



# EUDAT User Forum

## Session on Metadata & Simple Store Services

Daan Broeder, Erwin Laure

2nd EUDAT User Forum  
11, 12 February 2013

# Program

- 14.30-14.40 EUDAT Simple Store Service, Erwin Laure
- 14.40-14.50 EUDAT Metadata Service, Daan Broeder
- Community contributions:
  - 14.50-15.00 Metadata at ICOS, Timo Vesala
  - 15.00-15.10 Metadata at GBIF, Éamonn Ó Tuama
  - 15.10-15.20 Metadata at agINFRA, David Massart
  - 15.20-15.30 Metadata at BBMRI, Roxana Merino Martinez
  - 15.30-15.40 Metadata at CESSDA, Melanie Wright
  - 15.40-16.15 Discussion

## Goal of the Metadata and SimpleStore Session

At the EUDAT services part of EUDAT user conference, the second parallel session will have as subject “Metadata and Simple Store”. The Simple Store aims at providing researchers and citizen scientist with a simple solution to safeguard and share important data sets that would otherwise be in danger of being lost. The EUDAT metadata service wants to provide an interdisciplinary metadata catalogue that enables researchers to look for interesting data sets also in other disciplines. Especially in the last service it is important to have the research communities actively engaged and therefore we have invited representatives of different communities to give a short overview of the state of metadata infrastructure in their community and discuss with us some important questions.

- Does your community have a 'stable' metadata schema and infrastructure
- Can you see the benefits of an interdisciplinary catalog and if so can you name any other discipline from which you would like to use data sets
- What are important browsing dimensions & facets for catalog for your community



# EUDAT Metadata Services

Daan Broeder, TLA – MPI for Psycholinguistics

2nd EUDAT User Forum  
11, 12 February 2013

# Joint Metadata Domain

- Deliver a service for searching and browsing metadata across disciplines and repositories for users from different communities
  - Appropriate terminology for users of all disciplines when specifying queries
  - Access to the data when allowed
  - Useful visualization of results
  - Commenting facility to exchange experiences
- Aggregating metadata from different communities in one system gives rise to *semantic interoperability* and *granularity visualization* problems

