CLARIN+

Working with Digital Collections of Newspapers

Leuven, 19-21 September 2016

CLARIN
Common Language Resources and Technology Infrastructure

CLARIN Data, Services and Tools

Menzo Windhouwer
Dieter van Uytvanck
CLARIN ERIC
menzo.windhouwer@meertens.knaw.nl

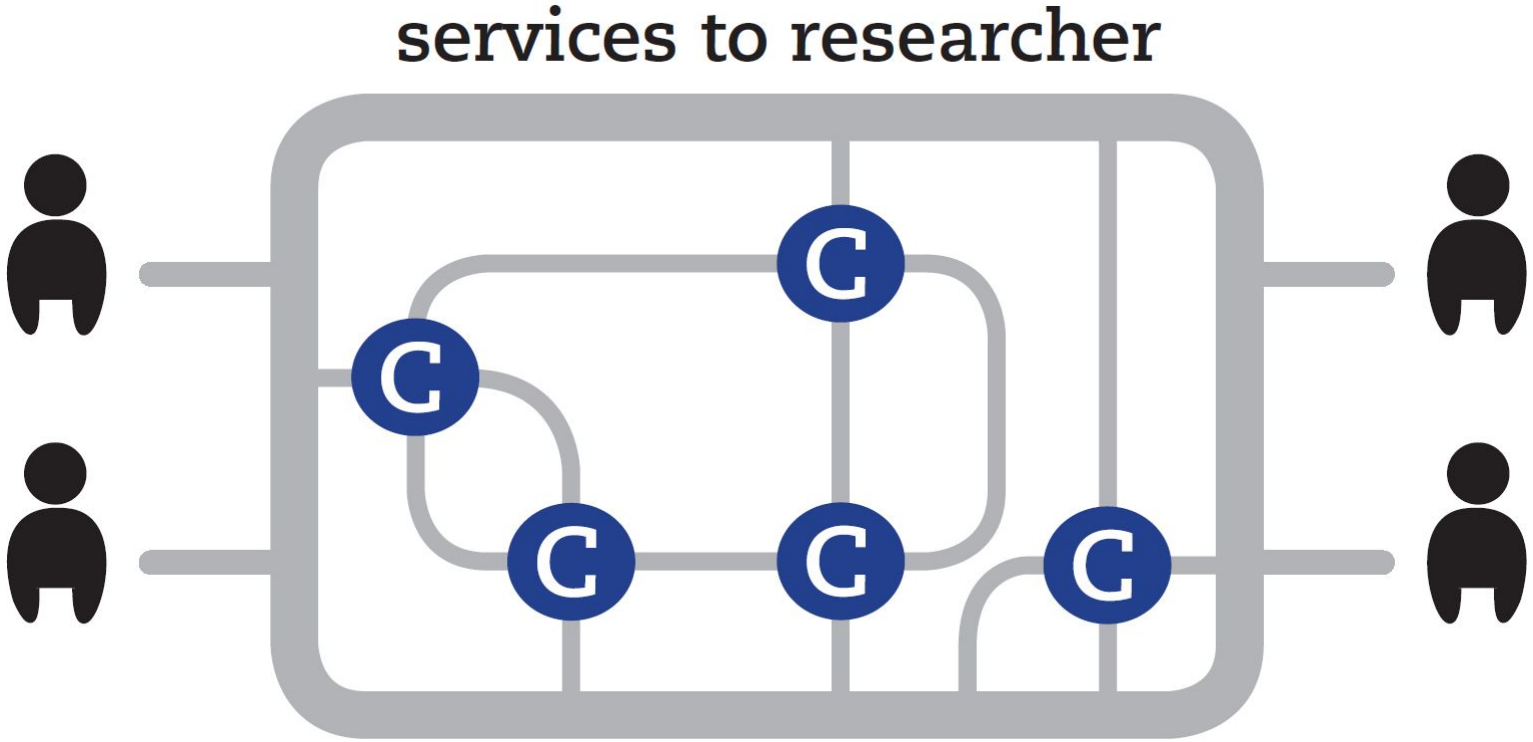
Working with Digital Collections of Newspapers
2016-09-19
Leuven

CLARIN?

