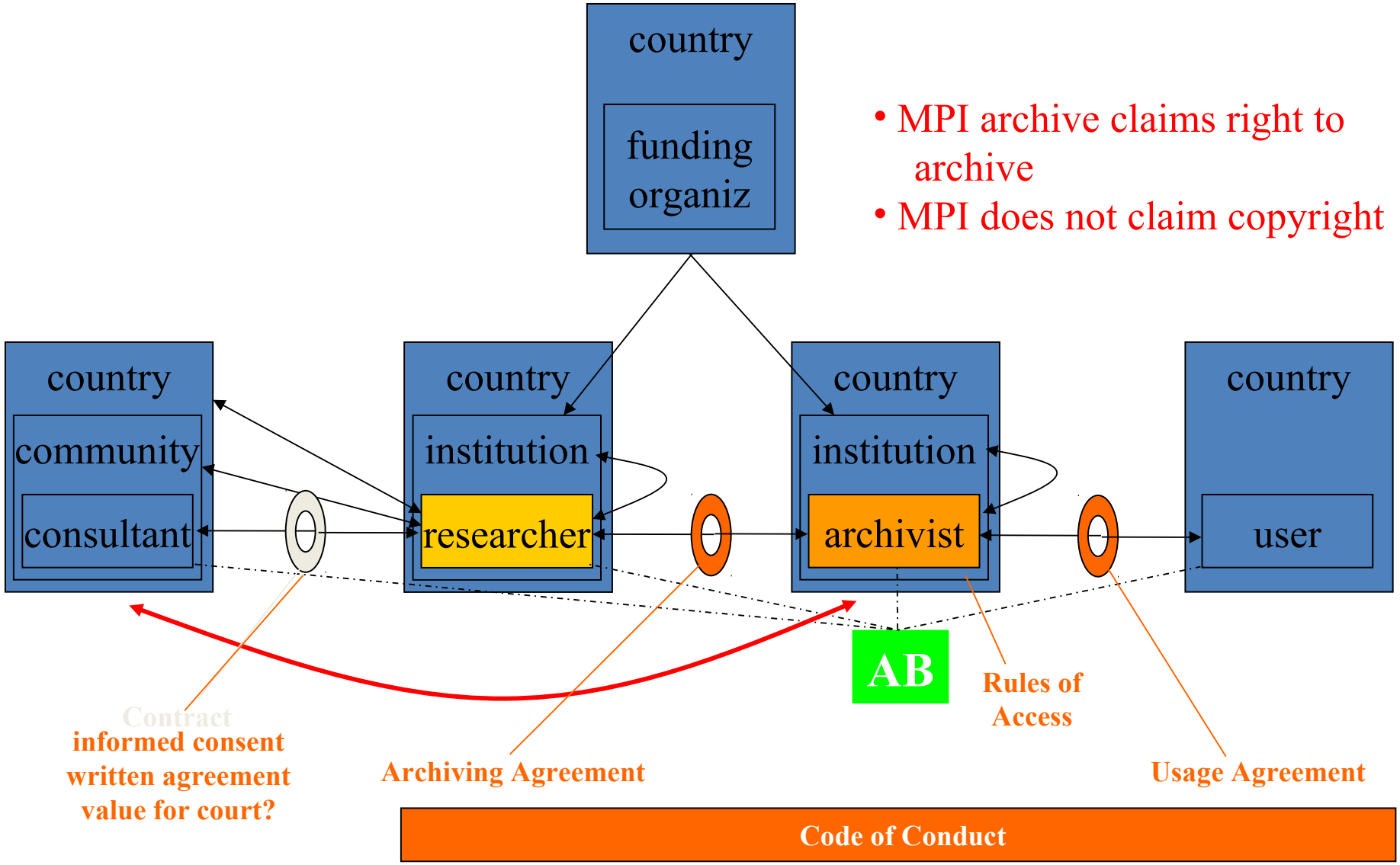


- highly compressed
- highly special
- what about authenticity?
- H.265 uses texture replacement

- can you go back (concatenation)

- so curation of digital formats can be challenge

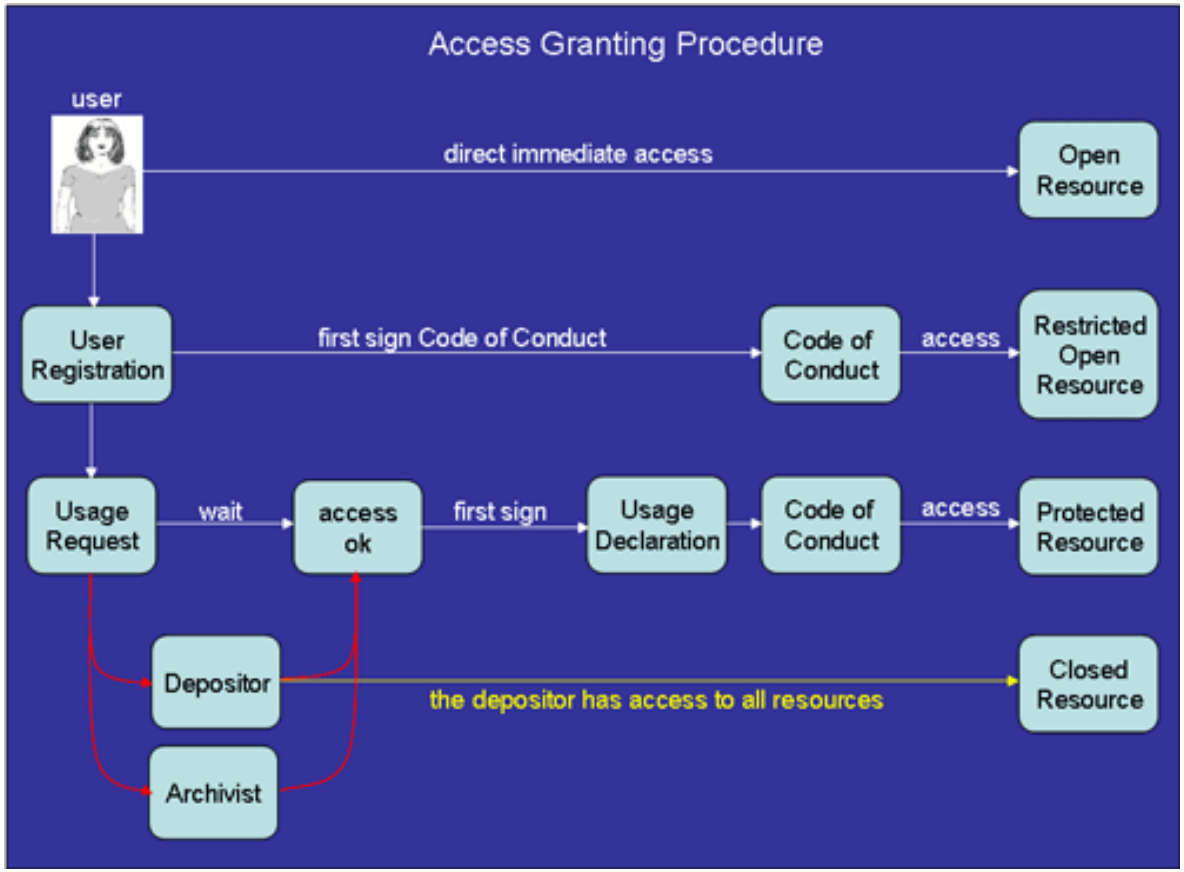
Rights Problem



- MPI archive claims right to archive
- MPI does not claim copyright

Access Levels

all support Open Access principle, but there are many obstacles such as IPR, copyright, ensuring dissertations, data as private capital, etc.