# User (Community) perspective

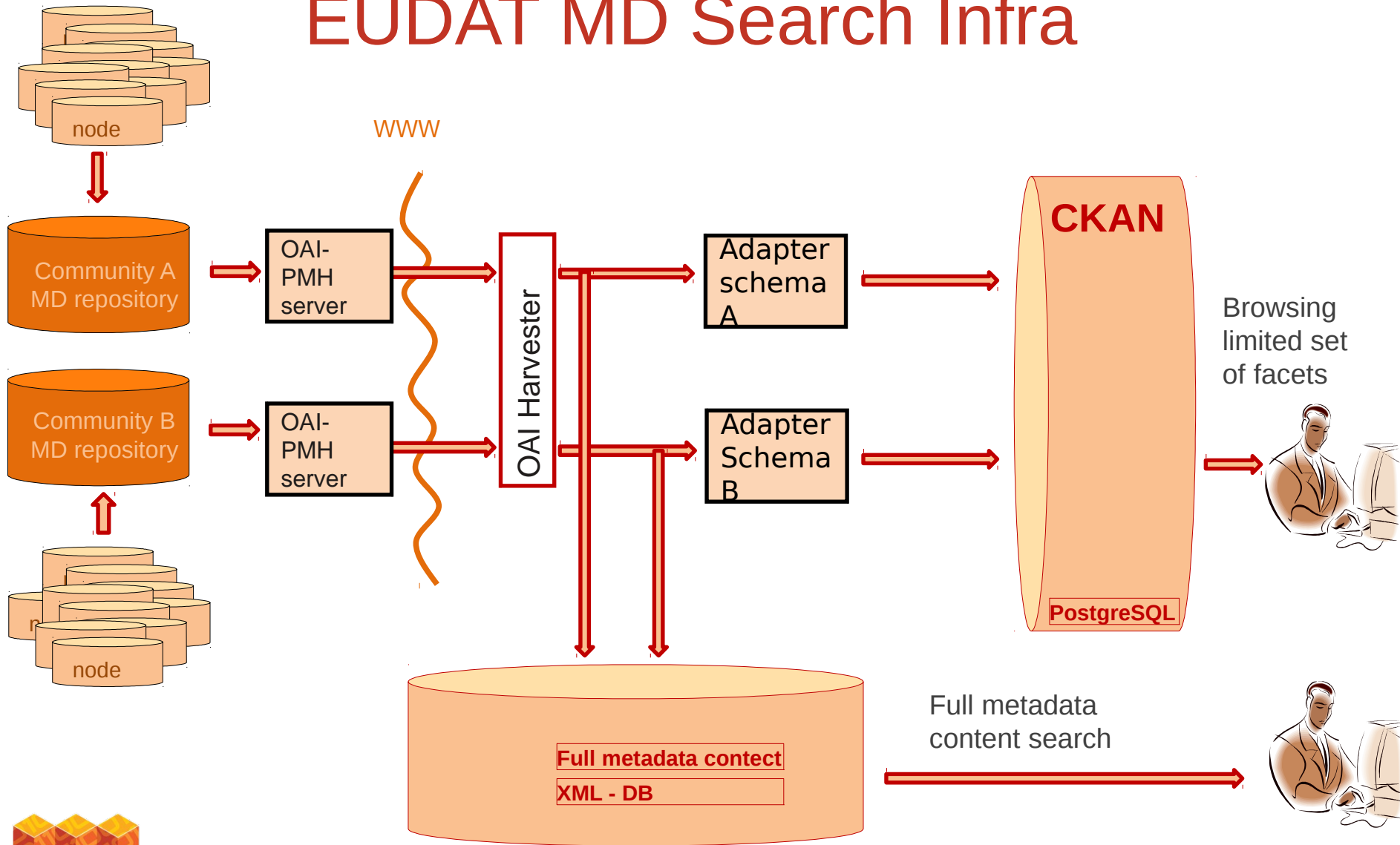
- From the user's perspective we think the following is important:
  - What are useful dimensions for searching & browsing?
  - How is the data presented e.g. what visualization options are available. e.g. geo-spatial, time line, taxonomies,
  - What metadata terminology can be used in the catalogue interface?
- Consider also the use for community's own purposes
  - Starting communities, that do not yet have set up a proper catalogue

# Interoperability

## Possible solutions:

- All repositories use the same metadata set
  - Not very practical. For sufficiently accurate description different (sub-)domains need different sets.
  - There are many existing metadata repositories with own sets & procedures that are difficult to change.
  - Already difficult to use one set only in a single repository
- Use a single set for exchange (->mapping by producer required)
- Use an interoperable metadata mapping framework
  - Hand crafted mappings
  - Ontology or concept registry based (community work required)
- Granularity problem is difficult to solve.
  - Filtering on metadata for data-sets versus metadata for single resources
  - Ranking search results, favoring the 'smaller contribution' communities

# EUDAT MD Search Infra





# EUDAT Metadata Catalog prototype I

- Tentative choice for CKAN as catalog software
  - Open Knowledge Foundation software
  - Choice made after a ‘limited’ comparison: large community, available documentation, proven track record
  - EUDAT will still be investigating other catalog software!
  - All should be modular & pluggable
- Currently the catalog contains datasets from ENES & CLARIN communities
- Working on adapting CKAN to EUDAT needs
- Priorities:
  - Proper use of OAI-PMH and adaptor modules for modularity
  - Switch on faceted browsing function next to tag based searching
  - Include more communities

# Metadata Catalogs

- EUDAT prototype

<http://eudat4.dkrz.de:5000>

- CLARIN VLO (Virtual Language Observatory)

<http://catalog.clarin.eu/ds/vlo/>




## Groups of Datasets

[List Groups](#) [Login to Add a Group](#)

Title	Number of datasets	Description
CLARIN	15	The Common Language Resources and Technology Infrastructure (CLARIN) project...
ENES	15	The project European Network for Earth System (ENES) supports the Earth...