- **C**ommon **La**nguage **R**esources and Technology **In**frastructure
- Research Infrastructure for the **humanities and social sciences**
- Provides easy and sustainable access for scholars
  - to **digital language data** (in written, spoken, video or multimodal form)
  - to **advanced tools** to discover, explore, exploit, annotate, analyse or combine them

CLARIN centres

- A **distributed architecture**: (http-accessible) files, web applications and web services spread all over Europe
- Nodes in the network: **centres** (http://clarin.eu/centres)



Organisation CLARIN

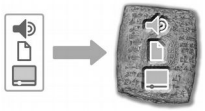
- European (ESFRI) Research Infrastructure
- ERIC since 2012
- Landmark since 2016
- **Members**:
  - Austria • Bulgaria • Czech Republic • Denmark • Dutch Language Union • Estonia • Finland • Germany • Greece • Hungary • Italy • Latvia • Lithuania • Netherlands • Norway • Poland • Portugal • Slovenia • Sweden • United Kingdom (observer)

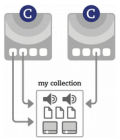
Benefits for countries

- **Access to the CLARIN Infrastructure**, i.e. to all CLARIN language resources and technology services
- **Access to expertise** via the CLARIN Knowledge Sharing Infrastructure
- **Embedding** in the humanities **research community**, with access to the same data
- **Better visibility of their language**, their research results, their resources and their **cultural heritage**
- **Opportunities**
  - for cross-lingual and -cultural **research**
  - to participate in **research projects** in which CLARIN ERIC participates as a beneficiary

The 34 CLARIN centres

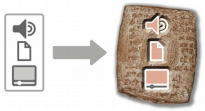
Our services for researchers

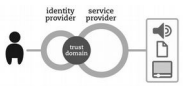


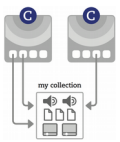


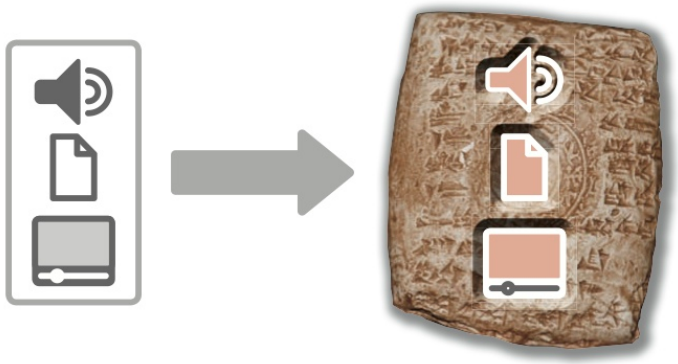
- Concrete and usable services
- All available via http://clarin.eu/services
  - (+ a whole set of technical services behind the scenes)

Our services for researchers



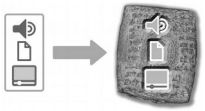


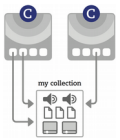




Depositing & Archiving

Our services for researchers





Virtual Language Observatory

VLO

Virtual Language Observatory Search Help CLARIN

VLO / Faceted search / Search results

newspaper

Showing 1 to 10 of 9381 results for newspaper × Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Collection

Type to search for more

- Leipzig Corpora Collection (7270)
- TLA: DiscAn (911)
- Institut für Deutsche Sprache, CLARIN-D Zentrum, Mannheim (674)
- TLA: Language and Cognition (106)
- TLA: MPI EVA corpora (88)
- UBU Clarin Set (77)
- European Language Resources Association (63)
- The LDC Corpus Catalog (29)
- CLARIN Centres (25)

<< < 1 2 3 4 5 6 7 8 9 10 > >>

newspaper

1 1

The text was recorded at Madison University in the 1960s. The text was recorded indoors.; The overall goal of the project is the documentation and preservation of the Hoocąk language. The project therefore includes the following sub-projects: (1) (audio- and video-)recording, analysing, processing and archiving a rep…

Newspaper articles

The Norwegian Newspaper Corpus

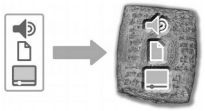
1

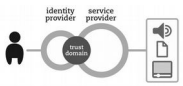
Dynamic, web-based newspaper corpus; 700 000 000 ws and growing; multitagged