Data & repositories in CLARIN

hat does a research infrastructure such as CLARIN do to
crease possibilities of sharing and re-using?

What is CLARIN?

LARIN is an electronic/Internet-based Infrastructure bringing linguistic resources & tools virtually together making them virtually available to interested users

some keywords

- aggregation of metadata for visibility
- storing & curating data for accessibility & usability
- managing permissions for accessing
- allow interpretation (syntax, semantics)
- allow re-use (understanding, purpose, etc.)

eal world:

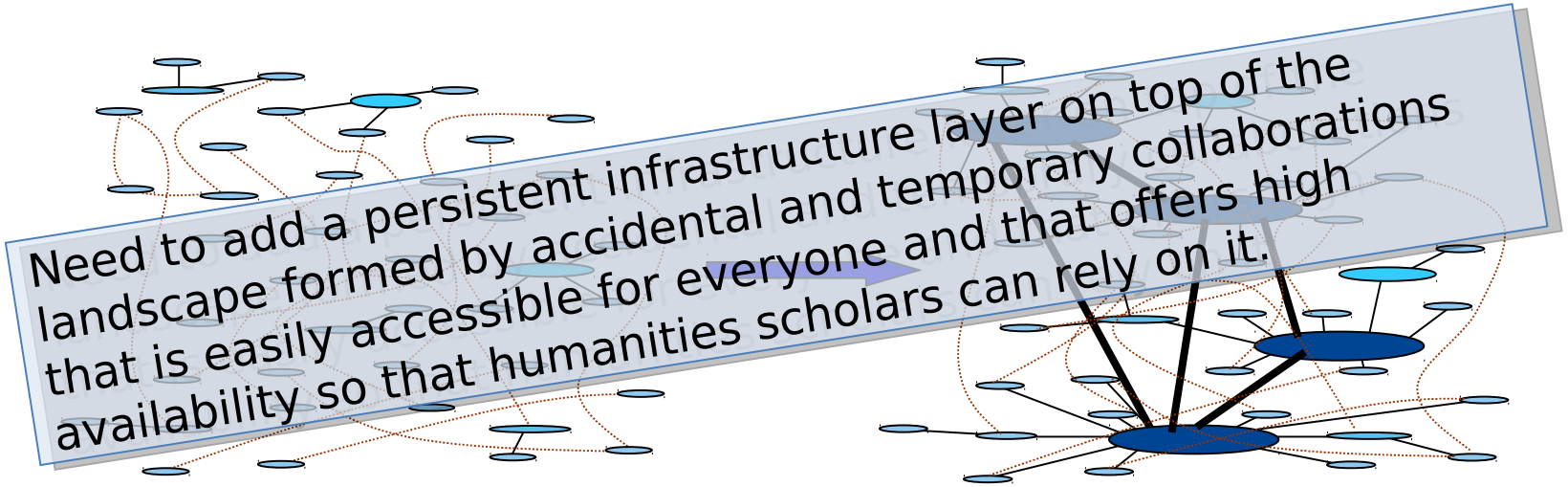
eating a framework where data and tools form an tegrated and interoperable domain allowing users to make se of all components without barriers

Restructuring in CLARIN

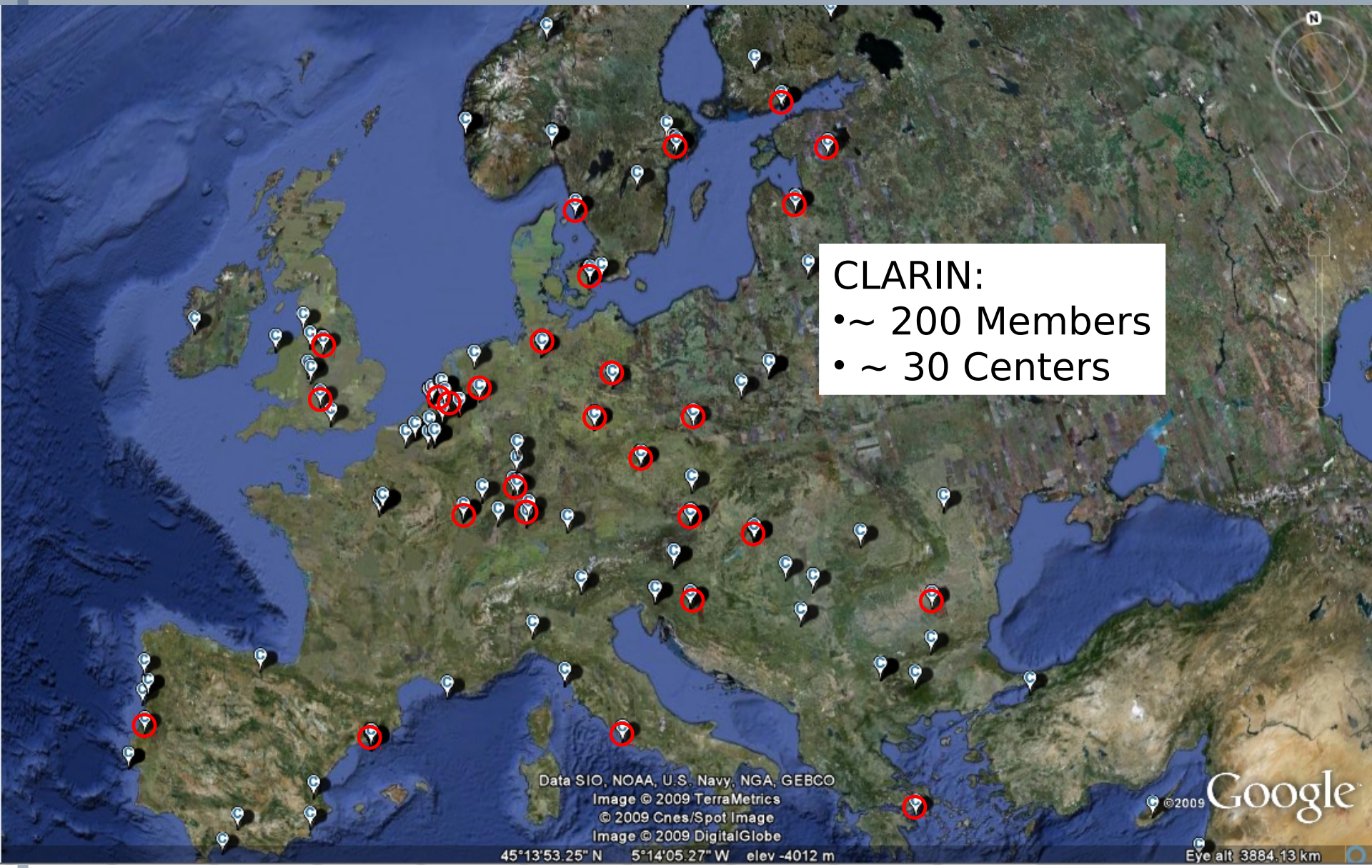
do we need centers (hubs) and what would be their role?