**What Are Groups?**

Whilst tags are great at collecting datasets together, there are occasions when you want to restrict users from editing a collection. A **group** can be set-up to specify which users have permission to add or remove datasets from it.

### About EUDAT Meta Data

#### Repository

- [About](#)
- [Twitter @ckanproject](#)
- [API](#)
- [API Docs](#)
- [Contact Us](#)
- [Privacy Policy](#)

### Sections

- [Users](#)
- [Tags](#)
- [Statistics](#)
- [Revisions](#)
- [Site Admin](#)

### Languages

- [English](#)
- [español](#)
- [português \(Brasil\)](#)
- [日本語](#)
- [français](#)
- [italiano](#)
- [한국어 \(대한민국\)](#)
- [čeština \(Česká republika\)](#)
- [català](#)
- [suomi](#)
- [Ελληνικά](#)
- [svenska](#)
- [Српски](#)
- [Norwegian](#)

### Meta

© 2012 [Open Knowledge Foundation](#) Licensed under the [Open Database License](#) [OPEN DATA](#)

Powered by [CKAN](#) v1.8.




Search - EUDAT Meta Data Repository - EUDAT Meta Data Repository

CLARIN Virtual Language Obser... x CLARIN Virtual Language Obser... x Search - EUDAT Meta Data Rep... x

eudat4.dkrz.de:5000/dataset?tags=childes

Google

 EUDAT Meta Data Repository — contains harvested meta data from EUDAT communities












[Search](#) [Groups](#) [About](#)

## Search - EUDAT Meta Data Repository

Search...

Tags: **childes**

9218 datasets found

<a href="#">t1-mpi-pl-023004</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023005</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023006</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023007</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023008</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023009</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023010</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023011</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023012</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023013</a>	 Not Openly Licensed
<a href="#">t1-mpi-pl-023014</a>	 Not Openly Licensed

### Tags

- [UnitedKingdom](#) (1967)
- [UnitedStates](#) (1583)
- [Netherlands](#) (1158)
- [Israel](#) (740)
- [Spain](#) (687)
- [Germany](#) (527)
- [Poland](#) (411)
- [France](#) (270)
- [SouthAfrica](#) (192)
- [Italy](#) (175)

### Other access

You can also access this registry using the [API](#) (see [API Docs](#)).



# Search - EUDAT Meta Data Repository

Language:Japanese Search

18 datasets found

- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed
- "Japanese - Sakura Corpus" Not Openly Licensed

## Tags

- talkbank (18)
- Japan (18)

## Other access

You can also access this registry using the [API](#) (see [API Docs](#)).



CLARIN Virtual Language Observatory - Resources

catalog.clarin.eu/ds/vlo/

# Virtual Language Observatory

Explore the world of language resources and technology from different perspectives

VLO Home >> Faceted Browser Resources

SEARCH

Showing 1 to 10 of 280703

name	description
"Barberstue I ""Punch""	Scene fra barberstue. "Punch". Negativ og aftryk gave fra magister Hans Lassen. Foto som forarbejde...
"Barberstue I ""Punch""	Scene fra barberstue. "Punch". Negativ og aftryk gave fra magister Hans Lassen. Foto som forarbejde...
"Barberstue I ""Punch""	Barberstue "Punch" 4. 7. 1878. Negativ og aftryk gave fra magister Hans Lassen. Foto Nationalmuseet...
"Bryggeriet ""Stjernen""	Bryggeriet "Stjernen", Dronning Olgas vej 61. (Trap 5. udg. ). Verfilmning van het in de Tweede Wereldoorlog illegaal uitgegeven gedicht van Jan Campert
"DE ACHTTIEN DOODEN"	"De... beelden van de bouw en de eerste steenlegging van het groot centraal gesticht van de Joodsche...
"DE JOODSCHE INVALIDE" GAAT BOUWEN	"Expertness" from Structured Text? RECONSIDER: A Diagnostic Promoting Program
"Form til ""bernekager""	Form til "bernekager" (Lokal åbenråsk benævneise) Bøg. På den ene side en rytter, på den anden dels...
"Form til ""bernekager""	Form til "bernekager" (Lokal åbenråsk benævneise) Frugttæ. I den ene side en spindende kvinde, i...
"Form til ""bernekager""	Form til "bernekager" (lokal åbenråsk benævneise). Bøg. Kun udgravet på den ene side, hvor der l et...