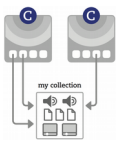
Romanian corpus of newspaper articles

1

Our services for researchers







Federated Content Search

Search text: newspaper

Search for: Any Language | Text Resources | in: All available collections | and show up to: 10 hits

23 matching collections found

Display as Key Word In Context | Download

Wikipedia – Institut für Deutsche Sprache

View

German Text Archive (DTA) – Berlin-Brandenburg Academy of Sciences and Humanities

View

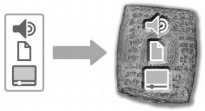
Ein gewaltiger Bau ist auch das Palais der Brisbane	**Newspaper**	Company , welches allein über 100.000 Pfund Sterling kostete und sieben Geschosse aufweist .
In diesem Augenblick , während englische Arbeiter mit Weib und Kind an Kälte und Hunger sterben , werden Millionen von englischem Geld , dem Produkt englischer Arbeit , in russischen , spanischen , italienischen und andern fremden Anleihen angelegt . " („ Reynold 's	**Newspaper**	, 20. Jan. 1867 .)

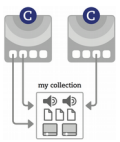
DWDS Core Corpus – Berlin-Brandenburg Academy of Sciences and Humanities

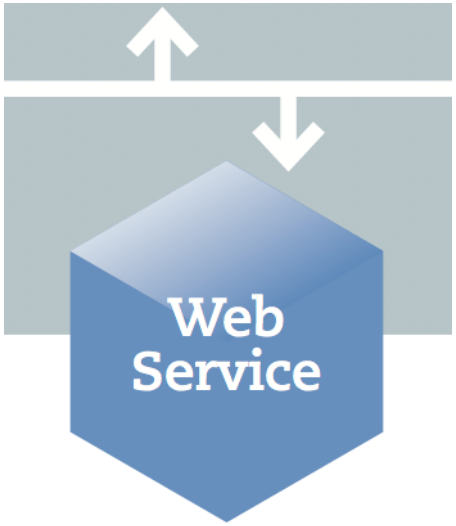
View

Die Verbreitung der Interviews wurde durch das Manhattan	**Newspapers**	Syndicate besorgt .
Die Engländer finden außer den örtlichen Zeitungen ihre andere Kost , Lloyds Weekly und Reynolds'	**Newspaper**	mit viel Verbrechen , Sport und Abenteuer .
Nusipepa = Brief , Schriftstück , Druckwerk (von	**newspaper**	, Zeitung) ; he savee look along newspaper = er kann lesen (er wissen schauen entlang Zeitung) .