resources & tools

- are created in a completely distributed manner
- would remain fragmented without hubs with responsibi
- would be inaccessible/un-interpretable without storing, curation and management effort
- would become inaccessible without an infrastructure



Repositories at the basis of CLARIN



CLARIN Overview

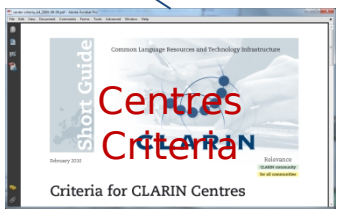
25 Centre Candidates

all are busy with restructuring plans

2 already give long-term preservation service

how to come to a persistent and stable	how to come to a federation and how to	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

CLARIN Centres



Centres Criteria

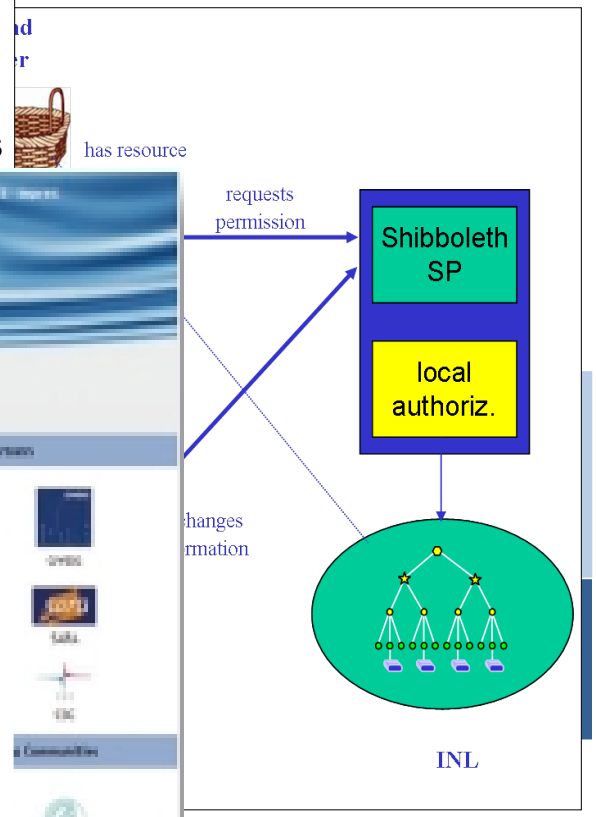
Long-term Preservation

REPLIX Replication

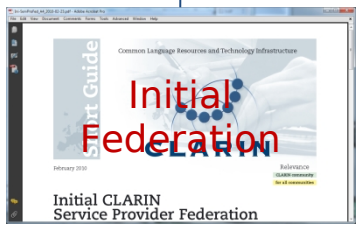
CLARIN Federation and PIDs

- Service Provider Federation
  - Agreement 1
  - n centers members

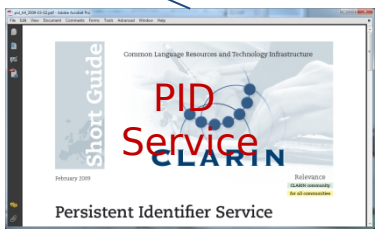
http://www.pidconsortium.eu



Trust Domain

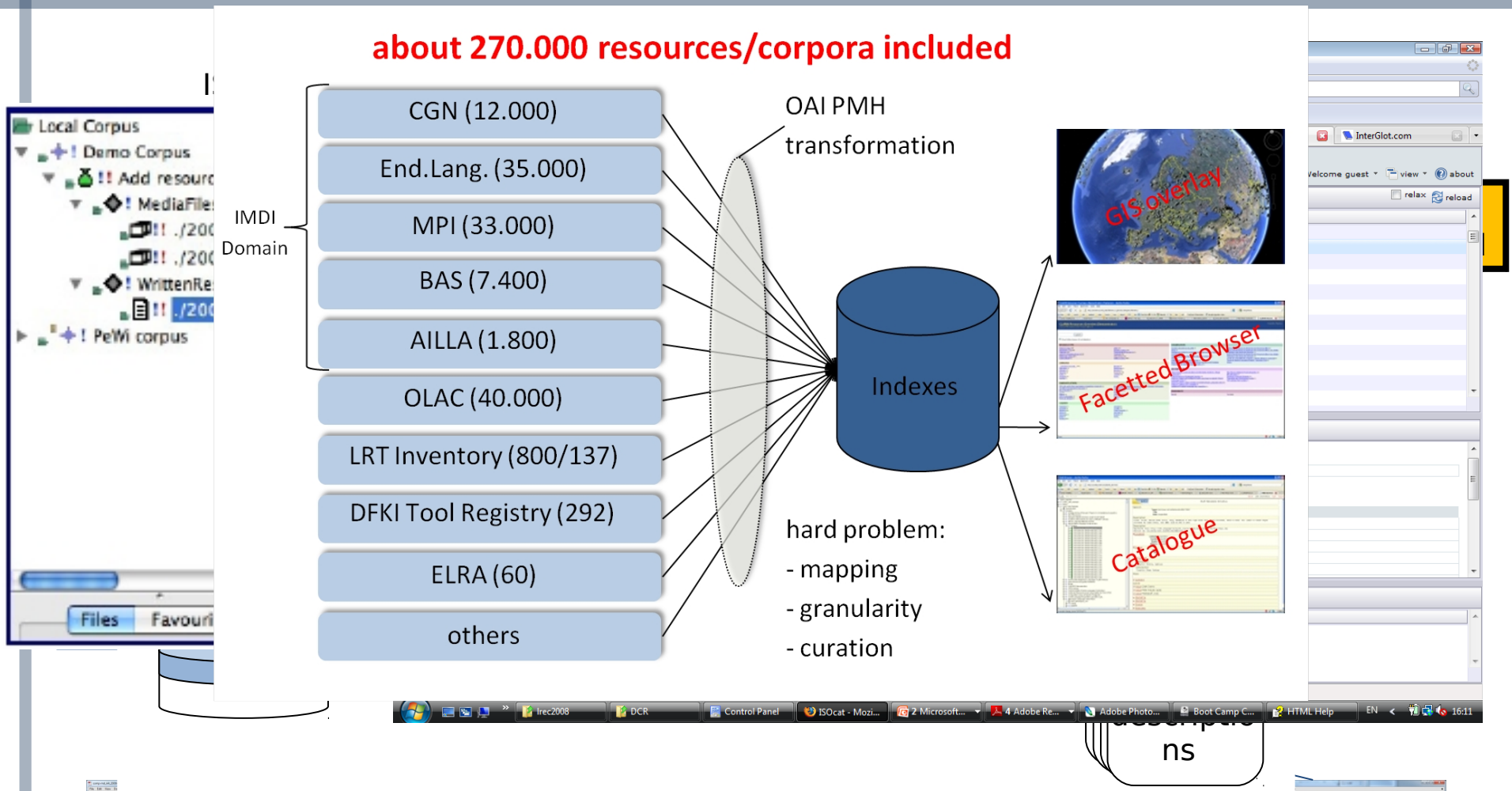


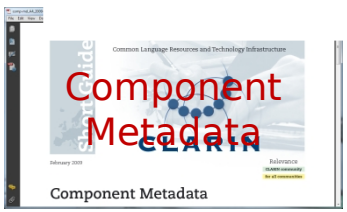
Initial CLARIN Service Provider Federation