### COLLECTION

- DK-CLARIN Repository (46378)
- Nederlands Instituut voor Beeld en Geluid Academia collectie (46156)
- childes (28595)
- Endangered Languages (21341)
- Language and Cognition (20545)
- talkbank (14243)
- Acquisition (13142)
- MPI CGN (12769)
- Bavarian Archive for Speech Signals (BAS) (11562)
- Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (8016)
- more...

### CONTINENT

- Europe (61806)
- North-America (21646)
- Asia (16920)
- South-America (8118)
- Oceania (5006)
- Africa (4410)
- Australia (2218)
- Middle-America (2176)
- North America (506)
- Australien (2)
- more...

### COUNTRY

- Germany (21118)
- United States (20340)
- Netherlands (19419)
- Japan (6953)
- United Kinodom (6460)
- Sweden (5815)
- Papua New Guinea (5542)
- Turkey (4476)
- Belgium (3965)
- France (3859)
- more...

### ORGANISATION

- CMU (42821)
- Max Planck Institute for Psycholinguistics (17520)
- NOS (11669)
- CLS-KUN (8859)
- Mangamania/Meios (7282)
- Nationalmuseet, Danmarks Nvere Tid (5944)
- sundhed.dk (5698)
- http://europa.eu/rapid/ (5330)
- NPS (5274)
- Ulrik Brinck (5147)
- more...

### DATAPROVIDER

- MPI IMDI Archive (132837)
- OLAC Metadata Providers (80493)
- CMDI Providers (66397)
- IPROSLA (976)

### LANGUAGE

- English (60436)
- German (28586)
- Dutch (24629)
- Spanish; Castilian (11161)
- French (9609)
- Japanese (7006)
- Turkish (6172)
- Swedish (6057)
- Chinese (2136)
- Polish (2065)
- more...

### GENRE

- discourse (78166)
- primary\_text (7376)
- language\_description (4310)
- story\_telling (3996)
- stimuli (3836)
- interview (3222)
- narrative (2697)
- question\_answering (1957)
- stimuli\_act-out (1569)
- movie\_description (1492)
- more...

### SUBJECT

- 9999999 (22829)
- sundhed og medicin (6971)
- dnt - danmarks nvere tid (6834)
- general (5365)
- Jura (2528)
- general linguistics (1536)
- lypology (1527)
- vooricthing (1499)
- politiek (1403)
- syntax (1149)
- more...

### RESOURCE TYPE

- annotation (80811)
- audio (69663)
- text (67936)
- video (43269)
- televsie (39303)
- unknown\_type (23251)
- image (9687)
- biocscop (2881)
- Televsie (1695)
- niet van toepassing (1168)
- more...

### NATIONAL PROJECT

- CLARIN-D (121700)
- CLARIN-NL (62588)
- DK-CLARIN (46378)
- CLARIN-EU (3701)
- LINDAT (CLARIN-CZ) (62)