Our services for researchers

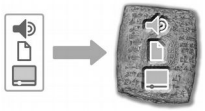


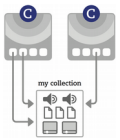


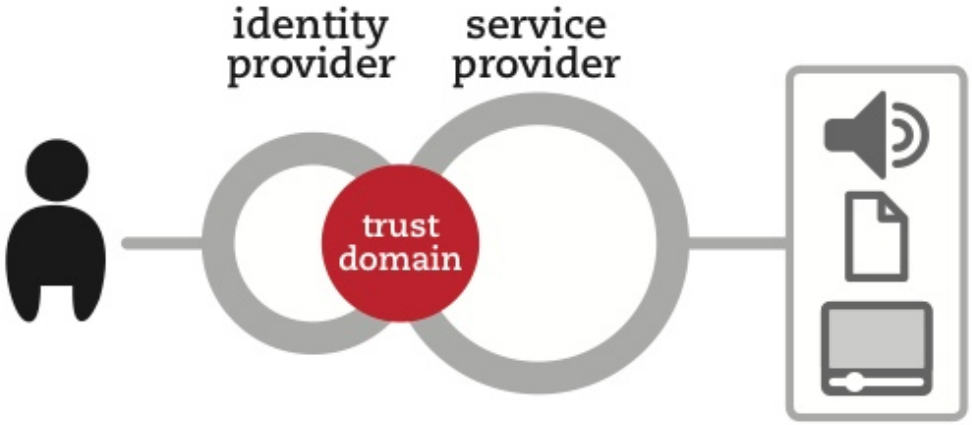


Web services and applications

Our services for researchers

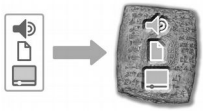


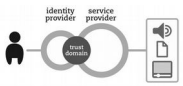


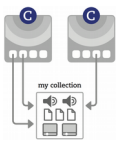


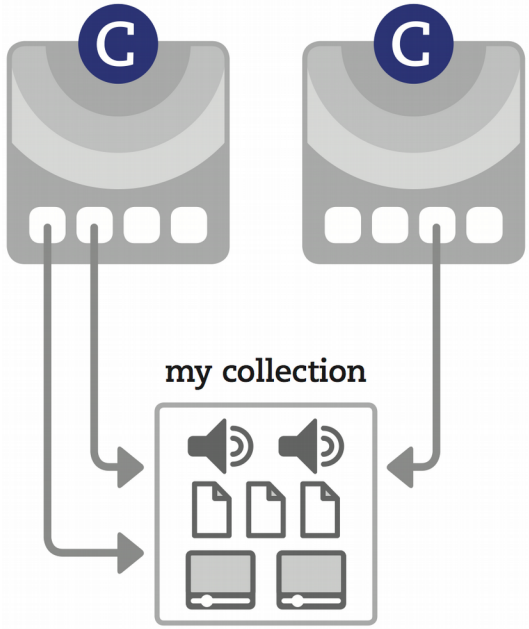
Easy access to protected resources

Our services for researchers









Virtual Collections

Absolute spatial deixis and proto-toponyms in Kata Kolok

General

Name: Absolute spatial deixis and proto-toponyms in Kata Kolok

Type: extensional

Creation Date: 2014-09-26

Description: Digital references for De Vos, C. (2014). Absolute spatial deixis and proto-toponyms in Kata Kolok. NUSA: Linguistic studies of languages in and around Indonesia, 56, 3-26.

Purpose: research

Reproducibility: intended

Persistent identifier: hdl:11372/VC-1001

Keywords:
- sign language
- Kata Kolok

Creators

Person: Connie de Vos

Organisation: Max Planck Institute for Psycholinguistics

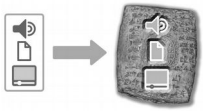
Website: http://www.mpi.nl/people/vos-connie-de

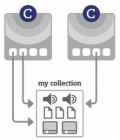
Role: Researcher

Resources

Reference	Type
Journal Article (fulltext) This paper presents an overview of spatial deictic structures in Kata Kolok, a sign language which is indigenous to a Balinese village community.	Resource
Footnote 3 - video Absolute versus absolute transpositional pointing signs	Resource
Footnote 4 - video COME-HERE-FROM-A and GO-FROM-HERE-TO-B	Resource