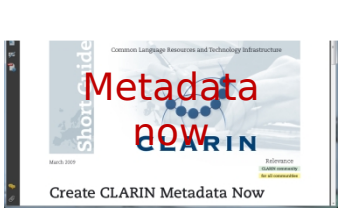


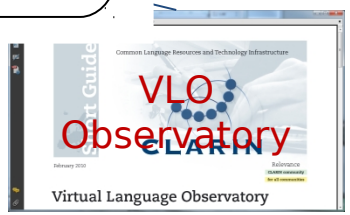
Persistent Identifier Service

CLARIN Metadata



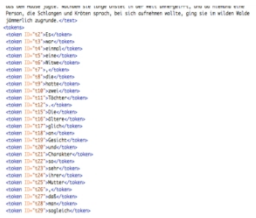






CLARIN Web Services

Stuttgart



Standard-conforma
Text Corpus Encod

Stuttgart Tübin

WebLicht: Web-Based Linguistic Chaining Tool

Tool Filters 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

Input Help

Build Chain

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

Add » Clear Run

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

Results

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

View As Table Download... 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>
```

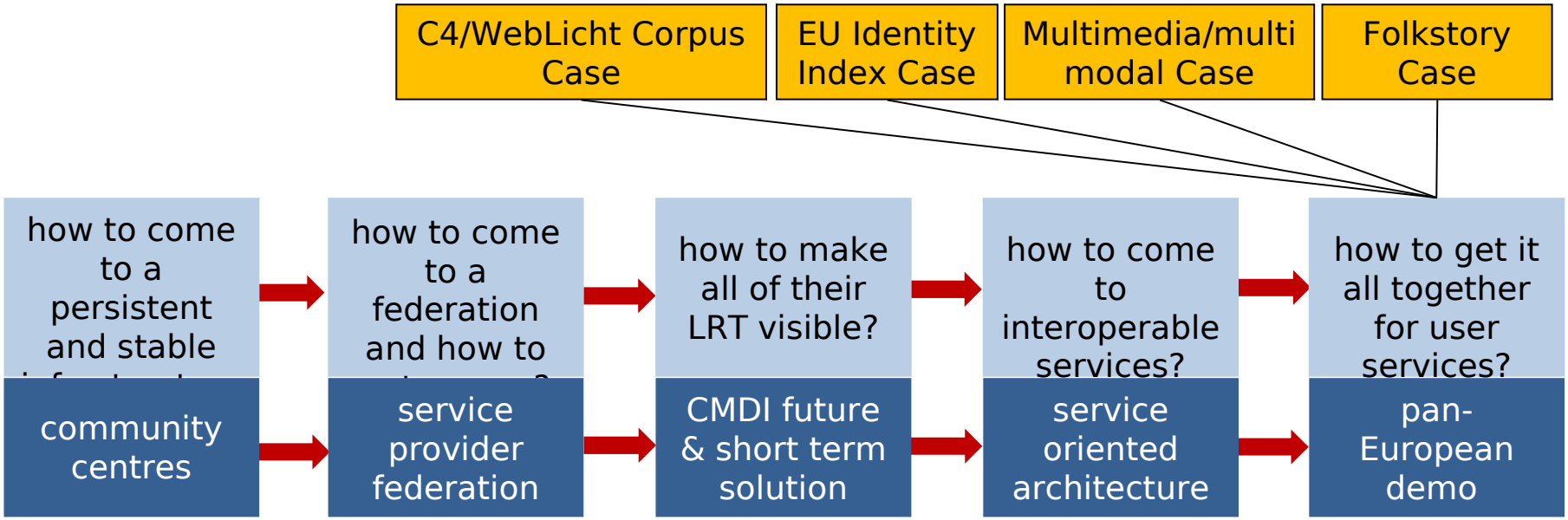
03EB1069277549D4AF3C0A90257FA856

Service Oriented Infrastructure

Web Services Interoperability

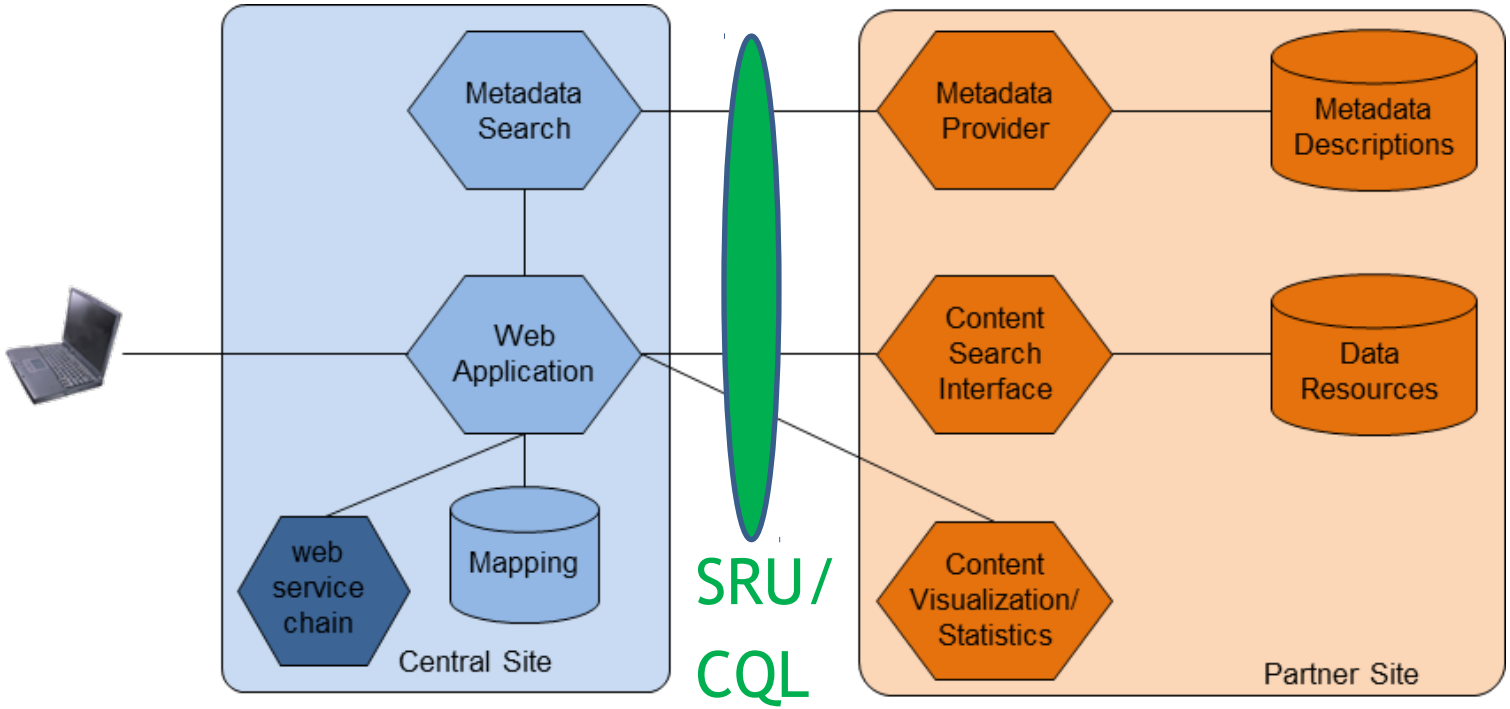
Standards and Best Practices

CLARIN Demo Cases



Distributed Search

- well Metadata is obvious -> Virtual Language Observatory
  - harvesting and mapping is not the problem
  - bad quality is the problem (as for Europeana etc.)
- planned is f.e. distributed content search



Center Types in CLARIN

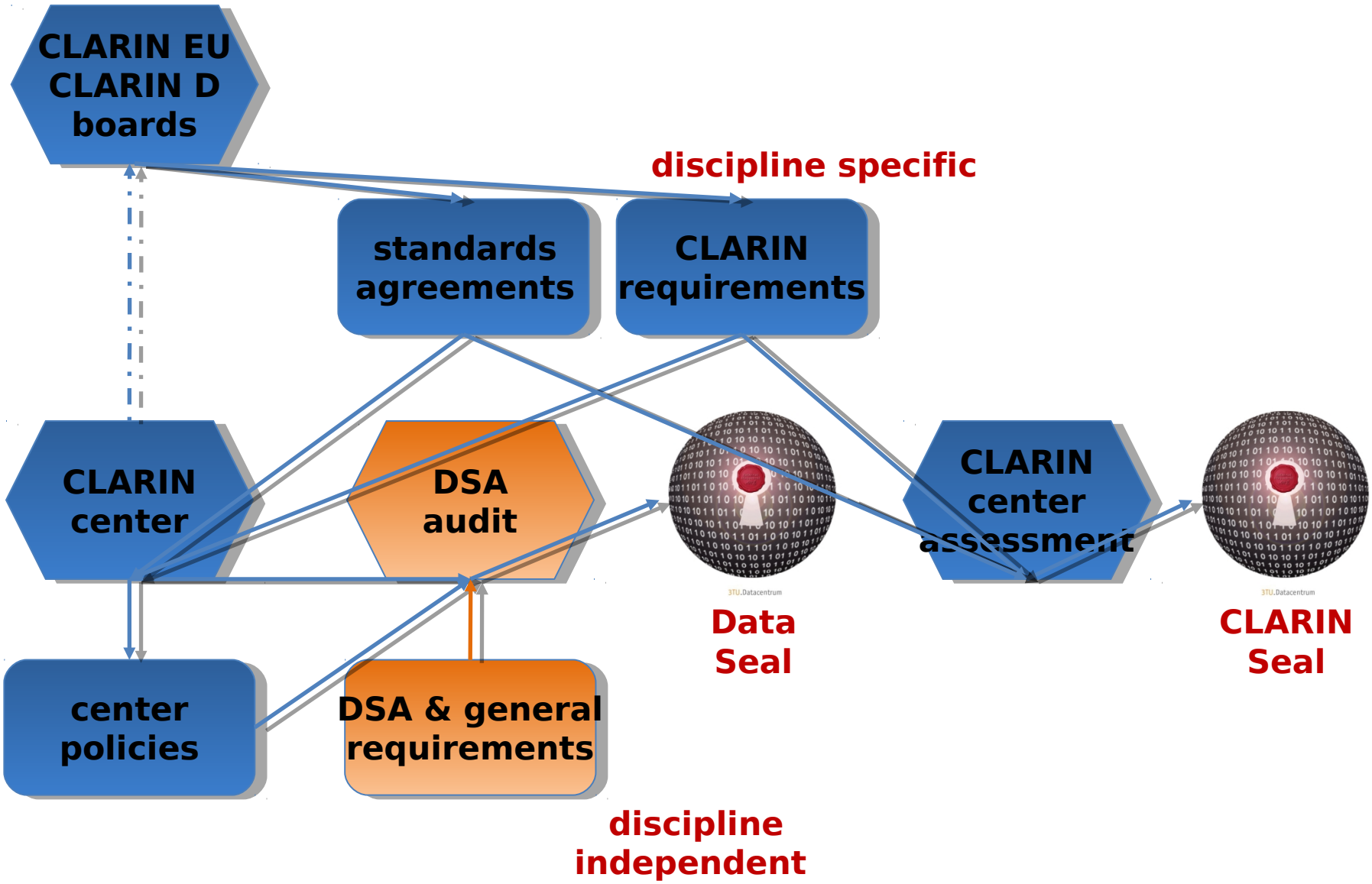
distinguish between center types

1. Recognized Centres (Type R) offer resources and
tools via standard web sites lacking facilities and commitment;
2. Metadata Providing Centres (Type C) offer machine readable metadata in a stable and persistent way;
3. Service Providing Centres (Type B) offer services to access resources and tools via specified interfaces in a stable and
persistent way;