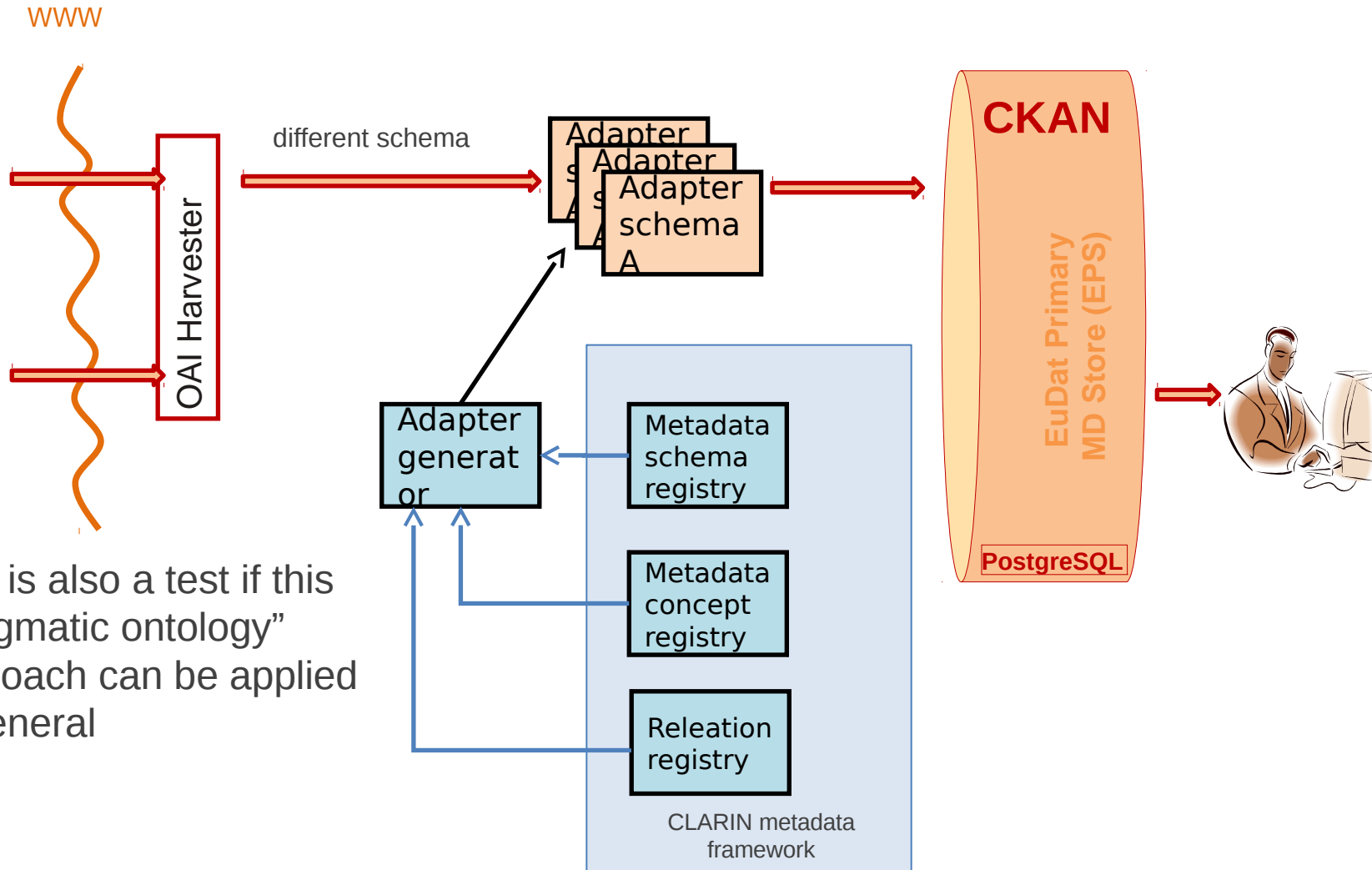
# Prototypes

- First prototype (target March 2013)
  - ENES, CLARIN metadata
  - CLARIN: 80.000 fixed schema record (C)IMDI
  - Limited number of adaptors.
- Second prototype
  - Specific topic related metadata:
    - Place: country, gps;
    - Time;
    - Subject (need subject taxonomy)

Thank you for your attention



# CLARIN particulars



This is also a test if this “pragmatic ontology” approach can be applied in general