Our services for researchers





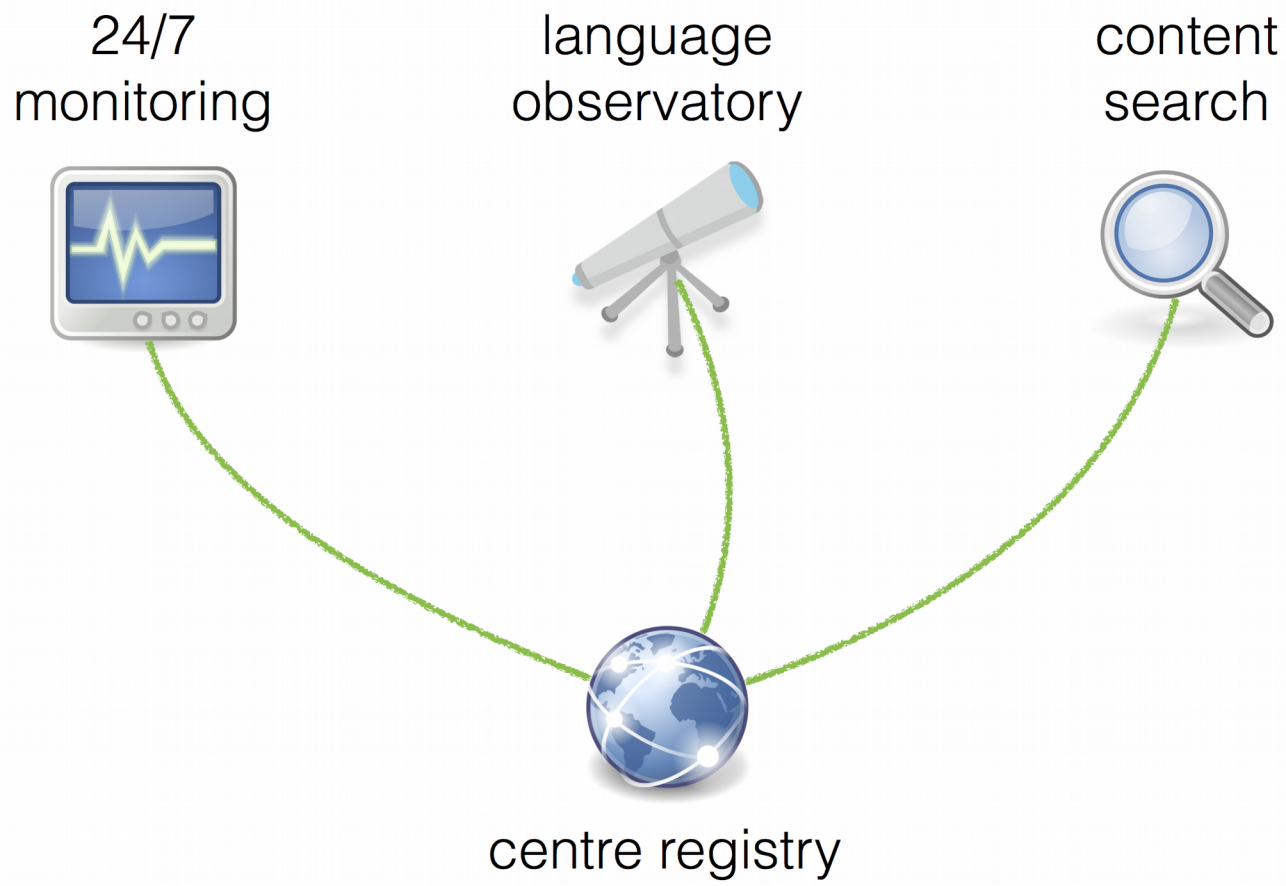
Consultancy

CLARIN technology pillars

- **Federated Identity** - letting users login to protected data and services with their own institutional username and password
- **Persistent Identifiers** - enabling sustainable citations of electronic resources
- **Sustainable repositories** - digital archives where language resources can be stored, accessed and shared
- **Flexible metadata and concept definitions** - to ensure semantic interoperability when describing language resources
- **Well-described and open protocols**, e.g.**:**
  - **Content search** - offering a search engine for a wide range of language resources
  - **Web service chaining** - giving users the possibility to freely combine language processing services