4. Infrastructure Centres (Type A) offer services relevant for the infrastructure as a whole;
5. External Centres (Type E) offer CLARIN

Requirements for Centers

enters need to offer useful services to the CLARIN community and to agree with the
sic CLARIN principles (own architecture choice, explicit statement about quality of
rvice, usage of persistent identifiers, adherence to agreed formats, protocols etc).
enters need to adhere to the security guidelines, i.e. its servers need to have accepted cer
enters need to join their national identity federation where available and be ready to join t
LARIN service provider federation to support single identity and single sign-on operation
sed on SAML2.0 and trust declarations.
enters need to have a proper and clearly specified repository system and participate in a
uality assessment procedure as proposed by the Data Seal of Approval or TRAC approache
enters need to offer component based metadata that make use of elements from accepted
gistries such as ISOcat in accordance with the CLARIN agreements, i.e. metadata needs
be harvestable via OAI PMH.
ch center needs to make clear statements about their policy of offering data and services
eir treatment of IPR issues.
ch center needs to make explicit statements about its technological and funding support
ate and its perspectives in these respects.
enters need to employ activities to relate their role in CLARIN to the research community i
der to guarantee a research based status of the infrastructure and allow researchers to
mbed their services in their daily research work.

Assessment in CLARIN

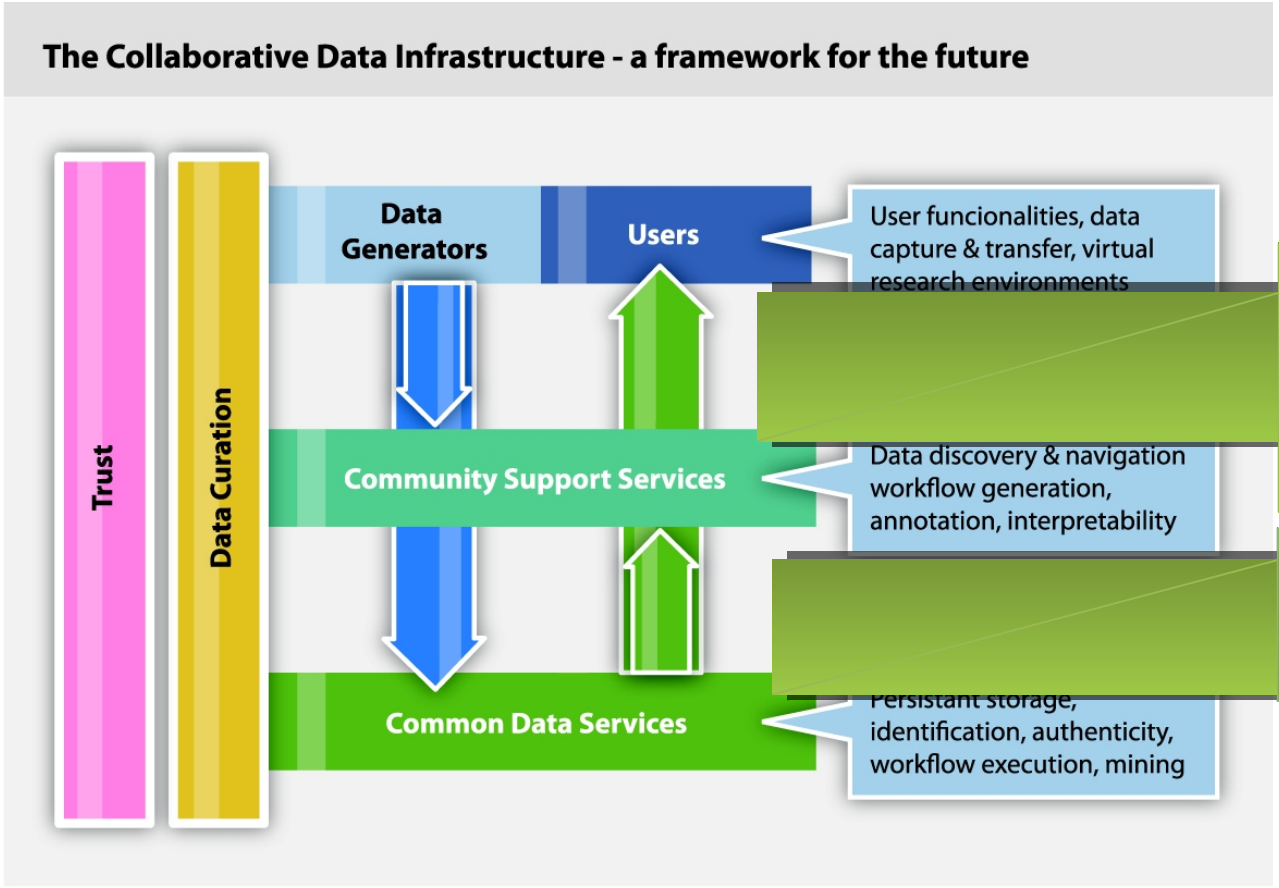


What about cross-disciplinary Initiatives

MAX-PLANCK-GESELLSCHAFT

What happens in cross-disciplinary projects wrt.
haring and re-using?
ake EUDAT as an example.
here are others: DataONE, DCF, OpenAIRE, etc.

Collaborative Data Infrastructrure

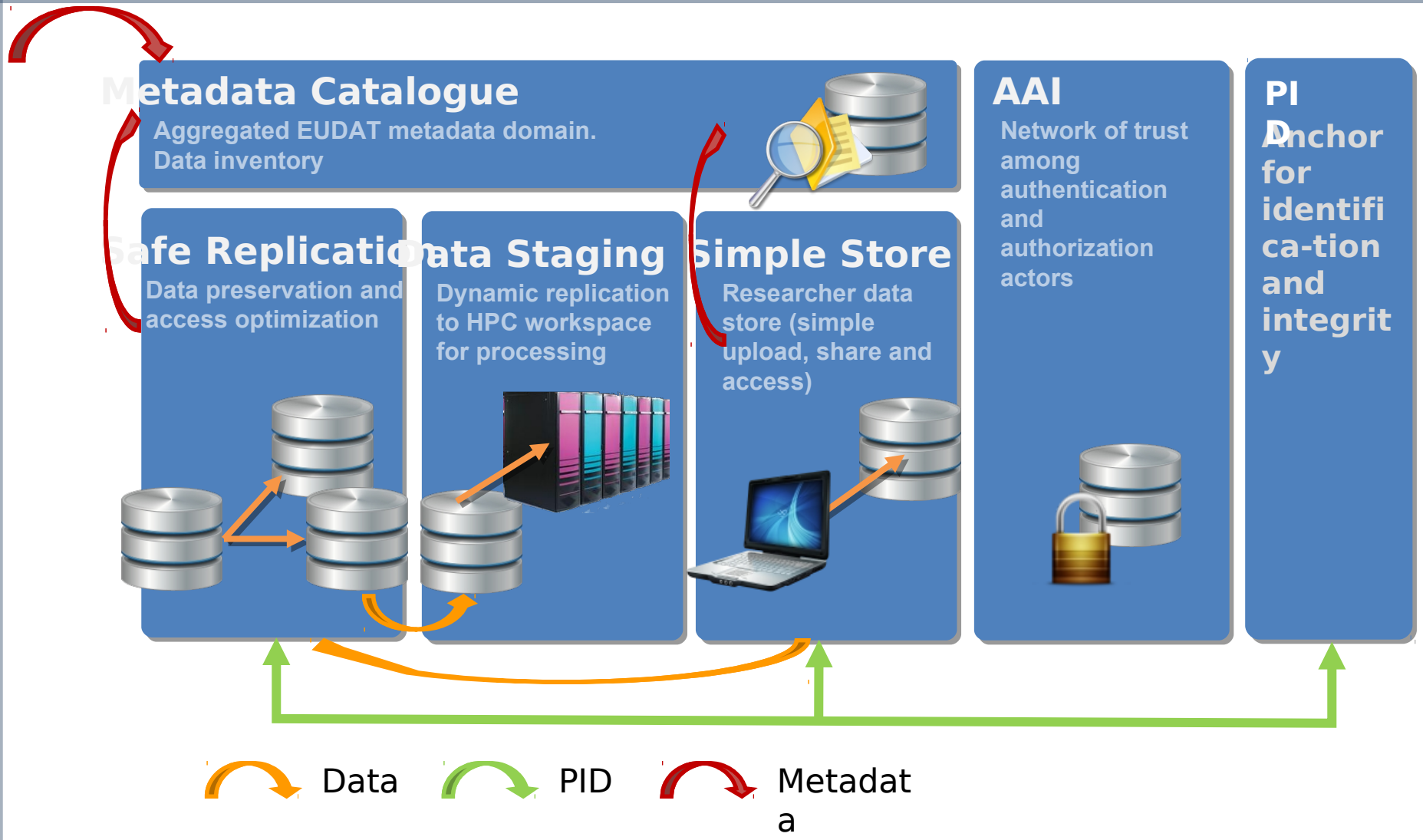


CLARIN, LifeWatch, ENES, EPOS, VPH, etc.
5 Core Infrastructures more second round infrastructures

=> 12 EUDAT data centers

Landscape in EUDAT

where are we in CLARIN EU



more to come: annotation, rt data, crowd sourcing, LTP & access, etc.

Cross-disciplinary work is hard

- all communities are working on defining or stabilizing their data landscapes
- big differences across communities and within communities
- just operating at the level of external objects is a challenge

- what about operating at content level?
  - domain of structural and semantic mapping
  - there is no golden way in science
  - but adhere to basic IT principles
    - use open standards
    - register your schemas
    - register your semantics
    - use PIDs
  - allow people to create and share their own relation sets

There is Research Data Alliance

- threats according to Alan Blatecky (NSF), Carlos Morais Pires (EC):
  - critical importance and the need to share data for next century science and education is not understood
  - urgency to address and create a global data infrastructure now is not understood
  - relying on more workshops, conferences, committees etc. to provide more recommendations
  - waiting on standards to be approved that will enable data sharing, interoperability and support data life cycle
- therefore
  - let’s start and do instead of talking and discussing
  - get a global layer of coordination to get things done -> RDA
  - have a simple and effective mechanism open for good ideas -> RDA
  - get out documents soon that are trusted -> RDA

Can RDA help

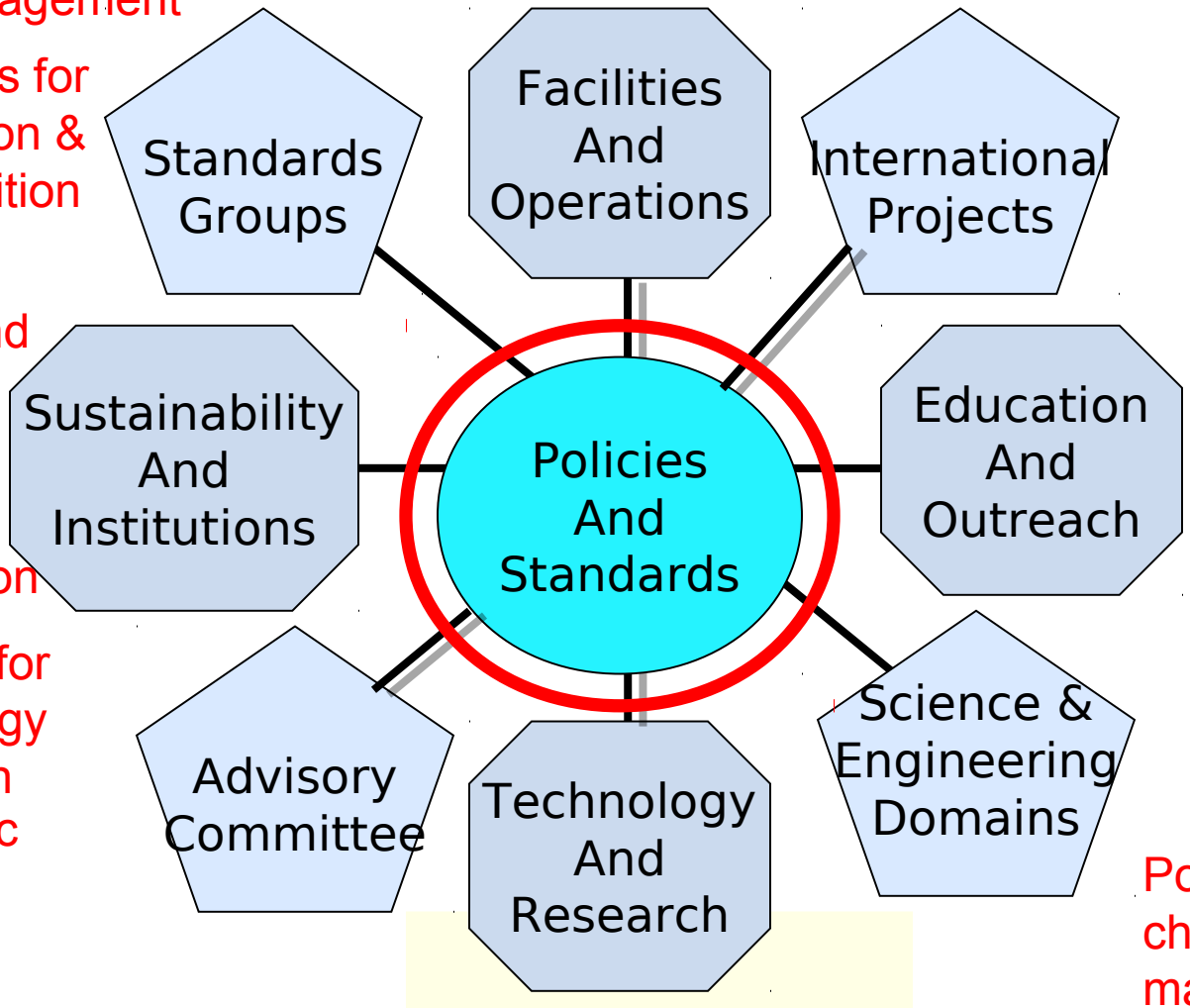
- RDA will have a great impact on cross-disciplinary enterprises as EUDAT
  - it is bottom-up and driven by ”data practitioners”
  - it’s focus is on removing concrete barriers on the way of sharing and interoperability – so it’s not another policy group
- I hope that RDA will also have implications on data organizations of communities
  - as usual – some argue that they solved the problems
- of course there are other important organizations we need to look at:
  - IETF — focus on networking
  - W3C — focus on the Web and its mechanisms
  - CODATA — focus on policies in area of data
  - World Data Systems — focus on proper data

Research Data Alliance WGs

- Data Foundation and Terminology (implies some agreed conceptualization)
- PID Information Type Harmonization
- Data Type Registry
- Practical Policy
- Metadata Normalization
- Pub/Data Citation/Linking
- Legal Interoperability
- Repository Audit and Certification
- The Engagement Group
- Marine Data Harmonization
- Defining Urban Data Exchange for Science

almost all group results would have an impact on EUDAT and simplify a lot

Policies in RDA (and in DCF in US)



ANDS in Australia

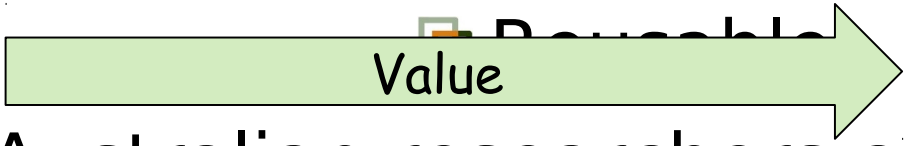
ANDS enables transformation of:

Data that are:

- Unmanaged
- Disconnected
- Invisible
- Single use

To Structured Collections that are:

- Managed
- Connected
- Findable
- Reusable



so that Australian researchers can easily publish, discover, access and use research data.

DataONE in US

Data Centers / Member Nodes

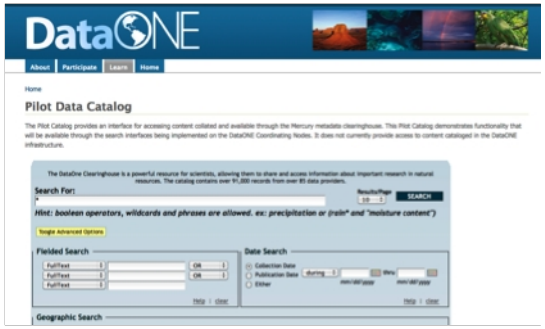
- USGS (BDP)
- LTER (EML)
- NCEAS (EML)
- ORNL DAAC (FGDC)
- OBFS (EML)
- IAI (FGDC)
- LP DAAC (DIF)
- LBA (FGDC)
- I-LTER (EML)
- TERN (EML)
- SAEON (EML)

Metadata Extraction

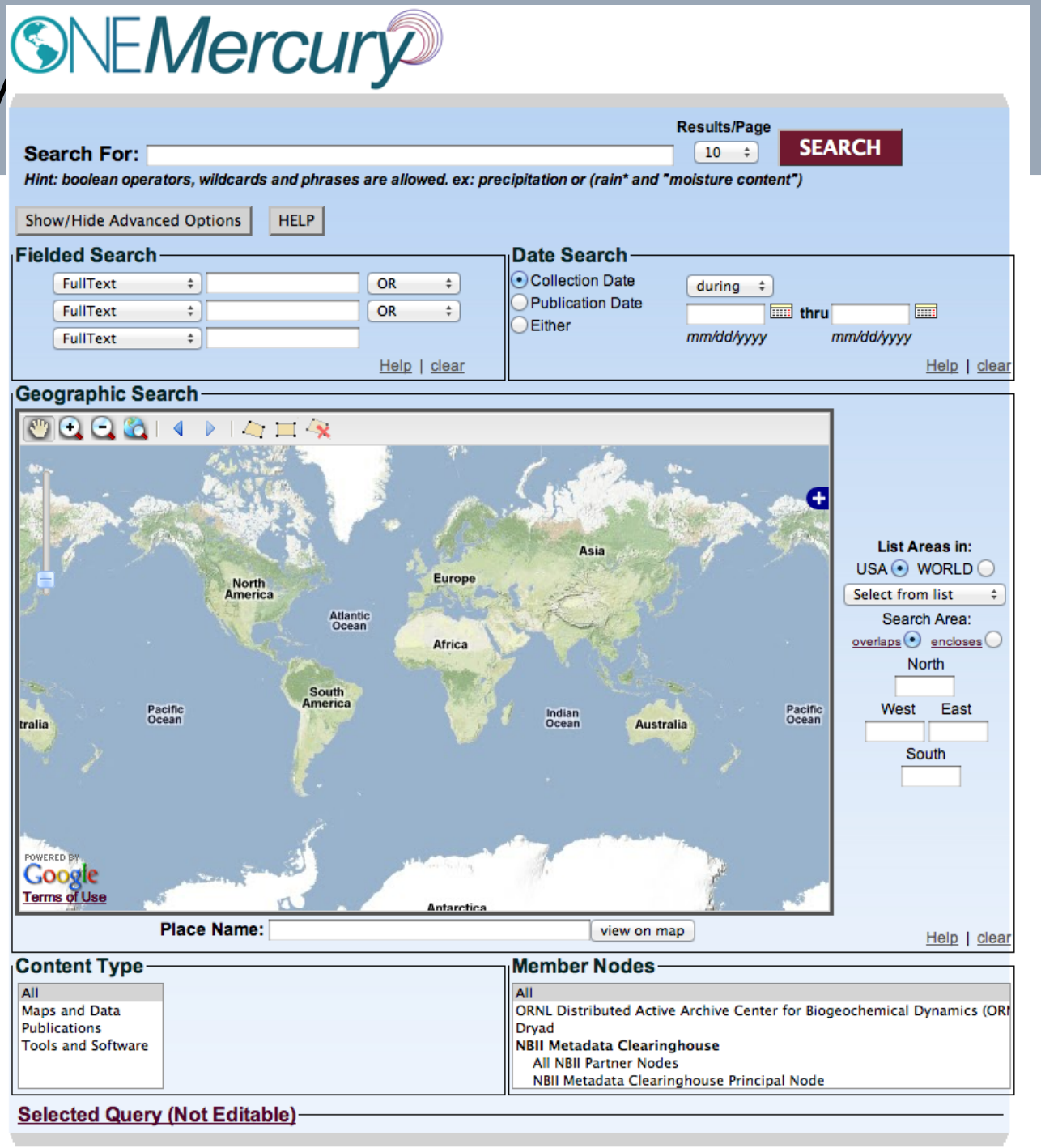
Catalog for Earth Observations

Internal Metadata Index

Stored at Coordinating Node



- Single portal
- Numerous search capabilities
- Search sharing functions (RSS, Web Services)
- Metadata has link to data, which reside at Data Center


ONEMercury
Results/Page
10
SEARCH
Search For:
Hint: boolean operators, wildcards and phrases are allowed. ex: precipitation or (rain* and "moisture content")
Show/Hide Advanced Options
HELP
Fielded Search
FullText
OR
Help | clear
Date Search
Collection Date
Publication Date
Either
during
thru
mm/dd/yyyy
Geographic Search
North America
Atlantic Ocean
Europe
Asia
Africa
South America
Pacific Ocean
Indian Ocean
Australia
Antarctica
Google
Terms of Use
List Areas in:
USA
WORLD
Select from list
Search Area:
overlaps
encloses
North
West
East
South
Place Name:
view on map
Content Type
All
Maps and Data
Publications
Tools and Software
Member Nodes
All
Dryad
NBII Metadata Clearinghouse
All NBII Partner Nodes
NBII Metadata Clearinghouse Principal Node
Selected Query (Not Editable)

Thanks for your attention.