Seamless integration *within CLARIN*

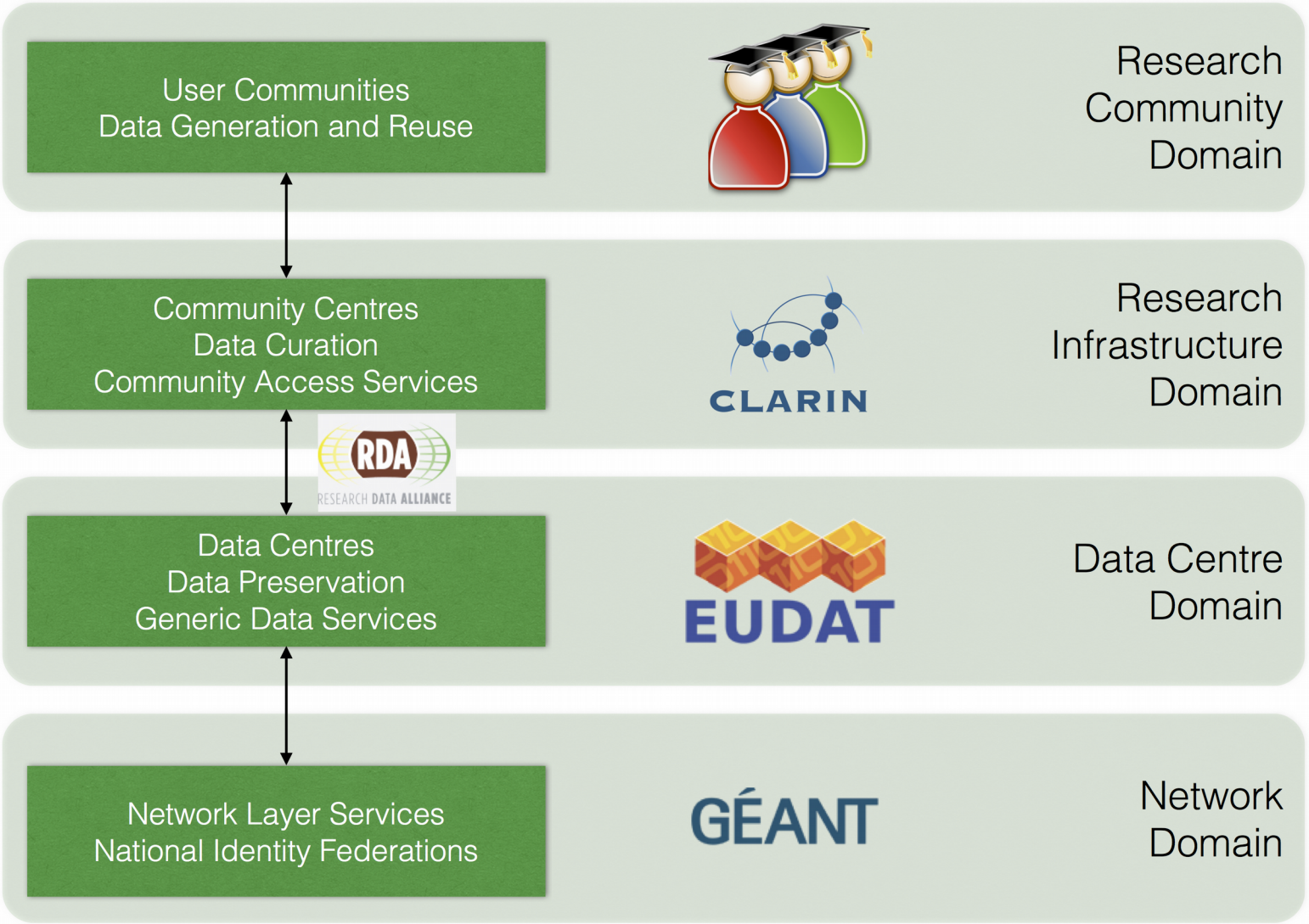
- Centres and Services are not isolated islands but part of a well-integrated setup



Seamless integration *within CLARIN*

- e.g. ELAN with WebLicht (tagging) and WebMaus (phonetic alignment)
- e.g. the Language Resource Switchboard

Seamless integration
in the infrastructure landscape



CLARIN for Digital Collections of Newspapers

- While CLARIN is based on a federation of long-term archives and services, it is not a static infrastructure
- This workshop is an excellent opportunity for us to:
  - listen to your needs
  - learn from your experiences
  - create additional bridges between the research infrastructure and potential users and data/service providers

Digital Collections of Newspapers – some first steps

- Would be nice to integrate them (all) into the Virtual Language Observatory
  - Challenge: gathering comparable metadata from the different sources (and maintaining it!)
    - at least dublin core via OAI-PMH endpoint
    - reuse/create a specific CMDI metadata profile?
  - What to do with non-digitized resources?
    - at least provide contact information for interested parties
  - What to do with access restrictions?
    - provide at least detailed information about how to obtain access
    - suggest single-sign-on
- Connect search engines to the Federated Content Search?

Digital Collections of Newspapers – some first feedback

- Most important: your ideas, suggestions and feedback!

Thank you for your attention!

For more details, please visit:
www.clarin.eu

or feel free to contact me at:
menzo.windhouwer@meertens.knaw.nl